THE PHONETICS AND PHONOLOGY OF SPEAKING STYLES AN OVERVIEW

K. J. Kohler
IPDS, Kiel, Germany

It is part of the competence of language users to adjust their phonetic output in structured ways to the demands of the different communicative situations and to be able to detect this situationally determined phonetic variability in the speech input. To refer to these phenomena the term 'speaking styles' is used.

It is a task for phoneticians to determine and analyse the physical parameters that have this stylistic signalling function and to associate them with phonetic and phonological categorization. Since there is growing interest in this field, the ICPhS is an appropriate platform to present its state-of-the-art and to develop guidelines for future research.

There are three basic areas of description that may be seen as factoring the manifestation of speaking styles:

- segmental reduction and elaboration
- prosodic patterns of timing, pitch and voice quality
- types and degrees of disfluency in speech production.

We cannot assume that there is a one-to-one correspondence between a classification into speaking styles and phonetic parameters. A remark such as "he has a teaching voice in his conversation" shows this discrepancy but also points to expectations listeners have. Since the differentiation between reading and spontaneous speaking is probably more clear-cut than finer gradings between speaking styles, and since researchers have embarked on precisely these areas - no doubt for methodological reasons - the contributions in this Symposium will also centre on them.

The team of participants was formed in a way to represent research activities with a wide geographical spread from Europe to Japan and to the United States and across a variety of institutions in academia and industrial application. The five contributions deal with the subject areas listed above:

W. J. Barry opens the round of presentations with a general discussion of the scope of speaking styles and of delimitations that go beyond the usually applied reduction and formality scales.

K. J. Kohler then concentrates on segmental articulatory reduction in scripted and in one form of spontaneous speech and topicalizes the balance between reduction of effort on the part of the speaker and cognitive-auditory constraints for the listener.

N. Campbell's paper represents the transition from the discussion of segmental to that of prosodic phenomena in speaking styles in that it introduces prosodic conditions (especially boundaries and prominences) for the realization of segmental strings in read and spontaneous speech.

G. Bruce's contribution then deals with prosodic structures of spontaneous dialogue vs. read speech in their own right.

J. Hirschberg's paper completes the discussion of exponency of spontaneous vs. read speech by adding disfluency indices to the prosodic ones.

A symposium of this kind has to impose limitations. The most serious one is the restriction to data from only five languages (English, German, Italian, Japanese, Swedish). But we hope that these shortcomings will be outweighed by the general principles we attempt to put forth. If we can stimulate interest in this area and motivate other researchers to expand our experience with this type of data the Symposium will have been a successful Congress event.

PHONETICS AND PHONOLOGY OF SPEAKING STYLES

W. J. Barry
Institute of Phonetics, University of the Saarland, Germany

ABSTRACT

This discussion attempts to identify some questions which have been neglected in the recent revived interest in speaking styles, and to discuss the theoretical implications of incorporating them.

THE SCOPE OF SPEAKING STYLES

Present approaches

In a recent paper on trends in speaking-style research, Maxine Eskanazi [1] selected three dimensions along which styles may be located.

a) the degree of (presumably phonetic) intelligibility required

b) familiarity between speaker and audiance, and

c) the social strata of those speaking.

She acknowledges the limiting nature of such a three-dimensional framework with a comment on the limits of the data available, and summarises a more comprehensive view of style with a complex of factors seen as operating mainly on the speaker:

"Style reflects the action of the environment upon the individual and the individual upon the environment. It is his perception of the various "status" levels of his listener and of the type of situation in which he finds himself. It is also the projection of himself, his background, and is a setting of the type and tone of conversation he wishes to have. All of this is a mixture of conscious and unconscious (...) effort on his part and is not always perceived in the same way as it is intended" ([1] p. 502)

The limits of the data available to redefine the framework, a scan of the literature suggests, are due to the scientific need for controls in data acquisition if differences in the results of analysis are to be attributed to the postulated differences in style. There is an understandably disproportionate emphasis on the difference between "read" or "prepared" speech and "spontaneous" speech. Another axis along which analytic comparisons are frequently made is "clear" vs. "casual" speech. Both these axes dominate, I suggest, because they can be turned into relatively easy instructions to, or situational parameters for the speakers during the acquisition phase. The effect on the three-dimensional framework suggested by Eskanazi is practically to collapse it into a one-dimensional "formal - casual" axis.

As an aside, it might be added that there are non-scientific interest-driven reasons for the restriction of speech types (I purposely avoid the term speaking style) of which we are all aware. It is therefore all the more important to ask whether there are convincing scientific reasons for *not* expanding the scope of speaking-style research.

My first question is, therefore, whether there is not an undue reduction of the concept of speaking style under discussion, and whether there are not important basic aspects which are not covered by the suggested descriptive framework?

Within this symposium, an extension of scope attempted in this paper is not intended as a focus for the other papers, but as a backdrop against which to place them in the perspective of the overall goal of characterising speaking style.

Other Dimensions of Speaking Style

An important consideration is that, in contrast to the production-oriented dimensions for categorising speaking styles summarised above, the concept of speaking style is fundamentally listener-oriented. Whether we consider the expressions used to describe someone's style of speaking, or the schooling of speakers for a particular task (politicians, managers, salespeople, etc.), it is the effect on the listener which is primary. The character actor is the epitome of audiance orientation.

Two assumptions are implicit in the term "speaking style", first that listeners have a personal neutral baseline against which they judge individual utterances, second that speakers' styles deviate to a greater or lesser degree as a product of their states of mind and their communicative intentions from their individual production baselines. If speaking styles are seen as communicatively significant, a considerable degree of isomorphism must be assumed between the production and listener baselines and the dimensions of variation around them.

Defining speech-production tasks and locating the resultant speech within the descriptive framework of the kind suggested in [1] does not allow for anomalous speaker performance in terms of effects on listeners. In particular the use of production categories such as "reading" and "spontaneous" as a basis for speaking-style analysis is misleading. The two categories cut across the dimensions of the Eskanazi framework, and conflate a vast number of different production styles. Radically differing performance at reading aloud, and a wide variety of situational and task variants within the category "spontaneous speech" make quantitatively meaningful statements across studies impossible.

For Speech Technology purposes, a differentiation of read and non-read speech may appear desirable, but it is not only very restricting in the search for phonetic correlates of communicatively relevant style features, it may well lead to nothing for Speech Technology. In most individuals, and in most groups within a given task definition, the difference may well be statistically clear, but the differences are not likely to be generalisable.

Phonetic Analysis

The dimensions of phonetic analysis, on the other hand, are common across speakers and tasks. They need to cover the three basic aspects of speech production: a) laryngeal excitation characteristics and related aspects of voice quality, b) modifications to the segmental structure of words and word sequences, and c) rhythmic and tonal properties of utterances resulting from the rate of speech, the grouping of words, and from the accentual and intonational patterns used.

Of these three areas of analysis, the second and third have received almost exclusive attention in recent work on speaking styles, as the other contributions in this symposium document. Voice quality has presumably been largely neglected for technical rather than theoretical reasons. Its contribution to the description of speakers and interaction in discourse are long established under the term paralanguage [2], and would add a valuable dimension to speaking-style analyses (cf. Campbell, this symposium), though its role as a functional prosodic feature also needs formal delimitation (at least in southern British English voice-quality alone ([±breathy]) can turn a rising contradictory query into a disbelieving query in the "He's not(?)" response to: "The bishop's coming to tea today.")

Practical Applications

The three phonetic analysis dimensions mentioned above form the basis of speaker analysis in the present-day, phonetically oriented approach to *forensic* speaker identification, though the short-term stuational and long-term speaker factors are conflated. As a first step, forensic phonetic work calls for the detailed auditory and instrumental analysis of any speech samples produced by the perpetrators, to give a profile of the person(s) being sought, before any suspects are asked to produce speech samples for comparison [3].

It is the listener-oriented side of speaking style that attracts the scrutiny of, and is modified by the image-building consultants engaged by politicians and other public figures. The consultants may not have a battery of phonetic facts available, but they are obviously able to identify, and communicate impressionistically, the parameters they think should be changed. Scientific phonetic goals should not be confused with PR (though commercially-oriented funding pressure has already taken its toll), but there is obviously an area of important practical application that mainstream phonetic research is neglecting at present.

Even in speech technology, the demands of which are to a large extent responsible for the restricted scope of work on speaking style in the past two decades, an open-minded consideration of the factors influencing speakers' production, once they depart from the

level of single-word utterances, is vital for open-ended progress. The implications for ASR are naturally greater than for speech synthesis because the recogniser in public service has to face the full range of voice-types and of personality-driven and situationally caused speaking styles. But in synthesis too, without quantitative knowledge of the phonetic parameters and the phonetic-phonological processes involved, progress towards task-differentiated synthesis will be blocked (see Campbell, this symposium).

The impact of these areas of professional activity on modern life is to my mind a sufficient reason for extending controlled study of speaking style beyond the formal-casual axis. This is by no means to say that the effects along that axis have been sufficiently explained, but it would, for example, be a disservice to engineers working in speech technology, to imply that the problems of spontaneous-speech will be solved by systematising that axis. To phoneticians, the vast complexity of speech production phenomena and their sensitivity to all aspects of the social, situational and psychological scenario goes (almost) without saying. But, as comments from non-linguists frequently document, to them speech is still the audible part of what can be written on a sheet of paper, and even for many engineers working in speech technology ta lot of effort has to be invested to show that signal variability can be related to systematic dimensions of the communicative situation, not just to the phonological, morphological and syntactic structure.

Difficulties of Analysis

The arguments for extending the classification base of speaking-style studies may convince without changing the unfeasibility of actually doing it. Past phonetic studies of emotional and attitudinal aspects of speech, as judged by listeners, are indeed not easily translatable into general statements of phonetic and/or phonological structure [4-9]. They *appear* to be much less easily controlled than speaker and production variables.

This is no doubt to some extent a valid objection, because if listener-oriented categories of speaking style are to be studied, control, in the form of listener-group judgements, has to follow the selection of a variety of spoken texts by each speaker to be analysed. As for the data collection itself, care has to be taken, a) to create a speech-production task and situation where the parameters under scrutiny are liable to occur, b) to obtain the validation by the speakers of the attitude, emotion, intention etc. underlying a particular utterance.

However, given these steps, control is in fact greater, due to formalised listener-group acceptance of the assumed category, than in solely production-oriented experiments. There, control is limited to instructions to the speakers (e.g. read, read/speak clearly, etc.) or definition of the interactional situation. The communicative effect of a speaker's behaviour in terms of a listener-group's descriptive categories is rarely formally checked.

PHONETICS AND PHONOLOGY

Another advantage of an extended scope of speaking style, though completely intrinsic to Phonetics and Phonology, would be a broader frame for the discussion of a topical theoretical issue, namely the question of phonetic vs. phonology within speaking-styles, with special reference to reduction phenomena.

To avoid misunderstanding, let me define my understanding of the part of phonology implicated in this discussion. I am considering the systematic basis of speech production at the level of motor programming, not merely a formal meta-system for the representation of parts of utterances (although that level of description is both necessary and important as an interface to morphology and syntax, as is a consideration of perception in phonology). The articulatorily based approach by Browman & Goldstein (B&G) [10, 11] exemplifies the aspect of phonology focussed on here. Their claims are currently under discussion and are eminently relevant to a discussion of speaking styles.

Briefly: B&G advocate a structure of gestures as phonological units which overlap and may be reduced, but which are not exchanged for other gestural units. Thus, all reduced forms have the same phonological form, and all reductions (e.g. along the formal-casual continuum) must be regarded as phonetic processes. Others (e.g. Holst & Nolan [12], Kohler [13]) while not arguing against a wide range of phonetic forms which can be explained as stages along a continuum of articulatory reduction (which may or may not be represented as different sequences of phonetic segments), argue for an ultimate form in the range of observed variants which reflects a categorical shift in the underlying phonological structure (see also Kohler, this symposium).

As so often in academic issues of principle, neither standpoint is provable because the ultimate answer is inaccessible to observation. In this case it lies in the production plans of the speaker. Nolan's and Kohler's indirect evidence in the form of careful analysis of surface behaviour is convincing from their own standpoint (see Kohler's extreme case of anticipatory assimilation "mit bunten Papierschlangen", this symposium), but apparently fails to convince the opposing camp.

The crux seems to be the degree of overlap and the degree of reduction that B&G's model allows, ultimately, how abstract their initially articulatorily interpretable model has become. However, that is no indication of what "gestural phonology" natural speakers actually possess, if gestures are what the organisational units of the production plan happen to be

Phonological Switch in Speaking Styles

Although not intended, the connection between reduction processes and the formal-casual speaking style axis can imply the general lack of phonological change with style change; i.e. style is a surface phonetic phenomenon.

Consideration of a wider range of stylistic phenomena shows, however, that changes only interpretable as "phonological" switches as well as changes in the surface-phonetic form are commonplace. Since many of these switches can be observed in connection with the same sort of socio-communicative variables as are implicated in the three dimensions behind the formal-casual continuum, the suggestion is, that phonological switches of the "gestural reorganisation" type are equally plausible.

Phonological Switches in Prosody

Before considering possible phonological switches involving segmental restructuring at the level of word sequences, we can illustrate that switches are commonplace at higher levels of phonological structure.

There are formalised speaking roles with recognized (and immediately recognizable) intonational patterns which differ radically from those used in any other form of speech. Examples of these, admittedly extreme in some cases, are certainly horse-race commentators, auctioneers, marketenders, sermons and community praying. The latter two cases have been described for English by Crystal [2].

Horse-race commentaries have not, to my knowledge, been described systematically, but the pattern is presumably familiar to many people; it seems to be similar across a number of languages: In British English there is a clear monotonization rule, with definite, race-stage-oriented resets (to a high pitch) with tempo and volume increases from one series of "intonation units" to the next, and with a sudden rallentando and decrescendo combined with a short series of resets to a lower pitch and a final low falling contour from the moment the winning horse finishes.

The other examples can be similarly characterised with a combination of local prosodic rules and an "operation-bound" pattern of intonation-unit sequences. The switch into any of these intonational systems is of course strictly situation-bound.

Another example of intonational switch, of a more subtle kind, has been observed during intonation work in Saarbrücken with southeast Italian (Bari) speakers [14]. In order to elicit spontaneous speech, a map-task is given [15, 16] in which one person verbally guides another along a route. Certain differences exist between the two maps, leading to frequent requests for information, confirmation etc. The two speakers "play" the game twice, once as guide, once being guided. The speakers

are unaware of the purpose of the recordings being made; only after the event do they fill in a form asking them about their language (dialect) habits. In one case, one of the speakers saw the questionnaire between the two "games". The researchers monitoring the recordings were puzzled that the speaker, who had been asking questions with a typical Bari rise + fall intonation [14] during the first game, consistently used the standard Italian rising contour in the second. When questioned about it, the speaker was in fact unaware of the manner of the"style" change, but confirmed that he had realized about the interest in accent from the questionnaire. This had triggered the unconscious switch to the more socially acceptable form.

A similar switch was observed in Palermo Italian in an earlier study in which dialect forms of question intonation were first elicited in a "game" situation and then presented in written form for controlled reading. The speakers again switched to the standard Italian intonation, and were unable to read with the dialect intonation until the questions were massively contextualized with dialectal precursors ([16] p. 143).

Segmental Switches

Both of the above examples are long-term switches; i.e. a phonological variant is selected which the speaker uses in a given situation. However, shorter-term switches are also common. It is extremely common for bilinguals to switch from one language to the other, and often back again, in the middle of a sentence. What is more, the switch is often unconscious, and the speaker can continue some time without being aware of having switched. This well-studied phenomena [18, 19] is always linked to a "trigger" word or event which is associated with the language that is switched into. This might be considered too far removed from speaking-style phenomena; after all, a sentence, or part of one, requires morphological and syntactic switches as well as phonological and phonetic ones. However, a minimal switch may extend for one word only, and the situational factors facilitating switches are likely to be located at the familiar end of the familiar - non-familiar dimension.

The examples so far indicate that, given the right situational factors, individuals generate utterances while selecting from different (parallel) sets of phonological rules.

Reducing the bilinguality to bi-dialectality, we get ever closer to normal speakers and their phonetic-phonological adjustment to situational factors. Presumably, most people have experienced dialect speakers' in a standard-language situation drifting into dialect as they discuss a point among themselves. Among highly educated people, who possibly spend more time speaking "standard" than dialect, it is a signal of solidarity and familiarity (even complicity, since it is often employed when negotiating a favour or service).

Important for the discussion of the "phonological switch" phenomenon is the considered use, above, of the term "drifting into dialect". My observation of Saarland dialect speakers, confirmed by phonetically aware dialect speakers (in fact the phenomenon appears to be general to bi-dialectal areas) is that the switch is seldom from standard German straight into full dialect. There is an intermediate form which is clearly close to the standard in lexical choice and even syntax but contains definite phonological and/or morpho-phonological modifications:

Standard		Saarbrücken
/...st/	-->	/...ʃt/
(weißt du	-->	[vɛ:ʃtə])
(erst	-->	[ɛɐʃt])
/..V(:)Cən/	-->	/..V(:)Cə/ or /..V(:)C/

(willst du messen --> [vɪlʃtə mɛs(ə)])

The use of standard words with a modified phonology is again important evidence for stylistically triggered phonological change. These alternative forms obviously cannot be explained by "reduction" processes. The speakers are either "switching on" a phonological rule which carries out the change to the dialectal sound structure, or they have both standard and dialectal forms in their lexicon to access appropriately, depending on the situation.

Finally, we find that non-bilingual, non-bidialectal speakers produce forms

of the same words and phrases which are far apart on a putative continuum of change, possibly from near the maximum form to near the most reduced, but do not produce forms in between. For example:

"going to" (/'gəʊɪŋ tu:/)
can easily be shown to reduce to
"gonna" (['gʌnə])
in a logical series of steps. However, the segmentally very reduced form is likely to appear at relatively slow speech rates, and even in accented position. This rules out the possibility of it emerging as a result of overlapping underlying gestures. The alternative, stylistically marked lexical entry may therefore be a more plausible explanation.

Similar examples are those given by Kohler (this symposium) as "stereotype formulae", which can also be interpreted as implying its own lexical (sub-) entry:
'kyou (/kju:/) for "thank you"
'nAbend (/'nɑ:bmt/) for "guten Abend"
In both cases a lexically stressed syllable is missing compared to the standard expression, which cannot be attributed to any normal elision rule.

Both "thank you" and "guten Abend" are written as two words but are presumably stored as single entries because of their formulaic function (stroke patients, unable to articulate more than single monsyllabic words, are known to produce formulaic expressions such as "thank you" spontaneously).

Lexical Entry or Gestural Recoding?

Explaining the phonological switch only in terms of alternative lexical access rather than an on-line change of phonological structuring during the articulatory planning stage of production is not satisfactory. The parallel between these examples and the critical issue of gestural overlap or phonological recoding only applies when the reduction is within a single lexical item. Words like:
greatcoat: /greɪtkəʊt/ <--> /greɪkkəʊt/
and
handbag: /hændbæg/ <--> /hæmbbæg/
satisfy this, but the similarity between this intra-word alternation and the possibility of the same alternation occurring across word boundaries, e.g:
He pulled his hand back:
/... hænd bæk/ <--> /... hæm: bæk/
makes an on-line phonological recoding mechanism rather than just style-based alternatives in the lexicon necessary, unless B&G can convince the other side of the plausibility of an absolute overlap + reduction principle.

SUMMARY

In the first section of this contribution to the speaking style discussion, we offered arguments for a need to extend the scope in the study of speaking styles, and argued both that the listener-orientation is basic to the concept of speaking-style, and that it has been largely ignored in recent work.

It was also argued that scientific stringency need not suffer and that the extended scope would, in fact, provide a step towards the refinement of the framework which Eskanazi considers necessary ([1] p. 501).

In the second section, we discussed the phonetics-phonology issue of gestural overlap and reduction vs. phonological reorganisation of gestures within the suggested wider scope of speaking style work. The commonplace occurrence of style-linked phonological switches was illustrated at different levels.

In the case of segmental variants, the alternative lexicon entries were seen as a possible explanation for single-word cases. Across word boundaries, however, similar variants could not be explained in this way, and style-based phonological processes were implicated. It was seen, though, that parallels with the original gestural issue were limited and highlighted it rather than resolving it.

REFERENCES

[1] Eskanazi, M. (1993): Trends in speaking styles research. *Proceedings Eurospeech 93*, Berlin. 501-509
[2] Crystal, D. (1975): *The English Tone of Voice. Essays in intonation, prosody and paralanguage*, London: Arnold
[3] Künzel, H.J. (1987): *Sprechererkennung. Grundzuege foren-sischer Sprachverarbeitung*. Heidelberg: Kriminalistik Verlag
[4] Fónagy, I. (1964): L'information du style verbal. *Linguistics* 4, 19-47
[5] Fónagy, I. (1969): Métaphore d'intonation et changement d'intonation.

Bulletin Soc. Linguistique de Paris 64, 22-42
[6] Fónagy, I. & Bérard, E. (1972): "Il est huit heutres": Contribution à l'analyse sémantique de la vive voix. *Phonetica* 26, 157-192
[7] Uldall, E.T. (1964): Attitudinal meaning conveyed by intonation contours. *Language & Speech* 3, 223-234
[8] Uldall, E. (1964): Dimensions of meaning in intonation. In: D. Abercrombie, D.B. Fry, P.A.D. MacCarthy, N.C. Scott & J.L.M. Trim (eds.), *In Honour of Daniel Jones*, London: Longmans, 271-279
[9] Hadding-Koch, K. & Studdert-Kennedy, M. (1974): Are you asking me, telling me, or talking to yourself? *J..Phonetics* 2, 7-14
[10] Browman, C.P. & Goldstein, L. (1992a): "Targetless" schwa: an articulatory analysis. In: G. J. Docherty and D. R. Ladd (Eds.), *Gesture, Segment, Prosody. Papers in Laboratory Phonology II.* Chapter 2, pp. 26-55. Cambridge: University Press.
[11] Browman, C.P. & Goldstein, L. (1992b): Articulatory Phonology: An Overview. *Phonetica* 49, 155-180.
[12] Holst, T. & Nolan, F. (in press): The influence of syntactic structures on [s] to [S] assimilation. In: B. Connell and A Arvaniti (eds.). *Phonology and Phonetic Evidence. Papers in Laboratory Phonology* IV. Cambridge: CUP.
[13] Kohler, K.J. (1994): Glottal stops and glottalization in German. Data and theory of connected speech processes. *Phonetica* 51, 38-51.
[14] Grice, M., Savino, M., (1995): Low tone target versus "sag" in Bari Italian intonation; a perceptual experiment. *Proc. XIII Int. Cong. Phonetic Sciences*
[15] Grice, M., Benzmueller, R., Savino, M., Andreeva, B. (1995): The intonation of queries and checks across languages: Data from Map Task dialogues. *Proc. XIII Int. Cong. Phonetic Sciences*
[16] Anderson et al. (1991): The HCRC Map Task Corpus, *Language and Speech* 34, 351-366
[17] Grice, M. (1995): *The intonation of interrogation in Palermo Italian. Implications for intonation theory.* Niemeyer (=Linguistische Arbeiten 334)
[18] Clyne, M.G. (1967): *Transference and Triggering.* The Hague
[19] Clyne, M.G. (1969): Switching between language systems. *Actes du Xe Congrès International des Linguistes.* Bucarest, 343-349

ARTICULATORY REDUCTION IN DIFFERENT SPEAKING STYLES

K. J. Kohler
IPDS, Kiel, Germany

ABSTRACT

This contribution outlines six principles for the insightful analysis of articulatory reduction and discusses the results of a comparison of some connected speech phenomena retrieved from two labelled data bases of German: the Kiel Corpus of Read/Spontaneous Speech.

PRINCIPLES

The study of articulatory reduction should take the following principles into account:

- reduction of effort in connected speech
- listener orientation: auditory constraints
- gestural reorganization
- control of global functional coordinative structures rather than of individual anatomical articulators: cognitive constraints
- constraints imposed by the varying demands of different communicative situations: speaking styles
- phonetic, rather than phonemic, processing of large acoustic data bases within a framework of complementary phonology.

These principles will be discussed one by one in the above order with reference to some examples from English, but I will in the main draw on reduction data from German scripted and spontaneous speech that has been collected and phonetically labelled at IPDS Kiel. The scripted data consist of acoustic records of isolated sentences and two texts and are from 53 Standard German speakers of a North German variety [1]. The spontaneous speech corpus was recorded by another 26 speakers, with the same dialect background, in an appointment scheduling dialogue scenario of the VERBMOBIL project [2,3]. The signal and label files, supplemented by phonetic variants lexica of all the word forms occurring in the corpus, are available on two CDROM's [4,5]. As the labelled data are in a modified SAMPA notation [1], this too will be used for illustrations in this presentation.

REDUCTION OF EFFORT

Assimilations and elisions at word boundaries, as well as reductions in function words follow a principle of reduction of effort in connected speech. For instance tongue tip gestures are eliminated in favour of lip and tongue movements (e.g. *it's in that box* [pb], *not in that case* [kk]). I have argued in several publications, e.g. [6], that speech articulation is characterised by constantly ongoing lip and tongue dorsum movements, with tongue tip gestures riding on them at specific points, almost exclusively associated with consonants. The two types of movements are also linked with different muscle sets. Apical gestures in this view are the special, marked feature in speech production, which is, on the one hand, put to use in the functional domain of languages (inflexions in Indo-European languages, deictics, articles), and is, on the other hand, weakened to reduce articulatory effort, related to the coordination of tongue tip to simultaneously ongoing tongue body movements. The assimilation of apical stops and nasals (/t/, /d/, /n/) to labials and dorsals is thus referred to articulatory-physiological mechanisms, which at the same time exclude assimilations of labial/dorsal to apical and those between labial and dorsal.

In German, '/@/ + nasal' in word-final syllables is a particularly interesting example of this apical to labial/dorsal assimilation, going either from right to left or from left to right. In this context, /@/ may be deleted, i.e. the apical and labial/dorsal gestures may get closely linked in time and thus be affected by the articulatory reduction principle. So canonical ['a:b@nth] *Abend* may be realized as ['a:bmth] or ['a:th@m] *Atem* as ['a:pm]. For canonical sequences of 'lenis plosive + /@/ + nasal' the reduction may go even further, producing a nasal gesture throughout the cluster and even reducing it to short duration; so we may, for instance, get the set of pronunciations, from most elaborated to most reduced, ['a:bEnth], ['a:b@nth], ['a:bnth], ['a:bmth], ['a:mmth], ['a:mth] (the first one being a reinforcement from canonical /@/). All these forms occur in connected speech.

These synchronic processes are to be differentiated from diachronic ones in sound change (e.g. Lat. octo > Ital. otto), where the prime cause is auditory, not articulatory, in the transmission from one generation of speakers to another. This leads to the second principle.

LISTENER ORIENTATION

The speaker-related control in synchronic processes is checked by listener orientation.

(a) First of all the position in the syllable is relevant: the initial place requires greater distinctivity for the listener's discrimination than the final place, especially in stressed word-initial syllables. They are landmarks for the listener to decode the incoming signal and to relate it to word sequences at higher processing levels. This excludes the assimilation of word-initial apicals to preceding word-final labials or dorsals (English *hat pin* [pp] vs. *tip toe* [pt], German *Lippen* [pm] vs. *gib nicht* [pn]). The acoustic-auditory distinctivity of the release burst and the aspiration strengthen the initial position and preclude assimilatory reduction.

(b) But there are also listener-oriented constraints related to segment features, namely the manner features stop/nasal/fricative. Stops only assimilate if they are not released (English *apt* [pth], *act* [kth], *picked* [kth], but *he picked me* [k(p)m] *a good one*; German *Akt* [kth], but *Zwischenaktmusik* [k(p)m], *Beamter* [mth6], but *Beamten* [mpm]). This is so because only the unreleased plosive has little perceptual distinctivity, with rather small differences between [t], [p], [k]. And what is not well differentiated for a listener anyway can be levelled more easily in the speaker's attempt to save effort. Released and aspirated stops with distinctive local acoustic frication properties do not meet this requirement. Nasals, being far less clearly differentiated according to place of articulation than released stops, can be assimilated in word-final position (English *happen* [pm], *organ grinder* [gN]; German *geben* [bm], *legen* [gN]). Fricatives [f], [x] vs. [s], on the other hand, are acoustically as well as auditorily very distinct and are not assimilated (English *this form* [sf] vs. *that place* [pp], German *Schußfahrt* [sf] vs. *Rundfahrt* [mpf] or *Schrottplatz* [pp]).

GESTURAL REORGANIZATION

Gestures disappear, and there is gestural reorganization. This differs from Browman & Goldstein's position [7], according to which only the timing and the amplitude of gestures are changed in a very mechanical way on the articulatory surface, whereas the gestures themselves remain as such, and new gestures cannot be created in the articulatory execution of a gestural score. However, with the disappearance of the apical gesture in the change [mth6n] > [mpm] for German *Beamten*, the stop formation and release are also changed: they are exclusively effected by velic control. There is no way of subsuming this under the variables of timing and amplitude. This becomes even more evident in the change [nth@np] > [mpmp] of German *mit bunten Papier-*

schlangen. As early as the first nasal the articulators have to be instructed for a labial gesture, although the triggering element comes last. Simple passive adjustment through contiguity cannot explain this phenomenon; there has to be active reorganization. Moreover, instead of the assimilated inter-nasal [p] we also find either a glottal closure or glottalization interspersed into a continuous nasal, i.e. the labial closure and the velum lowering are maintained throughout the sequence [m...m]. This means that the phonatory break, signalling a stop to a hearer, is achieved through replacing a velic closure by a glottal one or by creak [8]. An articulatory action is thus transferred from one movable structure of the vocal tract to another for gestural economy [9], adhering to a principle of functional equivalence for the output to a hearer. Again, only gestural reorganization at a more central level can explain this phonetic process. From the only possible explanation of these empirical data follows the fourth principle.

FUNCTIONAL COORDINATIVE STRUCTURES AND COGNITIVE CONSTRAINTS

Degrees of reduction in accordance with the balance to be struck between reduction of articulatory effort on the part of the speaker and perceptual discriminability on the part of the listener are governed by a reduction coefficient, which controls whole sets of articulators forming a global functional coordinative structure. So, e.g., in the series of reductions from [mIthde:m] through [mItm], [mIpm], [mIbm] to [mIm] in the German phrase *mit dem Bus* [10] we are dealing with a progression through three successive domains: I. the reduction of opening-closing movements, II. the coarticulation between apical and labial gestures and the increasing reduction of the former, III. the progressive shortening of the oral and velic closure configuration. In spite of reorganizations within and between these three domains, they constitute a continuous scale of reduction, along which the reduction coefficient is located for a speech act, i.e. differently for different speaking styles, e.g. scripted vs. spontaneous.

This reduction coefficient is governed by higher-level cognitive processing as an essential prerequisite to speech production: the gestural adjustments are primarily not constrained by the vocal tract and passive changes in timing and amplitude, but are conditioned to a large extent by an adaptation, on the part of the speaker, to the acoustic-auditory needs of the listener in given communicative environments and to the semantic and syntactic demands of the utterance. So speakers reduce less in response to the request for repetition or in unfavourable contexts of situation, e.g. in noisy and more formal conditions, when failing to be understood is rated high (cf. also Lindblom's H & H theory [11]). Moreover, the semantic content of an utterance and its syntactic structure check the degree to which articulatory reduction can operate. Content words are not normally subjected to the same simplification as function words, unless they are weakened semantically at the same time, as in greetings and other sterotype formulae in phatic communion (compare English *kyou* as against *many thanks* and German *n'Abend* versus *guten Appetit*).

The reduction of function words is furthermore governed by word class (e.g. German *ihr* as a personal pronoun is reduced more than in the function of a possessive pronoun) and by position in a syntactic structure (e.g. enclitic *ihr* in German *habt ihr (den Film gesehen)?* is reduced more than proclitic *ihr* in *ihr habt (mich enttäuscht)*). In the German sentence *Er ist der, der der Sache am meisten schadet.* ("He is the one that endangers the cause most."), *der* refers to the demonstrative or the relative pronoun or to the definite article, and all three can be represented by the phonological form /d'e:r/, but the first is realised as [d'e:6], the second as [dE6], the third as [d6], with progressive reduction in accordance with decreasing prominence in different syntactic slots, for which the reduction coefficient is set at different points on the reduction scale (see also [12]).

SPEAKING STYLES

Articulatory reduction is more frequent and more extreme, the closer the speaking style is located to the informal and spontaneous ends of the formality and spontaneity scales. There are two ways in which spontaneous speech differs from reading style speech production with regard to articulatory reduction phenomena: either the degree of gestural levelling is increased to produce more extreme articulatory simplification, or the reduction features that are found in scripted speech have a higher frequency of occurrence, they turn up more readily. Both possiblities can be illustrated by a comparison of data from the two German corpora.

Spontaneous dialogue provides many examples that go beyond the rules set up for German on the basis of scripted speech [6]. The following two instances from the Kiel VERBMOBIL corpus [5,15], which can be multiplied manifold, illustrate the phonetic processes at work.

(a) The phrase *wahrscheinlich ein bißchen* ("probably a little") has the canonical citation form (in SAMPA trancription):

va:6#S'aInlIC QaIn+ b'IsC@n

but is realized (labelled) as:

v a:6 #S 'aI n -MA l- I- C- Q- aI- n-m+ b 'I s C @- n,

where a symbol followed by '- space' means deletion, a symbol preceded by 'space -' insertion and 'symbol - symbol' replacement. So in this case the segments of the final syllable of *wahrscheinlich* are all deleted, with reference to the citation form, but their long component 'palatality' is preserved as a feature in the preceding nasal, which has palatal place of articulation. To mark this componential (prosodic) feature in a linear segmental transcription, -MA is inserted as a general marker that can be substantiated by subsequent speech signal analysis. -MA symbolizes componential features that are still present in spite of the disappearance of delimitable segments (e.g. nasalization, labiodentalization, vowel quality and duration residue, palatalization, velarization etc.). It receives its componential meaning from the segments marked as deleted in relation to the segment preserved before it. This subcategorization will eventually be performed automatically by computer programme based on distinctive feature representations of the segments.

(b) In the phrase *ich kann Ihnen das ja mal [sagen]* ("I can perhaps suggest this to you") the canonical form is:

QIC+ kan+ Qi:n@n+ das+ ja:+ ma:l+,

but its realization is labelled as:

Q- I C+ k -h a n+ -MA Q- i:- n @- n-+ d-n a s+ j a:+ m a: l+.

Here the high dorsum articulation for the first vowel in *Ihnen* is transferred to the alveolar nasal consonant following it. The apical nasal consonant articulation is maintained from the end of *kann* through *Ihnen* to the beginning of *da*, but inside it palatalization reflects the vowel of the function word *Ihnen*, although it is no longer realized as such.

These extreme reductions are either absent or very rare in speech reproduced from writing. There was only one speaker in the Kiel Corpus of Read Speech who provided instances of this type. For example his rendering of *morgen früh*, canonically represented as [m'O6g@n fr'y:], was [m'O~ fr'y:], with a nasalized vowel. So this speaker went beyond the well-known process of consonantal nasalization, found in scripted speech and derivable in the sequence [g@n]>[gn]>[gN]>[NN]>[N], by not only advancing the lowering of the velum in time but by changing the dorsum closing gesture as well. This informant is characterized by a general articulatory imprecision, which, in everyday interchanges, also prompts the re-

Table 1. Absolute frequencies of /@/ elision and various assimilation processes in '/@/ + nasal' of word-final syllables

(a) Read Speech		(b) Spontaneous Speech	
total number of words in corpus	*31,374*	*total number of words in corpus*	*9,291*
total number of sample words, incl. 52 compound double entries	*5,117*	*total number of sample words, incl. 16 compound double entries*	*1,605*
/@/ reinforcement	10	/@/ reinforcement	1
/@/ preservation	950	/@/ preservation	109
/@/ deletion	4,157	/@/ deletion	1,495
- place assimilation	1,146	- place assimilation	339
-- across word boundaries	52	-- across word boundaries	50
-- lenis stop nasalization	13	-- lenis stop nasalization	18
-- no lenis stop nasalization	824	-- no lenis stop nasalization	151
- no place assimilation	2,991	- no place assimilation	1,134
-- contextually not possible	2,414	-- contextually not possible	846
-- contextually possible, but not made	312	-- contextually possible, but not made	72
-- lenis stop nasalization	151	-- lenis stop nasalization	99
-- no lenis stop nasalization	333	-- no lenis stop nasalization	41
- treatment of /n/		- treatment of /n/	
-- deletion after nasalized stop	145	-- deletion after nasalized stop	93
-- deletion after canonical nasal	120	-- deletion after canonical nasal	119
-- preservation after any nasal, with place assimilation	314	-- preservation after any nasal, with place assimilation	136
- extreme reduction	20	- extreme reduction	22

quest for repetition more often than usual.

Although the two labelled corpora for German are of different sizes (31,374 word forms of scripted and 9,291 word forms of spontaneous speech) and from a different number (but homogeneous dialect group) of speakers, comparative relational statements about the occurrence of reduction phenomena are possible. I have selected the treatment of '/@/ + nasal' in word-final syllables, already mentioned. Table 1 provides an overview of the data, including non-final word components in compounds.

The proportion of the sample size to the total corpus size is comparable for the two corpora. Preservations and reinforcements of the vowel make up 7% of the sample in spontaneous and 19% in read speech, which is an indication that /@/ is dropped more readily in spontaneous interactions. Contrariwise, place assimilations of /n/ as a proportion of the total of possible place assimilation environments are 82% for spontaneous and 79% for read speech. It may thus be concluded that the two speaking styles do not differ in apical gesture reduction. However, if this parameter is looked at in combination with the nasalization feature in lenis stops there are clear differences: (a) nasalization is added to place assimilation in 11% of the possible cases of unscripted, but in only 2% of scripted speech; (b) in the absence of place assimilation, the incidence of nasalization is even more strikingly in favour of spontaneous speech: 71% vs 31%; (c) overall percentages of nasalization of lenis plosives are 38% vs. 12%.

The treatment of /n/ in 'nasal + /@n/' syllables is also characteristically different between the two speaking styles: in spontaneous speech the nasal is preserved - with place assimilation to the preceding nasal of any origin (canonical nasal or nasalized lenis stop) - in 39% of the cases, as against 54% in read speech. This reflects the greater tendency in spontaneous speech to reduce long syllabic consonants to short non-syllabic ones. These data also indicate that the reduction of 'lenis stop + /@n/' syllables is more likely to go beyond the place assimilation and nasalization stages in that speaking style. Finally, the proportion of more extreme reductions, beyond these well-established categories of /@/ elision, place assimilation, nasalization and syllabic nasal deletion, is greater in spontaneous speech; in reading style the relevant instances are largely due to the one speaker already mentioned: the example from his speech discussed above constitutes an extension of the reduction scale in 'lenis + /@n/' syllables at issue here.

The data suggest that articulatory features show different degrees of susceptibility to reduction. The elimination of a separate opening-closing movement in /@n/ syllables, i.e. a simple change of gestural timing, seems to be common in both speaking styles, but reading clearly restrains this tendency. Similarly, place assimilations reflecting gestural reorganization, although not as frequent as /@/ elision, are also quite regular when the special contextual conditions prevail, but the two speaking styles do not seem to differ. On the other hand, nasalization of lenis stops and reducing syllabic consonants, again due to gesture timing, are less common, and there is a large difference between the two speaking styles in the frequency of occurrence. The further reduction involving another gestural reorganization to eliminate a closing movement of, e.g., the tongue dorsum is even less likely to occur and closely linked to a spontaneous speaking style, even if it is introduced into reading by speakers who obviously have a less well defined separation of phonetic registers.

PHONETIC PROCESSING

The corpora of data used as a basis for the aforegoing discussion are labelled within a linear segmental phonemic framework. A powerful grapheme-to-phoneme conversion module within the RULSYS/INFOVOX TTS system [1,13] is used to automatically generate phonemic notations from the orthographic input (scripted text files read by the subjects and transliteration files of spontaneous recordings [14], respectively). After manual correction (appr. 3% error rate for running text), canonical transcription files result, which, together with the corresponding speech signal files are input to an adaptation of the KTH/ Stockholm MIX programme [13] for linear segmentation and labelling, with guidance from graphic signal plots (oscillogram, spectrogram) and acoustic output. The result is a canonical transcription file representing the actual pronunciation within a segmental description. All the labels of the canonical transcription are taken over, modified as deletions or replacements, if necessary, and supplemented by marking insertions. It is at this stage that additional phonetic labels are introduced at the subphonemic level (plosive releases, nasalization, creaky voice etc.). Many, but by no means all, of these markers are again segmental and linear (but cf. the use of -MA in the section on 'Speaking Styles').

This linear segmental frame allows the systematic and economical representation of lexical items in canonical citation form, to which actually spoken word forms can be related. This makes it possible to search label files generated in this way for such phonetic processes as assimilations and elisions, as exemplified in the preceding paragraphs. But in spite of these great advantages, which the linear approach offers, it also has serious drawbacks:

- Segments may not always be discernible in the signal as sequential elements, but their reflexes may nevertheless still be present as componential modifications of remaining segment strings, referable to such processes as palatalization, velarization, nasalization etc. The examples discussed in the preceding section illustrate this.

- If only the segmental deletions are marked there is a loss of contrastive phonological information because the signal contains more relevant features than this kind of symbolization represents, but the strictly linear segmental approach with a phonemic orientation is not capable of capturing this distinctivity.

As examples of this non-segmental residue of deleted segments are particularly common in spontaneous speech, a theoretical solution has to be found to deal with them adequately in the labelling of dialogue data bases. But the solution cannot consist of abandoning the linear segmental approach altogether and adopting a non-linear componential one instead because this theoretical reversal would forego all the clear advantages the linear concept has. So instead of an 'either-or' an 'as-well-as' is needed: besides individual segmental building blocks in their own right there are reciprocal influences on their concatenation, which manifest themselves in long phonetic components, even if the segments have been deleted in fast and reduced speech.

This integration of non-linear concepts into a linear phonological frame results in complementary phonology [15], a theoretical approach that requires both linearity and non-linearity - segments and components - for an adequate phonological representation of speech. The two concepts are only different aspects of the same phonetic-linguistic phenomenon. The segmental part establishes the relationship between phonetic manifestations in connected speech and canonical phonemic representations in a lexicon; the componential part takes into account the phonetic processes that characterize the concatenation of segments, especially in spontaneous speech and in its more numerous and more extreme articulatory reductions. This frame of complementary phonology provides a more adequate link between the symbolic and signal levels of speech analysis, interfacing symbolic descriptions of speech phenomena with events at the level of acoustic signals, which must in turn be related to the underlying articulatory processes.

For a more thorough pursuit of questions of segmental reduction in different speaking styles much larger data bases are required for more sophisticated statistical evaluation than have been analysed so far. And they have to be processed in such a way that they can give detailed phonetic information at the symbolic level in close correspondence with the acoustic signal and its various analyses. The symbolic representation should be within the framework of complementary phonology and the data entered into a data bank that allows the efficient and quick retrieval of acoustic signal data, also for further signal processing, in relation to symbolic strings contained in the label files and derived variants lexica. The Kiel data bases are structured along these lines and being put to use for the study of articulatory reduction in different speaking styles. What we need are data banks of this type for many languages in order to put the analysis of what has been termed 'phrase-level phonology' on a broader and comparative as well as typological level.

Although we can gain a great deal of insight into gestural dynamics under different speaking style conditions from the acoustic record, there are clear limitations as to what it can tell us. It is therefore mandatory to supplement our labelled acoustic data bases by labelled articulatory ones, but before this can be done successfully a methodology for representative data collection and processing will have to be developed.

The important conclusion to be drawn from the study of speaking style phenomena is the realization that the phonemic switches contained in a series of reduced forms along the reduction scale from least to most do not capture the essentials of what goes on in articulatory modification. It is the phonetic perspective associated with the biological and social constraints in communicative systems of sound producing humans that provides the answer to the question as to why articulatory reduction works the way it does.

ACKNOWLEDGEMENT

Part of the work reported here was carried out with financial support from the German Ministry of Education, Science, Research and Technology (BMBF) under VERBMOBIL contract 01IV101M7. My special thanks go to KTH/Stockholm for making MIX and related software available.

REFERENCES

[1] Kohler, K.J., *Lexica of the Kiel PHONDAT Corpus. Read Speech*, vols I, II, AIPUK 27, 28, Kiel: IPDS.

[2] Pätzold, M., Simpson, A. (1994), *Das Kieler Szenario zur Terminabsprache*. VERBMOBIL Memo Nr. 58, Kiel: IPDS.

[3] Kohler, K.J., Pätzold, M., Simpson, A. (1994), *Handbuch zur Segmentation und Etikettierung von Spontansprache - 2.3*. VERBMOBIL Technisches Dokument Nr. 16, Kiel: IPDS.

[4] IPDS (1994), *CD-ROM#1: The Kiel Corpus of Read Speech*, vol. I, Kiel: IPDS.

[5] IPDS (1995), *CD-ROM#2: The Kiel Corpus of Spontaneous Speech*, vol.I, Kiel: IPDS.

[6] Kohler, K.J. (1990), "Segmental reduction in connected speech in German: phonological facts and phonetic explanations", in W.J. Hardcastle and A. Marchal (eds.), *Speech production and speech modelling*, pp. 69-92, Dordrecht/ Boston/London: Kluwer Academic Publishers.

[7] Browman, C.P. Goldstein, L. (1992), "Articulatory phonology: an overview", *Phonetica*, vol. 49, pp. 155-180.

[8] Kohler, K.J. (1994), "Glottal stops and glottalization in German. Data and theory of connected speech processes." *Phonetica*, vol. 51, pp. 38-51.

[9] Kröger, B.J. (1993), "A gestural production model and its application to reduction in German", *Phonetica*, vol. 50, pp. 213-233.

[10] Kohler, K.J. (1991), "The phonetics/phonology issue in the study of articulatory reduction", *Phonetica*, vol. 48, pp. 180-192.

[11] Lindblom, B. (1990), "Explaining phonetic variation: a sketch of the H & H theory, in W.J. Hardcastle and A. Marchal (eds.), *Speech production and speech modelling*,pp.403-439,Dordrecht/ Boston/London: Kluwer Academic Publishers.

[12] Kohler, K.J. (1979), "Kommunikative Aspekte satzphonetischer Prozesse im Deutschen", in H. Vater (ed.), *Phonologische Probleme des Deutschen, Studien zur deutschen Grammatik 10*, pp. 13-39, Tübingen: Gunter Narr.

[13] Carlson, R., Granström, B., Hunnicut, S. (1990), "Multilingual text-to-speech development and applications", in W.A. Ainsworth (ed), *Advances in Speech, Hearing, and Language Processing*, pp. 269-296, London: JAI Press.

[14] Kohler, K.J., Lex, G., Pätzold, M., Scheffers, M., Simpson, A., Thon, W. (1994), *Handbuch zur Datenaufnahme und Transliteration in TP 14 von VERBMOBIL - 3.0* VERBMOBIL Technisches Dokument Nr. 11, Kiel: IPDS.

[15] Kohler, K.J. (1994), "Complementary phonology: a theoretical frame for labelling an acoustic data base of dialogues" *Proc. ICSLP94*, vol. 1, pp. 427-430.

FROM READ SPEECH TO REAL SPEECH

W. N. Campbell

ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan.

ABSTRACT

This paper describes differences in speaking style between read and spontaneous speech from the viewpoint of synthesis research and discusses the development of a set of labels to encode prosodic and segmental variation. Spontaneous speech confronts us with phenomena that were not encountered in corpora of prepared or read speech, and to label them we are increasingly having to identify higher-level units of discourse structure and speaker involvement.

INTRODUCTION

The relationship between speech synthesis and phonetic research is an old one, but we have yet to hear synthetic speech that sounds natural. Some isolated vowels and consonants can be very well replicated and, with careful hand-tuning, even whole utterances can be mimicked, but I am aware of no speech-synthesis-by-rule system that I could yet mistake for a human speaking. Perhaps the best rule-generated synthetic speech that can be heard today is concatenative, using small segments from recorded sequences of real speech and joining them to form novel utterances, but even then the resulting speech loses much of its original naturalness.

I maintain that the reasons for this loss of naturalness are two-fold: a) that typically only a limited number of speech tokens are used to generate a variety of speech, so degradation results from the signal processing required to put them together and modify their prosody, and b) from constraints in the recording of the original speech sequences themselves. Almost all sequences for concatenation are taken from recordings of carefully read speech, and although they may be *phonemically representative*, they are *prosodically constrained* and invariant. Thus, what they encode perhaps models the relevant configurations of the vocal tract for a given speech sequence but fails to adequately model the dynamic articulatory characteristics of the speech.

This paper, focuses not on the *modelling* of speech but on its *characterisation* (or labelling) instead. Using examples illustrating durational characteristics, it shows some effects of the differences in speaking style, and attempts to address the phonetics of spontaneous speech. In doing so, it describes the development of a small set of features sufficient for the description of natural speech such that its variety can be emulated in a synthesis system.

Natural speech

Scientific analysis requires controls, but as Barry has pointed out [1], in the acquisition of speech recordings, these are too often controls on production, with not enough concern for communicative effect. In its natural form, speech is inter-personal and often functionally goal-directed, but in recordings of lab speech (or of speech units for synthesis), where the listener is replaced by a microphone, the speech becomes production-based rather than listener-oriented. As a consequence, the materials we collect and analyse may not be representative of what people do when they speak normally.

For the analysis of natural speech, it is necessary to replace production controls with statistical controls, and use these to study instead large representative corpora of spontaneously produced spoken material. Such corpora are now becoming widely available but the tools for their analysis were developed for a more restricted speaking style. To cope with large volumes of speech, the processing must be automatic, requiring a minimum of manual intervention.

LABELLING SPEECH

Kohler [2] (see also Coleman [3]) has shown that although the articulation of a given sequence of words can vary considerably under different speaking styles according to a cognitively-based reduction coefficient that is dependent on speech act type, a linear segmental representation of canonical citation forms can account well for such phonological reorganisation of speech. He shows that although a segment may be elided or deleted in the production of fluent speech, a non-segmental residue remains to colour the articulation of the remaining segments. Such a canonical representation is easily accessible from a machine-readable pronunciation dictionary. Thus, given the orthographic transcription of a speech corpus, segmental labelling can be automated to a large extent by using speech recognition technology.

Segmental labelling

By training hidden Markov models (HMMs) corresponding to the phonetic labels in a machine-readable pronunciation dictionary, and generating networks of possible pronunciations for each word, we can use Baum-Welsh re-estimation [4] to model the HMMs closely on each corpus, using the orthographic transcriptions to constrain the alignments, and thereby achieve segmentation accuracy comparable to human transcription [5]. Separate lexical sub-entries are included for some particularly different pronunciation variants such as 'gonna' for 'going to'.

What can be predicted does not need to be labelled. Since the articulation varies according to speaking style, it is sufficient to model the speaking style (or its prosodic correlates) in order to be able to predict the reduction coefficient. The canonical segmental labels allow access to phone-sized portions of the speech waveform from which we can extract prosodic information in order to account for the finer articulatory differences and thus enable us to encode phonation-style characteristics without the need for marking them explicitly.

Materials

The materials referred to in this paper come from four corpora. The first contains readings of 5000 citation-form English words, a subset of these words read one at a time in the form of 200 meaningful sentences, the same sentences read continuously, and 20 minutes of spontaneous interactive monologue (*i.e.*, dialogue with a passive partner). These are in British English from a young adult female speaker, and represent an extreme range of production variation.

The second corpus contains forty-five minutes of professionally read American radio-news speech [6]. It comes from one speaker and exhibits a consistent marked style of production typical of professional announcer speech.

The third, consisting of 300 focus-shifting sentences, produced by an American speaker, illustrates contrastive focus. A set of sentences were produced in three utterance styles: a) read in grouped order by set, b) read in randomised order, and c) produced spontaneously by elicitation in interactive discourse. Each set of sentences contained syntactically and semantically identical word-sequences that differed only in the emphasis given to each word in different renditions. Shifts of emphasis in the read speech were controlled by use of capitalisation to signal different interpretations and, in the interactive discourse, by (deliberate) misinterpretations on the listener's part.

Finally, from a speaker of American English, is one side of a series of twenty task-related dialogues, performed in a multi-modal environment, alternatively with and without a view of the interlocutor's face [7]. These allow us to compare the speech of one individual, in a highly restricted domain, under a variety of interaction styles.

These corpora were variously labelled at different sites using different transcription conventions, and to compare them it was necessary to relabel all to a uniform style. To do this economically requires definition of a small set of labels that suffice for the complete characterisation of their perceptually salient characteristics. Needless to say, this work is ongoing.

Prosodic labelling

Traditionally, speech has been labelled separately for prosodic and segmental characteristics, but while these features are independently variable, the interaction between them is strong. Segments vary consistently in relation to the prosodic environment so that this segmental variation, in conjunction with the prosodic variation, plays a functional role in chunking the speech and signalling prominence relationships. In read speech at least, boundaries and prominences appear to be the most basic elements of prosodic structure. In locating segments relative to these two dimensions, we can predict much about their articulation.

For example, a given speech segment immediately before a prosodic phrase boundary is likely to be very different from an equivalent one immediately after; it may be considerably lengthened, its amplitude low and decaying, and it may exhibit vocal fry. The segment will also be lengthened in a nuclear accented syllable, but there will be a different profile of lengthening [8] [9] and also increases in spectral tilt resulting from changes in vocal effort [10] [11] [12] [13] and in supraglottal phonation arising from local hyperarticulation [6] [14].

The BU Radio News corpus [6] has been prosodically labelled by hand according to the ToBI conventions [16] to differentiate high and low tones at intonational boundaries and on prominent syllables, and to mark the degree of prosodic discontinuity at junctions between each pair of words. Campbell & Black [17] used this corpus as the basis for a resynthesis test of the assumption that labels of prosodic and canonical segmental context suffice to encode the lower-level spectral and articulation characteristics.

The test was done by iteratively removing sentences from the radio-news corpus and resynthesising them by concatenation of similar segment sequences selected from the remaining utterances according to suitability of their prosodic environment. This test confirmed that much of the spectral variation in the segments was adequately coded [17]. In one salient example, a sequence of segments across a prosodic phrase boundary were resynthesised using tokens selected from pre- and post-pausal locations such that the 'silence' between them also included an appropriate sharp intake of breath, which made the resulting synthesis sound even more 'natural'[1]. When equivalent tokens from the same segmental sequences were selected from less appropriate prosodic environments, the resulting synthetic speech showed considerable degradation.

Using the radio corpus as training data, Wightman & Campbell [18] defined a set of acoustic, lexical, and segmental features derivable from the phone labels, the dictionary, and the speech waveform, that achieved automatic detection of 86% of hand-labelled prominences, 83% of intonation boundaries, and 88% correct estimation of break indices (at ±1). This was trained using a hybrid combination of a tree quantiser with Viterbi post-processing to maximise the output likelihoods, operating directly on the aligner output.

The acoustic features extracted from the speech waveform for the autolabelling of prosody include (in order of predictive strength) silence duration, duration of the syllable rhyme, the maximum pitch target[2], the mean pitch of the word, intensity at the fundamental, and spectral tilt (using a harmonic ratio). Non-acoustic features included end-of-word status, polysyllabicity, lexical stress potential, position of the syllable in the word, and word-class (function or content only). These latter are all derivable directly from the dictionary used in the aligning.

[1] It should be mentioned here though that because of the limited size of this source database, simple concatenation of these selected units produces noisy synthetic speech, and some (distorting) signal processing would still be required to reduce discontinuities between the selected units.

[2] Pitch targets were calculated using Daniel Hirst's quadratic spline smoothing to estimate the underlying contour from the actual f0.[19]

SPONTANEOUS SPEECH

As an illustration of the contrasts between read and spontaneous speech in British English, we can examine durational structuring, as shown in figures 1 – 4, which plot mean segmental duration against the coefficient of variance (*i.e.,* the standard deviation expressed relative to the mean) for each phone class.

We can see from Figures 1 and 2 that in the isolated-word citation-form readings, there is a good dispersion in the mean durations for each phone class, and relatively constant variance in their durations. Figure 3 shows the opposite to be the case for the same sequence of words in read sentences. Here the variance increases and there is considerable shortening so that segments are no longer so distinct. Separate examination of segments in word-initial and word-medial position confirmed that this is not just a result of more phrase-final lengthening (isolated words being complete phrases).

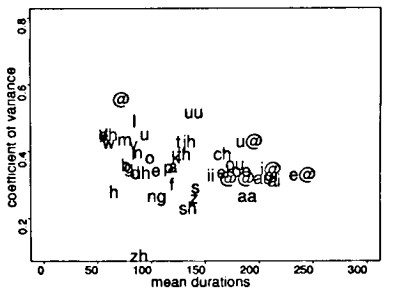


Figure 1: Citation-form words

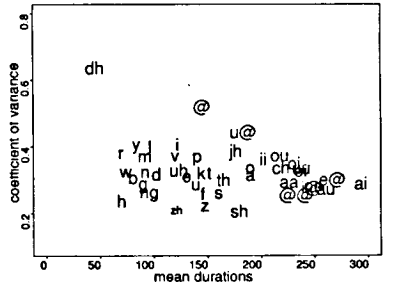


Figure 2: Isolated-word sentences

Rather, the articultion of the citation-form words was generally slower and more distinct.

When the speech contains little contextual information, and the speaker is concerned to be clearly understood, then segmental durations are maximally separated, exaggerating the difference between the phone types, but as the style becomes more natural and the listener can rely on prosodic phrasing to aid in the interpretation of the speech, then we find more variance in the durations and less distinction between their means; all words tend to be shorter and more varied than in the citation-form readings.

The spontaneous monologue from the same speaker, in Figure 4, shows the same trends exaggerated. we find not only that the mean durations for all segment types are low and uniform,

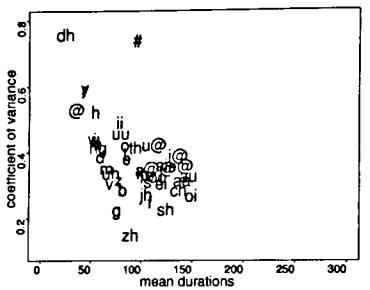


Figure 3: Continuous sentences

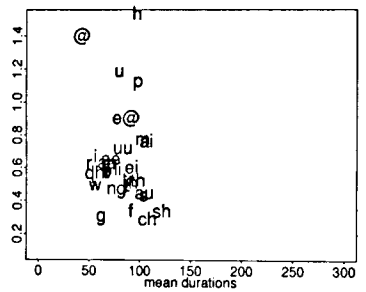


Figure 4: Spontaneous speech

but that the variances are huge - note too that the vertical scale of Figure 4 is *twice* that of the other three figures. When the listener is present, the speaker has a better understanding of the extent of their mutual information, and can rush through some parts of the speech, and linger before others. There is much greater range of variance in the natural speech than was found in the more carefully prepared 'lab' speech.

Spectral correlates

To show that this difference of style is not unique to durations, nor specific to one speaker, we can compare changes in spectral tilt associated with prominence and focus in read *vs.* interactive speech.

A study of prominence detection, described more fully in [15], showed that speakers change their phonation according to the discourse context and the type of information they impart. The detection algorithm used both duration and spectral tilt[3]. The corpus varied prominence in two ways: (1) by asking the speaker to "emphasise" capitalised words in reading, and (2) by eliciting emphatic corrections of feigned misinterpretations.

Table 1: Prominence detection

	A	B	C
dur'n & energy:	92%	78%	72%

Showing percentage correct detection of prominence using normalised measures of duration and energy.
Key: A: read grouped, B: read in randomised order, C: interactive

Using normalised duration and energy to detect prominences achieved 92% recognition in the clearly read speech, but only 72% in the interactive dialogue (Table 1.). The elicited corrections resulted in a clearer articulation, but durational organisation was more varied, and prominence was not easy to detect automatically using this alone as a cue.

It is interesting to note that although the durational cues to prominence were weakened by greater variance in the interactive speech, the spectral measure was apparently strengthened, as Table 2 shows. We can suppose that this trade-off is not coincidental, and that

Table 2: ± prominent spectral tilt

	student's t	df
read grouped	35.63	7676
read randomised	19.01	6110
interactive	42.76	6974

Showing the separation in mean spectral tilt between prominent and non-prominent syllable peaks.

the speaker varies her production according to the needs of the discourse context.

Interactive speech

Although much of the labelling of significant levels of information can be done to a large extent automatically for read speech, when extending these techniques to the dialogue corpus and to the spontaneous monologue, we realise the need for extra levels of information to describe the structuring of discourse events and to indicate the switches in speaking style [1] that affect the global prosodic characteristics. Whereas the read speech was highly predictable, the unplanned speech is characterised by bursts of faster and slower sections where the speaker varies her role, and much greater variation in f0 range and pausing as she expresses different degrees of confidence, hesitation, and uncertainty.

Hirschberg has pointed out [20] that the major differences between lab speech and spontaneous speech appear to be prosodic (concerning speaking rate and choice of intonation contour), but there are also significant segmental differences. She notes for example that some disfluencies in spontaneous speech are marked by characteristic phonetic effects, such as *interruption glottalisation*, which is acoustically distinct from articulatorily similar laryngealisation. In labelling to include these characteristics, we need to have not just the prosodic and segmental information derived from HMM alignment, but also an indication of fluency and commitment. That is, we need to label not just what the speaker says but what she is doing in saying it, and how she feels about what she is saying.

Transcribing the orthography of spontaneous speech requires more than the skills of an audio-typist, and decisions have to be made about marking disfluencies and repairs. To include this information in our labelling, two extra levels of information were added to the ToBI set. One after Nakatani and Shriberg [21], extended the miscellaneous tier of the ToBI transcription to bracket interruptions in the speech flow, and one after Stenström [22]), added IFT (illocutionary force type) speech-act labels. The following set was used:
inform, expressive, good wishes response, apology response, invite, vocative, suggest, instruct, promise, good wishes, yn question, do you understand question, wh question, yes, no, permission request, acknowledge, thank, thanks response, alert, offer, offer follow up, action request, laugh, greet, farewell, apology, temporize, hesitation, confirmation, cough, breath, silence.

As an example, in the dialogue corpus the word 'okay' was said 140 times. It was variously labelled as 'acknowledge', 'confirmation', 'offer follow up', 'accept', and 'do you understand qn.', etc., twelve categories in all. The intonation, duration, and articulation varied considerably; sometimes short, sharp, and rising, on a high tone, sometimes slow and drawn out on a falling tone. Since we were able to find significant correlations between the intonation and the speech act label for most of these cases (see [23] for a discussion), we continue in our assumption that instead of trying to predict and model the lower-level acoustic variations, we should instead be accessing them through higher-level labels.

Spontaneous speech appears to be most marked in terms of its rhythmic structuring, exhibiting greater ranges of variation with corresponding differences in phonation style. These prosodic changes appear to have clear correlates in the speech-act labels that we are now using. However, to more fully describe them, we need also to formalise a measure of the speaker's commitment to her utterance. Impressionistic com-

[3] as measured by the relative amount of energy in the top third of an ERB-scaled spectrum (2kHz - 8kHz).

ments such as 'she's thinking ahead', 'her mind's not on what she's saying', 'she's said this many times before', and 'she doesn't quite know how to put this' are triggered by such differences in speaking style, but none of the labels we have considered so far are sufficient to mark such differences. The first step in this work is to determine the appropriate labels, then we can categorise their prosodic and articulatory correlates. Since human listeners can respond consistently to small speaking-style changes, then the clues must be present somewhere in a higher representation of the signal.

SUMMARY

To summarise the main points of this paper, I have argued that for the efficient characterisation of speech sounds (at least in the context of concatenative speech synthesis), it may not be necessary to label fine articulatory details, nor to attempt a numerical description of the prosodic attributes, but instead to use a higher-level specification of the environment in which they occur.

In read speech, knowing the triphone context of a segment, its position in the syllable, and whether that syllable is prominent, prosodic-phrase-final, or both, allows us to predict much about its lengthening characteristics, its energy profile, its manner of phonation, and whether it will elide, assimilate, or remain robust. Thus for adequate characterisation of speech it is not necessary to label the fine phonetic features explicitly since the higher-level description suffices to include them implicitly.

In the case of *real* speech, however, a significant part of the message lies in the *interpretation* of *how* it was said, and to encode that level of information, we need to incorporate labels for discourse and communication strategy; we need to estimate the state of mind of the speaker, her commitment to the utterance, and the role of that utterance in a greater discourse. This is future work.

Finally, to return to synthesis, if we consider a hugely finite corpus of natural speech as a source of units for concatenative synthesis then, instead of disruptively warping a segment to fit a predicted context, it would be possible to select an appropriate segment from amongst the available variants. Furthermore, if that corpus were adequately labelled in terms of all the contributing factors (i.e., with phonemic, phrasal, prosodic, speech-act etc., labels), then it would no longer even be necessary to predict fine details of the speech at all; it would be enough to select a segment with the same labels to characterise the desired target speech. The durations and other relevant acoustic features would be contextually appropriate and natural by default.

The remaining challenge is to label large corpora of real speech according to a small and sufficiently descriptive set of features so that all the relevant variations can be indexed and retrieved. This reduces to a definition of the *perceptually salient* characteristics of speech, which in turn enables us to use only a large speech corpus instead of a huge one without loss of naturalness.

REFERENCES

[1] Barry, W. J., (1995) "Phonetics and phonology in speaking styles". In *Symposium on speaking styles, Proc ICPhS 95*, Stockholm, Sweden.

[2] Kohler, K, (1995) "Articulatory reduction in different speaking styles". In *Symposium on speaking styles, Proc ICPhS 95*, Stockholm, Sweden.

[3] Coleman, J. C., (1992) "The phonetic interpretation of headed phonological structures containing overlapping constituents". *Phonetics Yearbook 9*, pp 1-44.

[4] Entropic Research Laboratory, Inc, (1993) *HTK - Hidden Markov Model Toolkit* 600 Pennsylvania Avenue, Washington DC 20003.

[5] Talkin, D., & Wightman, C. W., (1994) "The Aligner: text-to-speech alignment using Markov models and a pronunciation dictionary". In *Proc. ESCA Workshop on Speech Synthesis*, Mohonk, NY. pp 89-92.

[6] Ostendorf, M., Price, P., & Shattuck-Hufnagel, S., (1995) *The Boston University Radio News Corpus*, Report No BCS - 95 001.

[7] Fais, L., (1994) "Conversation as collaboration: some syntactic evidence", *Speech Communication 15*, pp 230-242.

[8] de Jong, K., (1995) "The supraglottal articulation of prominence in English: linguistic stress as localised hyper-articulation". in *Journal of the Acoustical Society of America 97(1)*, pp 491-504.

[9] Campbell, W.N. (1993) 'Predicting segmental durations for accommodation within a syllable-level timing framework', *Proc Eurospeech-93,*, Berlin, Germany pp 1081-1084.

[10] Pierrehumbert, J. & Talkin, D. (1992) "Lenition of /h/ and glottal stop". In **Papers in Laboratory Phonology II,** eds. G. J. Docherty & D. R. Ladd, Cambridge University Press.

[11] Gauffin, J. & Sundberg, J. (1989) "Spectral correlates of glottal voice source waveform characteristics", **JSHR 32**, pp 556-565.

[12] Sluijter, A., & van Heuven, V. J., (1993) "Perceptual cues of linguistic stress: intensity revisited", *Proc. ESCA workshop on Prosody,*,Lund University, Sweden. pp 246-249,

[13] Campbell, W. N., & Beckman, M. (1995) "Stress, Loudness, and Spectral Tilt", *Proc Acoustical Soc. Japan*, Spring meeting, 3-4-3.

[14] Lindblom, B. E. F. (1990) "Explaining phonetic variation: A sketch of the H&H theory". **Speech Production and Speech Modelling** edited by H. J. Hardcastle and A. Marchal (Kluwer, Dordrecht), pp 403-409.

[15] Campbell, W.N. (1995) "Loudness, spectral tilt, and perceived prominence in dialogues", In *Proc ICPhS 95*, Stockholm, Sweden.

[16] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J., (1992) "ToBI: a standard for labelling English prosody". In *Proceedings of ICSLP92*, volume 2, pp 867–870.

[17] Campbell, W. N., & Black, A. W., (1994) "Prosody and the selection of source units for concatenative synthesis". In *Proc. ESCA Workshop on Speech Synthesis*, Mohonk, NY.

[18] Wightman, C., W., & Campbell, W., N., (1995) "Improved labelling of prosodic structures", IEEE Trans. Sp. & Audio, submitted.

[19] Hirst, D., (1980) "Automatic modelling of fundamental frequency using a quadratic spline function" In *Travaux de l'Institut de Phonétique 15*, Aix en Provence, pp 71-85.

[20] Hirschberg J., (1995) "Acoustic and prosodic cues to speaking style in spontaneous and read speech". In *Symposium on speaking styles, Proc ICPhS*, Stockholm, Sweden.

[21] Nakatani, C., & Shriberg, L., (1993) "Draft proposal for labelling disfluencies in ToBI". paper presented at 3rd ToBI labelling workshop, Ohio.

[22] Stenström, A., (1994) *An Introduction to Spoken Interaction.* Longman, London.

[23] Black, A. W., & Campbell, W. N., (1995) "Predicting the intonation of discourse segments from examples in dialogue speech". In *Proc. ESCA Workshop on Spoken Dialogue*, Hanstholm, Denmark.

MODELLING SWEDISH INTONATION FOR READ AND SPONTANEOUS SPEECH

Gösta Bruce
Department of Linguistics and Phonetics, Lund, Sweden

ABSTRACT

Our approach to the study of dialogue prosody involves analysis of dialogue structure, prosodic analysis (auditory and acoustic-phonetic) and synthesis. A comparison of spontaneous and read speech shows variation in global pitch level and range to reflect topic structure more stereotypically in read speech, and variation in local pitch range on a focussed item to be a signal of feedback seeking and interaction in spontaneous speech.

INTRODUCTION

The focus of attention of our current prosody research is the prosody of spontaneous speech and dialogue. The framework for this prosody research is the project *Prosodic segmentation and structuring of dialogue* centered around the description of Swedish [1]. This project is a cooperation between phonetics at Lund and speech communication at KTH, Stockholm with support within the second phase of the Swedish Language Technology Programme 1993 - 1996.

The ultimate goal of the project is to create a more powerful prosody model, in particular for the description of intonational patterns of spontaneous speech and dialogue. In order to achieve this goal we will have to increase our understanding of the structuring of spontaneous dialogue, particularly how intonation contributes to the development of such a dialogue.

Background

Our current prosody model is based on two decades of prosody research within the project group with experience mainly from read, laboratory speech. Thus our modelling of intonation in Swedish is typically based on the analysis of speech material elicited from informants who have been asked to perform specific tasks in a phonetics laboratory environment.

The intention behind this laboratory setting has not been the study of the prosody of reading or even less so elocution as a particular speaking style. Instead the idea has been to create a situation simulating natural speech to a reasonable extent while taking advantage of the laboratory setting with its high degree of experimental control and the possibility of varying test parameters one at a time.

The drawback is the somewhat artificial situation with albeit phonetically well balanced speech material (particularly in terms of microprosody) but often semantically unusual utterances and pragmatically strange situations. These constraints will put fairly heavy demands on the informants and their acting skills. Even if the informant turns out to be a good actor, the recorded utterance will clearly be an instance of prepared, read speech and will probably be recognized as such, even if the more or less pronounced non-sense character of the utterance is disregarded and only its prosody is considered.

A case in point is our way of eliciting phonetic prominence (focus) in different positions of a test utterance. The method exploited is a question - answer paradigm, where a variable context utterance (Question) is used to elicit a particular word in focus (or a whole phrase) of a fixed response utterance (Answer) [2].

Example
Question: Vad vill man lämna för några nunnor? (What nuns do they want to leave?)
Answer: Man vill lämna några LÅNGA nunnor. (They want to leave some TALL nuns.)

By the use of sonorant consonants and vowels of approximately the same degree of opening, differences in F0 that may be due to microprosodic variation are neutralized. Such a test paradigm presents a kind of simulation of a small portion of a very simple dialogue, where the informant has been instructed to perform both the roles of asking and answering.

Until fairly recently there have been relatively few phonetic studies of prosody in spontaneous speech and dialogue, i.e. the kind of situation where prosody has its proper function and usage. The reason for this state of affairs is to be looked for in the relative complexity of prosody. Spontaneous speech and dialogue offer such a richness of prosodic variation that its study seems to require a basic understanding of prosody in the more controlled context of laboratory speech.

THE PROSODY MODEL

The fundamental question addressed in my paper is: what kind of differences in terms of intonation and pitch patterns are typically found when we compare natural, spontaneous speech with prepared, read speech of the type referred to above?

In order to be able to find an adequate answer to this question, we will have to understand what counts as a difference, i.e. we will have to devise some kind of measure of variability. An instrument that may help us to this effect is our prosody model which has been implemented in a text-to-speech system for the generation of synthetic F0 contours and timing patterns [3]. Thus one reference for our current research effort within dialogue prosody and spontaneous speech is the Swedish prosody model, based on experience from prepared speech in the phonetics laboratory simulating natural speech as described above.

A question related to the one raised above will then be: how suitable is the intonation model which has been built upon our knowledge about prepared, laboratory speech also for the description of natural, spontaneous speech, and further how could the model be accommodated and elaborated to encompass the pitch patterns of spontaneous dialogue?

A fundamental assumption behind our modelling of prosody is that prosody can express a number of different, communicative functions, although the relationship between a certain function and its phonetic expression is typically indirect and quite complex.

Our modelling of the intonational structuring has been particularly concerned with the basic functions: grouping (signalling of coherence and boundary) into prosodic words, compounds, phrases, utterances and paragraphs, and prominence (foregrounding and backgrounding) of words and phrases, as well as their interaction [3]. A basic component of the intonational model is the tonal inventory in terms of tonal turning points (H, L, their combinations and diacritic symbols for alignment with prominences and boundaries) used as a phonological / abstract phonetic notation for both prominence relations and grouping. The input string to the model is then a tonal transcription (symbolic notation) on which the phonetic implementation rules operate to create the output F0 contour.

RESEARCH METHODOLOGY

One fundamental distinction between prepared, read speech and unrehearsed, spontaneous speech is the amount of planning involved. The on-line planning of spontaneous as opposed to the completely preplanned, read speech will typically result in well known prosodic differences such as the number and distribution of pauses, more variation in speech tempo, voice quality and voice intensity as well as repetitions, false starts and corrections characteristic of spontaneous speech. What effects this difference in planning has specifically on intonation and choice of pitch patterns is less clear, however.

A true spontaneous dialogue is often described as an act of negotiation between the two (or more) interlocutors. Important aspects of the structure of the dialogue are at least the following: textual aspects [topic structure, semantic focus], initiative / response structure, feedback seeking [are you with me?, do you see what I mean?] and giving ['mm', 'yeah'] as well as turn regulating [keeping, yielding, taking and struggling for the turn]. Our analysis of the structure of a dialogue attempts to take these aspects

into account as well as other aspects like signalling of attitudes and rhetoric activity, which seem to be represented in all kinds of speech to a varying degree.

An important starting point for our work is to initially regard dialogue and prosody as independent. This means that we assume that it is convenient at first to make an analysis of the structuring of dialogue and a corresponding analysis of prosodic categories. Only then is a coupling made between the prosodic analysis and the dialogue analysis which enables the establish-ment of possible, interesting inter-connections and correlations. Therefore, we do not a priori assume that there would be, for example, a special question intonation obligatorily used by a speaker taking a strong initiative in a conversation, or that the introduction of a new conversation topic necessarily needs to be signalled prosodically.

Our prosodic analysis is divided into an auditory analysis in the form of a prosodic transcription and an acoustic-phonetic analysis. The prosodic transcription relevant here is a broad, phonetic analysis containing symbolization of both prominence and grouping. It consists of two parts. The first part is the tonal tier for the notation of two levels of prominence (accented, focussed), including the lexically determined distinction between the two word accents in Swedish, as well as junctures (initial and terminal boundary tones). The second part is the grouping tier with two levels of phrasing (minor and major phrase, corresponding to prosodic phrase and prosodic utterance respectively) being symbolized. Our transcription is reminiscient of the ToBI transcription system [4], but unlike ToBI we rely exclusively on an auditory analysis.

The acoustic-phonetic analysis is based on F0 and waveform information, whereby both global features (in terms of, for example, F0 level and F0 range) and local features (in terms of direction and timing of F0 events) are taken into consideration and interpreted in our current prosody model.

tonal structure

accented	accent I (HL*)	
	accent II (H*L)	
focussed	accent I ([H]L*H)	
	accent II (H*LH)	
	compound (H*L...L*H)	
juncture	initial (%L; %H)	
	terminal (L%; LH%)	

grouping

boundary	minor	|
	major	‖

The three types of analysis - analysis of dialogue structure, auditory analysis, acoustic-phonetic analysis - involving both symbol and signal information are combined and synchronized with each other in the same ESPS/Waves+ environment. The labelling used (symbol information) consists of an orthographic tier (marking the end of words), a tonal tier (symbols of tonal structure), a boundary tier (symbols of grouping), dialogue structure tier (hierachical topic structure), and a miscellaneous tier (with extralinguistic and other information).

An important part of our research methodology is the use of speech synthesis. In addition to the text-to-speech synthesis as described above, one method currently being developed is the implementation of our intonation model in the ESPS/Waves+ environment which will be used as an analysis-by-synthesis tool. The input is the prosodic transcription with information about type and time location of tonal turning points. This information (with few segmentation marks) together with phonetic rules according to our intonation model are fed into a modifed version of the ESPS / Waves+ synthesizer. The synthesis module will be exploited both to verify the prosodic transcription and to develop the prosody model itself.

COMPARING INTONATION IN READ/SPONTANEOUS SPEECH

In the present paper we do not intend to give any conclusive answer to the question about typical differences between read and spontaneous speech. Instead we will try to come up with a few hypotheses about the specific, interactive contribution of intonation, features that would be typically lacking in prepared, non-interactive speech, as well as features typical of the intonation of both read and spontaneous speech.

Experimental design

One way of tackling this question experimentally is to make a direct comparison between the original, spontaneous version of a section of a dialogue and a corresponding, read version of the same portion of speech (cf. for example an early study by Gårding [5] and a recent study by Ayers [6]).

In our material, the original dialogue is a friendly conversation between two adult speakers, a daughter and her mother, who were talking spontaneously for around 13 minutes about partly predetermined topics. The read version is an acted dialogue involving the same speakers who took part in the spontaneous conversation. In the read version, they simulated their own, original dialogue from scripts. The recording of the acted version took place a few weeks after the original recording. The section from the dialogue which we have chosen for specific comparison and illustration in the present paper and which appears particularly interesting from a dialogue prosody point of view lasts for about a minute. In this section a transition from one conversational topic to another takes place (the very end of the topic 'the trip to Berlin' and the initial part of 'the blouse'), where the daughter takes the initiative and is thedominating speaker in the initial, mainly monologue part of this topic (Figure 1).

D: ja fast vi hinner inte så mycket	D: yes but we can't do all that much
M: nej	M: no
D: på	D: in
M: nej	M: no
D: en och en halv dag	D: a day and a half
M: Tiergarten och	M: Tiergarten and
D: men det ska bli jättespännande i alla fall	D: but it'll be tremendously exciting anyway
M: mm	M: mm
D: mm	D: mm
M: ja	M: yeah
D: ska jag berätta om min om den dära blusen som jag tänkte jag skulle sy av det där rutiga tyget	D: shall I tell you about my about that blouse that I thought I'd sew out of that checkered material
M: mm	M: mm
D: mm det är ju så jätterutigt så jag tror det blir jättekonstigt om man bara syr en vanlig skjorta	D: mm it's really so very checkered so I thought it would be really strange if I just sewed an ordinary shirt
eller ja det blir inte jättekonstigt men det är så fint tyg så det är synd om man inte gör nåt av det då	or well it wouldn't be really strange but it's such nice material so it would be a shame if I didn't do something with it
så har jag tänkt eh att jag skulle köpa ett vitt tyg till och ha dubbla kragar	so I thought ah that I would buy some white material and have a double collar
M: mm	M: mm
D: eh så jag ska ha en jag skulle gärna vilja ha en v-ringning och sen så eh lite snibbig krage så här	D: ah so I'll have a I'd really like to have a v-neck and then ah sort of pointed collar like this

Figure 1. Extract from spontaneous dialogue: friendly conversation between mother [M] and daughter [D]; Swedish original to the left and English translation to the right.

Tentative findings and discussion

The two versions of the dialogue section are, not surprisingly, auditively clearly distinct. The prepared, non-interactive but coherent character of the acted version of the dialogue as opposed to the characteristic interactive, on-line planning of the spontaneous version is quite striking. While we can assume that this impression is at least partly due to differences in pausing, variation in speech tempo, voice intensity, and voice quality as well as in the degree of reduction / elaboration, our specific task here is to try to isolate the contribution of pitch patterns and intonation.

The prosodic transcription, which is the broad, auditory analysis described above (basically a phonological analysis of accentuation and phrasing) displays apparent similarities between the two versions investigated. We can observe some differences in accentuation and focus locations as well as in phrasing between the versions. Some of these differences are clearly 'accidental', while others are probably more stable differences between speaking styles. In our search for regularities we will have to neutralize for differences between the versions in for example focus placement and phrasing that are clearly optional and not dependent on the specific speaking style.

Judging from the prosodic transcription, the most stable difference between the versions appears to be in phrasing. In the read version there is a tendency for a phrase to accomodate more words than in the spontaneous version as signalled by pitch and other cues. This may be thought of as due to the difference in planning between the speaking styles. The chunking into smaller units characteristic of the spontaneous speech is likely to be a reflection of the on-line planning. It is clear, though, that the broad, auditory categories used in our prosodic transcription do not reflect the apparent difference in intonation between the two speaking styles.

In our phonetic analysis of prosody we will concentrate on how the pitch patterns of the dialogue section may reflect two potentially important aspects of dialogue structure, namely the textual aspects (in particular topic structure) and the feedback dimension.

A number of studies have shown how variation in global F0 range reflects the hierarchical organization of a discourse and the segmentation into topics or subtopics, for example [7], [8], [9], [10], [11], [12]. An expansion of F0 range at the beginning and a compression of F0 range values towards the end of a text unit is typical. This global downtrend over a text unit has been modelled, by way of extrapolating from similar phenomena occurring over the course of a single utterance, as declination [10], downstep [13] or initial raising / final lowering [11].

The study by Grønnum-Thorsen [10] is particularly instructive. In text units containing four utterances the first utterance has higher F0 values and the last utterance has lower F0 values, while the two medial utterances tend to cluster around the same, intermediate F0 values. The generalization may be that the beginning of a new topic / speech paragraph is signalled by pitch raising and the end by pitch lowering, while the ongoing speech in between has no particular textual signalling by pitch.

Figure 2 shows F0 contours of the transitional phase between the major topics ('the trip to Berlin' and 'the blouse') for the read and spontaneous versions of the dialogue, including two utterances before and one utterance (consisting of two prosodic phrases) after the topic shift. In the acted, read version the major topic shift is clearly signalled by means of F0. A marked shift from a fairly low F0 level and compressed F0 range to a high F0 level (increase by half an octave) and wide F0 range at the discourse boundary ties in with cited and expected relations. The decrease in F0 level and F0 range from the first to the second phrase of the utterance beginning the new topic is also apparent.

Figure 3 shows another example utterance containing two prosodic phrases representing the beginning of a subtopic somewhat later in the dialogue. Also here the read version displays

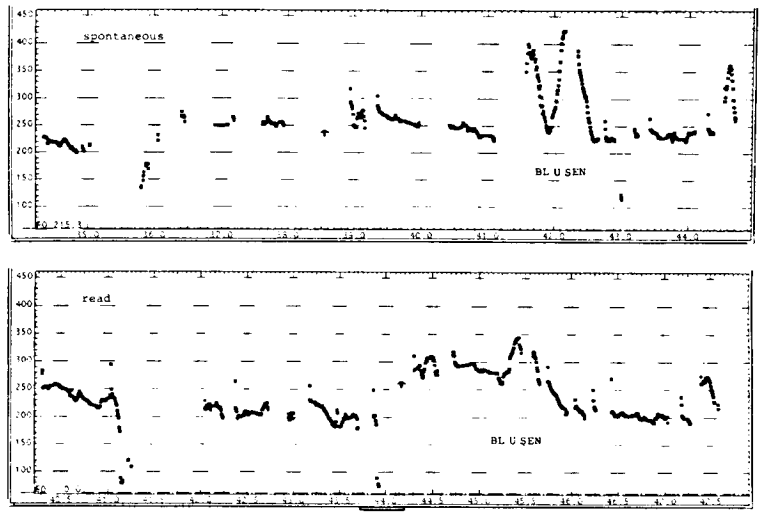


Figure 2. The effect of major topic shift. F0 contours of the same utterances from the spontaneous version (upper part) and the read version (lower part) by speaker [D]. The arrow indicates the time location of the topic shift. Focal word is in capital letters.

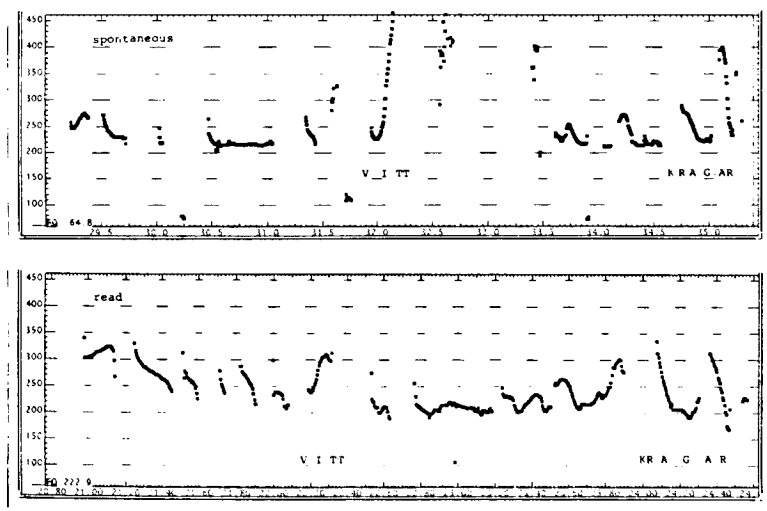


Figure 3. The effect of interactivity. F0 contours of the same utterance from the spontaneous version (upper part) and the read version (lower part) by speaker [D]. Focal words are in capital letters.

higher F0 values of the first phrase as compared with the second phrase of the utterance, reflecting the further textual organization of a subtopic.

In the spontaneous version (see Figures 2 and 3) we also find signs of textual organization in the distribution of pitch patterns including the major topic shift. Also here it is primarily global F0 level and F0 range which seem responsible for this organization. But this signalling appears to be less marked and maybe less stereotypic than in the acted version. Similar findings are reported in a corresponding study of American English read and spontaneous speech by Ayers [6].

A case in point is the signalling of the major topic shift in the spontaneous version. Unlike the read, acted version, the first utterance of the new topic 'the blouse' begins on approximately the same F0 level as the final utterance of the old topic, albeit with a wider F0 range. The last utterance of the topic 'the trip to Berlin' is namely characterized by a clear increase in F0 level as compared with the immediately preceding utterance. It functions prosodically as a transition utterance, both rounding off the old topic and, as it were, anticipating the topic shift, and also as a turn-keeping signal. Another exemplification from spontaneous dialogue of an utterance constituting a transition between major conversation topics, which textually belongs to the old topic but prosodically is also clearly affiliated with the new topic, is found in [14].

The difference in interactivity between the original, spontaneous version and the acted, read version is reflected in the feedback dimension. The seeking of feedback by the speaker having the turn seems to be one characteristic of the spontaneous dialogue which is typically lacking in the acted, prepared dialogue.

In the spontaneous version as exemplified in Figures 2 and 3 there appears to be a particularly wide, local F0 range on the two focussed items in each Figure (affecting mainly the focal H), while the corresponding items in the acted, read version have a moderately wide F0 range. The difference in range between the read and spontaneous versions for the focussed items exemplified amounts to about half an octave.

One possibility is that the difference in F0 values is related to differences in interactivity, specifically in the feedback dimension. The concentration of feedback seeking to certain focussed items is evidenced by the fact that it is at these points in time that feedback is also given through the use of support items of the 'mm', 'yeah' type. The particularly wide F0 range on the focussed items thus seems to be a reflection of the speaker seeking feedback from the listener.

This is reminiscient of the high rising tone sequence (H H%) found for spontaneous as opposed to read speech in American English by Ayers [6], what Sacks and Schegloff call 'try marker', and what Clark and Schaefer term a 'trial constituent' [15].

It should be pointed out, however, that in the spontaneous speech studied the exploitation of an extra wide F0 range on focussed items appears to be quite variable for successive phrases and utterances, apparently depending on the speaker's need for feedback.

CONCLUSION

Our comparison of the two versions of the dialogue section examined here serves as an illustration of possible intonational differences between read and spontaneous speech. The most apparent differences from this comparison are summarized here.

Pitch appears to play a major role in the signalling of textual organization, topic structure and division into speech paragraphs. This is evidenced in the spontaneous and acted versions which both display variation in global pitch level and pitch range as a reflection of this organization. A possible difference between read and spontaneous speech may be that in read speech textual organization is more stereotypic and rigidly marked. The initial raising of pitch level and range of the first phrase / utterance, intermediate values in between, and the final lowering of the last phrase / utterance of a text unit may represent the reading stereotype.

A marked increase in local pitch range specifically on focussed items may serve as a means of seeking feedback from your interlocutor. This is a feature characteristic of spontaneous speech, while it appears to be absent in acted, read speech.

These and other features of intonation are currently being modelled in our prosody model and tested for their assumed significance as signals of read and spontaneous speech.

ACKNOWLEDGEMENTS

This work was carried out under a contract from the Swedish Language Technology Programme (HSFR-NUTEK). I want to acknowledge the cooperation with my colleagues in the research project 'Prosodic segmentation and structuring of dialogue', namely Björn Granström, Kjell Gustafson, Merle Horne, David House and Paul Touati. Gayle Ayers, Dept of Linguistics, Ohio State University was a guest researcher in Lund for several months during 1993 and 1994 and has contributed to the project. She prepared the recordings and the analysis of the dialogue investigated.

REFERENCES

[1] Bruce, G., B. Granström, K. Gustafson, D. House and P. Touati (1994), "Modelling Swedish prosody in a dialogue framework", *Proceedings of ICSLP 94*, pp. 1099-1102, Yokohama.

[2] Bruce, G. (1977), *Swedish word accents in sentence perspective*, Lund: Gleerups.

[3] Bruce, G. and B. Granström (1993), "Prosodic modelling in Swedish speech synthesis", *Speech Communication* 13, pp. 63-73.

[4] Silverman, K., M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg (1992), "ToBI: a standard for labelling English prosody", *Proceedings of ICSLP 92*, pp. 867-870, Edmonton.

[5] Gårding, E. (1967), "Prosodiska drag i spontant och uppläst tal", G. Holm (ed.), *Svenskt talspråk*, pp. 40-85, Uppsala: Almqvist & Wiksell.

[6] Ayers, G. (1994), "Discourse functions of pitch range in spntaneous and read speech", *OSU Working Papers in Linguistics*, Vol. 44, pp. 1-49.

[7] Lehiste, I. (1975), " The phonetic structure of paragraphs", A. Cohen and S. Nooteboom (eds.), *Structure and Process in Speech Perception*, pp. 195-206, Berlin: Springer-Verlag.

[8] Brown, G., K. Currie and J. Kenworthy (1980), *Questions of intonation*, London: Croom Helm.

[9] Bruce, G. (1982), "Textual aspects of prosody in Swedish", *Phonetica*, Vol. 39, pp. 274-287.

[10] Thorsen, N. (1985), "Intonation and text in Standard Danish", *JASA*, Vol. 77, pp. 1205-1216.

[11] Hirschberg, J. and J. Pierrehumbert (1986), "The intonational structuring of discourse", *Proceedings of the 24 Meeting of the Association of Computational Linguistics*, 136-144, New York.

[12] Silverman, K. (1987) *The Structure and Processing of Fundamental Frequency Contours*, Ph.D. Thesis, Cambridge UK: Cambridge University.

[13] Berg, R. van den, C. Gussenhoven and T. Rietveld (1992), "Downstep in Dutch: Implications for a model", G. Docherty and R. Ladd (eds.) *Papers in Laboratory Phonology II. Gesture, Segment, Prosody*, pp. 335-359, Cambridge University Press.

[14] Bruce, G, P. Touati, A. Botinis and U. Willstedt (1988), " Preliminary report from the KIPROS project", *Working Papers*, Vol. 33, 23-50, Lund University: Dept. of Linguistics and Phonetics.

[15] Clark, H.H. and E.F. Schaefer (1989), "Contributing to discourse", *Cognitive Science*, Vol. 13, pp. 259-294.

Prosodic and Other Acoustic Cues to Speaking Style in Spontaneous and Read Speech

Julia Hirschberg
AT&T Bell Laboratories

INTRODUCTION

Corpus-based approaches to the study of speaking styles seeks to identify observed differences as potentially perceptually salient. Particularly when reliable differences are found across speakers in such corpora, we hypothesize that such differences may in fact cause subjects to judge one style to be "spontaneous", or "conversational" and another to be "formal" or "read" or "laboratory speech". So the question of what distinguishes one style from another is addressed by examining similar corpora for systematic differences in lexical choice, syntactic constructions, and acoustic and prosodic phenomena. In this paper, results from three American English corpus studies of spontaneous and read speech are discussed, to examine the extent to which they support commonly held beliefs about which prosodic and other acoustic phenomena commonly distinguish between these speaking styles.

DIFFERENCES IN RATE AND CONTOUR

Two phenomena widely believed to distinguish spontaneous from read speech are speaking rate and choice of intonational contour. An examination of 395 read and spontaneous utterances in the ARPA *atis0* training corpus, collected at Texas Instruments for use in the DARPA Air Travel Information System spoken language system evaluation task from seventeen speakers (0), supports these hypotheses. These speakers interacted with a simulated voice response system to make air travel plans for several different hypothetical customers. They later returned to read transcriptions of their own speech.

A simple analysis of speaking rate in this corpus indicates considerable differences in rate for the read vs. spontaneously produced utterances. Of the seventeen speakers in this corpus, the average speaking rate (in syllables per second) was faster for read sentences than for spontaneous sentences in sixteen cases. Ratios of read to spontaneous speaking rate averages ranged from .93 (for the seventeenth speaker) to 1.57. Whether this difference is due primarily to the presence of pauses within spontaneous utterances or simply to an overall reduction in rate remains to be examined.

With respect to choice of contour as a differentiator between spontaneous and read speech, there are similar but less clear-cut findings. It has been claimed that yes-no questions are commonly uttered with rising intonation in spontaneous speech, but not so reliably in read speech. In a small study of yes-no questions in read speech (0), inverted yes-no questions and *wh*-questions from the training and test data of the speaker independent DARPA *Resource Management* (RM) database and from the TIMIT database were sampled for purposes of comparison with standard contours for those sentence types in spontaneous speech (0). While only 55% of yes-no questions in the RM database were uttered with rising intonation, over 80% of yes-no questions in the TIMIT1 sample rose. Production of *wh*-questions appears similar in both databases, with only 8% of the TIMIT1 *wh*-questions and 9% of the RM *wh*-questions uttered with final rise. However, this study examined only read speech, inferring the comparison with

spontaneous.

Comparing spontaneous with read utterances in the *atis0* corpus, we do find apparent differences between the two styles for a broader array of contours. There is a greater tendency in spontaneous speech to utter declaratives with some form of rising intonation; while fully 93.1% of read declaratives are uttered with final fall, only 70.5% of spontaneous declaratives are so uttered. Similarly, while 83.8% of read *wh*-questions are uttered with falling contours, only 62.2% of spontaneous *wh*-questions are. Although the numbers for yes-no questions are even smaller than for *wh*-questions, the observed behavior shows distinctions between read and spontaneous speech which one might not have predicted: more spontaneous yes-no questions (43.3%) are uttered with falling intonation than read yes-no questions (30%), for utterances of the same tokens.

CHARACTERIZING DISFLUENCY IN SPEECH

The presence of various phenomena known collectively as DISFLUENCIES is commonly thought to distinguish spontaneous speech from read speech. However, identifying what constitutes disfluency in spontaneous speech has proven a difficult topic in itself. While intuitively we may know disfluency when we hear it, the impression of disfluent speech may be described in various ways, and mapped to a large number of distinct lexical, acoustic, and prosodic realizations.

How important a cue disfluency is in identifying speech as spontaneous or not must depend in part on how frequently disfluencies occur, a question that corpus-based studies are well-suited to addressing. Blackmer and Mitton (0) report a rate of one disfluency per 4.6 seconds for callers to a Canadian radio talk show. Studies of large recorded speech corpora have found that approximately 9–10% of spontaneous utterances contain one particular form of disfluency, SELF REPAIRS (0; 0; 0). Yet even where a particular type of disfluency was the subject of study, the acoustic, prosodic, and lexical phenomena which realized the disfluency varied considerably from token to token.

Several multi-speaker corpus-based studies of disfluent speech (0) sought to identify the range of acoustic and phonetic features associated with one type of disfluency, self-repairs, and compared some of these features with fluent read speech. The corpus for these studies consisted of 6414 utterances from the ARPA *atis2* database (0) collected at AT&T, BBN, CMU, SRI, and TI and produced by 122 speakers. 346 (5.4%) utterances contained at least one repair, defined as the self-correction of one or more phonemes (up to and including sequences of words) in an utterance. The speech was labeled for intonational prominences and phrasing following Pierrehumbert's description of English intonation (0). Disfluencies had been categorized independently of these studies as REPAIR (self-correction of lexical material), HESITATION ("unnatural" interruption of speech flow without any following correction of lexical material), or OTHER DISFLUENCY; only the first category were examined in detail.

To provide a framework for the investigation, we further labeled each repair instance using the REPAIR INTERVAL MODEL (RIM) (0). This model divides the repair event into three temporal intervals and identifies critical time points within the intervals. A full repair comprises three contiguous intervals, the REPARANDUM INTERVAL, the lexical material which is to be repaired), the DISFLUENCY INTERVAL (the DI, extending from the termination of the fluent portion of the utterance to the resumption of fluent speech, and containing any number of silence, filled pauses and CUE PHRASES, such as "I meant"), and the REPAIR INTERVAL(the correcting material, which is intended to 'replace' the reparandum). The end of the reparandum coincides with the termination of the fluent portion of the utterance, which we term the INTERRUPTION SITE (IS).

Characteristics of the Reparandum Interval

The most reliable cue found to the presence of a self-correction in utterances in the corpus was the speech fragment. However, the fragments observed in the corpus did *not* represent a homogenous class. Most fragments did tend to occur in content words (43%) rather than function words (5%).[1] Also, 91% of fragments were one syllable or less in length. But there were distributional differences in the phonetic composition of fragments. Single consonant fragments that are fricatives occurred more than six times as often as those that are stops. However, fricatives and stops occur almost equally as the initial consonant in single syllable fragments. Furthermore, we observed two divergences from the underlying distributions of initial phonemes for all words in the corpus. Vowel-initial words were less likely to occur as fragments and fricative-initial words were more likely to occur as fragments, relative to the underlying distributions for those classes in the corpus as a whole. Both the overall and repair distributions and the single consonant and single syllable distributions differed significantly.

In addition to speech fragments, we found evidence that glottalization and coarticulation may be associated with reparanda offsets, especially those ending in fragments;[2] hence, these too may serve as cues to speaking style. In our corpus, 30.2% of reparanda offsets are marked by what we will term INTERRUPTION GLOTTALIZATION. Although interruption glottalization was usually associated with fragments, not all fragments were glottalized.[3] Interruption glottalization appears acoustically distinct from LARYNGEALIZATION (creaky voice), which often occurs at the end of prosodic phrases; the latter typically extends over several syllables at the end of an intonational phrase and is associated with a decrease in energy and low fundamental frequency (0). Interruption glottalization on the other hand tends to occur only over the interrupted syllable, and does not appear to be associated with a sustained decrease in energy and fundamental frequency.[4]

A final feature characterizing the end of some reparanda intervals is the presence of coarticulatory gestures preceding silence. Sonorant endings of both fragments and non-fragments in our corpus may exhibit coarticulatory effects of an acoustically unrealized subsequent phoneme. A related feature is the lack of phrase-final lengthening on the last few segments in the reparandum for many cases of repairs. More generally, both of these features are cues to disfluency in the rhythmic structure of pre-pausal segments.

To summarize, findings from the study of reparanda in self-corrections indicate that this spontaneous speech phenomenon can be realized via word fragments, interruption glottalization, and articulatory gestures that are not fully realized.

Characteristics of the Disfluency Interval

In the RIM model, the DI includes all cue phrases and all filled and unfilled pauses from the offset of the reparandum to the onset of the repair. These phenomena are commonly seen as indicators of spontaneous speech in general, and repair phenomena in particular (0; 0; 0; 0). In a superset of our corpus, only 2.8% (333/11900) spontaneous utterances contained one or more filled pauses and 4.8% (572/11900) contained fragments, while 7.1% (844/11900) contained at least one example of either. In fact, while fragments were reliable cues to repair phenomena in our study, filled pauses were not. We did however find that the duration of silent pauses as the DI was a reliable indicator of the presence of a self-repair, and in fact served also to distinguish between non-fragment and fragment repairs. Overall, silent DIs are significantly shorter than fluent pauses, and the DI duration for fragment repairs is significantly shorter than for non-fragment repairs. The fragment repair DI duration is also significantly shorter than fluent pause intervals, while there is no significant difference between non-fragment repairs DIs and fluent phrase boundaries. So, DIs in general appear to be distinct from fluent phrase boundaries. The presence of such unusual boundaries may thus also be seen as a way of distinguishing spontaneous from read speech.

Characteristics of the Repair Interval

One final distinction between spontaneous utterances containing corrections and fluent utterances was in intonational phrasing. We tested the hypothesis that repair interval offsets are marked by the presence of intonational phrase boundaries by examining whether phrase boundaries observed at that offset differed in their occurrence from those observed in fluent speech for the TI corpus as a whole (0). Using Wang and Hirschberg's (0) phrase prediction procedure, with prediction trained on 478 sentences of read, fluent speech from the *atis* TI read corpus, we estimated whether the phrasing at the repair offset was predictably distinct from this model of fluent phrasing. To see whether these boundaries were distinct from those in fluent speech, we compared the phrasing of repair utterances with the phrasing predicted for the corresponding corrected version of the utterance as identified by *atis* transcribers.[5] We found that in these 63 utterances the repair offset co-occurs with minor or major phrase boundaries for 49% of repairs. For 40% of all repairs, an observed boundary occurs at the repair offset where one is predicted in fluent speech; and for 33% of all repairs, no boundary is observed where none is predicted. For the remaining 27% of repairs, observed phrasing diverges from that predicted by a fluent phrasing model. In 37% of these latter cases, a boundary occurs where none is predicted, and, in the remainder, no boundary occurrs when one is predicted.

We also found more general differences from predicted phrasing over the entire repair interval. Two strong predictors of prosodic phrasing in fluent speech are thought to be syntactic constituency (0; 0; 0), especially the relative inviolability of noun phrases (0), and the length of prosodic phrases (0). In our repair utterances, we observed phrase boundaries at repair offsets which occurred within larger NPs when only a portion of the NP was being corrected. We also found cases in which intonational phrases observed in repair utterances were much longer than phrases observed in fluent speech. In such cases, the absence of intonational phrase boundaries appeared to identify the entire repair as a substituting unit.

So, differences in the location of intonational phrase boundaries as well as the realization of boundaries themselves, fragmentation, interruption glottalization, and coarticulatory phenomena may all be found in spontaneous speech but are found more rarely (not at all in our corpus) in read speech. However, when we attempt to use these observations for the purpose of actually distinguishing between spontaneous and read speech, we find some limitations. While a sizeable fraction of spontaneous utterances contain disfluencies (about 20%, for example, of the *atis0* corpus utterances), a much larger portion do not. Even those containing a disfluent phenomenon exhibit vastly different acoustic and prosodic manifestations of that phenomenon. While the presence of fragments, filled pauses, cue phrases,

[1] 52% of intended words were not recoverable by the transcribers.

[2] See also (0; 0).

[3] In our database, 62% of fragments are *not* glottalized, and 9% of glottalized reparanda offsets are *not* fragments.

[4] We suspect that this phenomenon is similar to that of HOLDING SILENCES investigated by Local and Kelly (0), which occur on discourse connectives; Local and Kelly speculate that these serve the general communicative function of holding the floor.

[5] Results reported here are for prediction on only the 63 TI repair utterances, since the prediction tree we used had been developed on TI utterances.

and differences in location and realization of intonational phrasing may variously distinguish some spontaneous utterances from non-spontaneous ones, the large majority of spontaneous utterances in our corpora contained no such cues.

INDICATORS OF DISCOURSE STRUCTURE

In a third set of studies, read and spontaneous elicited speech were compared to see whether the way speakers convey discourse structural information differs from one style to the other, as previous studies have suggested. Our corpus comprises elicited monologues produced by multiple non-professional speakers, who are given written instructions to perform a series of increasingly complex direction-giving tasks. Speakers first explain simple routes such as getting from one station to another on the subway, and progress gradually to the most complex task of planning a round-trip journey from Harvard Square to several Boston tourist sights. The speakers are provided with various maps, and may write notes to themselves as well as trace routes on the maps. For the duration of the experiment, the speakers are in face-to-face contact with a silent experimental partner (a confederate) who traces on her map the routes described by the speakers. The speech is subsequently orthographically transcribed, with false starts and other speech errors repaired or omitted; subjects return several weeks after their first recording to read the transcribed speech. Both sets of recordings are then acoustically and prosodically labeled, the latter using the ToBI labeling convention (0; 0). Results are given for the spontaneous and the read speech for one speaker, performing five direction-giving tasks.

Discourse segmentations based on Grosz & Sidner's theory of discourse structure were obtained from three subjects labeling from text alone (group T) and three labeling from speech and text (group S). Consensus labels (all subjects in the group agreeing) were obtained for segment-initial (SBEG), segment-final (SF), and segment-medial (SCONT, defined as neither SBEG nor SF). The segmentations of group S differ significantly from those of group T. Unlike results of earlier experiments (0; 0), those who listened to speech while segmenting produced more consensus boundaries for both read and spontaneous speech than did those who segmented from text alone. When the read and spontaneous data are pooled, labelers from speech and text agree upon significantly more SBEG boundaries. Spontaneous speech is generally claimed to exhibit less reliable prosodic indicators of discourse structure than read speech (cf. (0)). Yet, in our corpus, spontaneous speech actually produced significantly more SCONT consensus labels than did read speech, for groups S and T combined. The higher overall percentages of consensus labels for spontaneous speech are attributable to this difference in SCONT labelings.

We examined the following acoustic and prosodic correlates of consensus labelings of intermediate phrases labeled as SBEGs and SFs: f0 maximum and average f0; rms maximum and average; speaking rate; and duration of preceding and subsequent pauses. We compared segmentation labels not only for group S versus group T, but also for spontaneous versus read speech. As noted, while intonational correlates for segment boundaries *have* been identified in read speech, they have been observed in spontaneous speech rarely and descriptively.

We found strong correlations for consensus SBEG and SF labels for groups S and T in both spontaneous speech and read speech.[6] Results on consensus SBEG labels were as follows: given group T segmentations, we found significantly higher maximum and average f0, and maximum and average rms, and shorter subsequent pause for both spontaneous and read speech; for read speech we also found significant correlations for preceding pauses. Given group S segmentations, we found significantly higher maximum and average f0, higher maximum rms, longer preceding and shorter succeeding pauses for read and spontaneous speech; we found higher average rms as well for read speech. Results on consensus SF labels were as follows: given group T segmentations, we found significantly lower average f0 and rms maximum for both read and spontaneous speech, and lower rms average and subsequent pause in addition for read speech. Given group S segmentations, we found lower average f0, rms maximum and average, shorter preceding pause, and longer subsequent pause for both read and spontaneous speech, and in addition, lower f0 maximum for read speech.

So, in this exercise, while we found some differences between read and spontaneous speech, what was *not* different was that speakers indeed seemed to use prosodic and acoustic cues to convey discourse structure.

[6] T-tests were used to test for statistical significance of difference in the means of phrases, e.g. beginning and not beginning segments. Results reported are significant at the .025 level or better.

DISCUSSION

Data from several large corpora of American English read and spontaneous speech thus do provide some comparative information on speaking style issues and raise some interesting additional questions for future analysis.

A study of corpora collected for several DARPA tasks supports the hypotheses that read speech is more rapid than spontaneous speech and that choice of intonational contour does indeed appear to be different for at least some sentence types in read vs. spontaneous utterances. However, a precise characterization of how differences observed in speaking rate between the faster read and slower spontaneous ATIS utterances difference remains to be determined. And, while an examination of read speech found that indeed only 55–80% of yes-no questions were read with rising "yes-no question" contours, thus confirming the hypothesis that read speech does not produce "natural" intonational contours, a comparison of read and spontaneous versions of sentences in the *atis* corpus shows, paradoxically, that over 43% of spontaneous yes-no questions were uttered with **falling** intonation, compared with only 30% of read yes-no questions. So, the notion of a "natural" association between sentence type and intonational contour for yes-no questions in spontaneous speech may need to be re-examined. Similarly, in spontaneous speech, about 30% of declarative sentences were uttered with some type of rising intonation, while over 93% of read declaratives were uttered with final fall. And *wh*-questions, commonly believed to be normally uttered with falling intonation in English, were only so uttered about 62% of the time in spontaneous speech, while in read speech about 84% of such questions were uttered with falling contours. One may speculate that the association between contour type and sentence type may be something that speakers adhere to more consistently when they are asked to read, particularly longer texts, such as the *atis* dialogues, rather than when they produce the same dialogues naturally.

Other questions about comparative speaking style are raised by studies of disfluency in spontaneous speech. While hesitations and self-repairs may occur in collections of read speech, they are rarely preserved; and their presence in speech presented to listeners does seem to provide a useful cue to listeners that the speech they have been presented with is spontaneous. However, do these phenomena, which are quite difficult to define precisely, actually occur often enough to provide reliable cues in general as to speaking style? In the studies of disfluency discussed above, only about 20% of spontaneous utterances contained any form of disfluency — from filled pauses to self-repairs to labeler-observed "hesitations". And the auditory cues that marked such disfluencies themselves range widely, from "abnormal" phrasings to speech fragments to glottalization and coarticulatory devi-

ations from fluent speech. Determining which of these cues are primary indicators of disfluent speech is clearly important, but only speech fragments and filled pauses appear to occur with any great frequency, even for the 20% of spontaneous utterances that contain disfluencies. So, it is clearly important to search for other cues to speaking style variation in addition to acoustically observable indicators of disfluency.

Finally, it has been speculated that a difference between read and spontaneous speaking styles might be inferred from the fact that regularities observable in read speech in the use of prosodic cues to discourse structure, such as pitch range, speaking rate, and pausal duration, had not also been observed for spontaneous speech. However, the comparisons discussed above in the Boston Directions Corpus suggest that this speculation may be incorrect. Statistically reliable associations between prosodic variation and subject-labeled discourse structure have indeed been found for spontaneous as well as read productions of speakers giving directions.

However, before we in fact conclude that the use of intonational variation is not a distinguishing characteristic of read speech, it should be noted that the spontaneous speech collected in this study is not entirely unplanned, but has been elicited from subjects given a task and some time to plan it. While the speech itself is thus spontaneous — subjects did not write out their directions — the productions did follow a planning period. This observation also raises the important point that "spontaneous" speech may vary along many dimensions, including the element of prior planning involved and the presence, size, and type of interlocutors involved.

References

DARPA. *Proceedings of the Speech and Natural Language Workshop*, Hidden Valley PA, June 1990. Morgan Kaufmann.

Julia Hirschberg. Distinguishing questions by contour in speech recognition tasks. In *Proceedings of the Speech and Natural Language Workshop*. Morgan Kaufmann, Cape Cod MA, October 1989.

P. Price, W. M. Fisher, J. Bernstein, and D. S. Pallett. The DARPA 1000-word Resource Management Database for continuous speech recognition. In *Proceedings*, volume 1, pages 651-654, New York, 1988. ICASSP88.

Elizabeth R. Blackmer and Janet L. Mitton. Theories of monitoring and the timing of repairs in spontaneous speech. *Cognition*, 39:173-194, 1991.

Donald Hindle. Deterministic parsing of syntactic non-fluencies. In *Proceedings of the 21st Annual Meeting*, pages 123-128, Cambridge MA, 1983. Association for Computational Linguistics.

Elizabeth Shriberg, John Bear, and John Dowding. Automatic detection and correction of repairs in human-computer dialog. In *Proceedings of the Speech and Natural Language Workshop*, pages 419-424, Harriman NY, 1992. DARPA, Morgan Kaufmann.

C. Nakatani and J. Hirschberg. A corpus-based study of repair cues in spontaneous speech. *Journal of the Acoustical Society of America*, March 1994.

MADCOW. Multi-site data collection for a spoken language corpus. In *Proceedings of the Speech and Natural Language Workshop*, pages 7-14, Harriman NY, February 1992. DARPA, Morgan Kaufmann.

Janet B. Pierrehumbert. *The Phonology and Phonetics of English Intonation*. PhD thesis, Massachusetts Institute of Technology, September 1980. Distributed by the Indiana University Linguistics Club.

John Bear, John Dowding, and Elizabeth Shriberg. Integrating multiple knowledge sources for detection and correction of repairs in human-computer dialog. In *Proceedings of the 30th Annual Meeting*, pages 56-63, Newark DE, 1992. Association for Computational Linguistics.

Joseph Olive, Alice Greenwood, and John Coleman. *The Acoustics of American English Speech: A Dynamic Approach*. Springer-Verlag, New York, 1993.

John Local and John Kelly. Projection and 'silences': Notes on phonetic and conversational structure. *Human Studies*, 9:185-204, 1986.

Willem Levelt. Monitoring and self-repair in speech. *Cognition*, 14:41-104, 1983.

Douglas O'Shaughnessy. Analysis of false starts in spontaneous speech. In *Proceedings of the International Conference on Spoken Language Processing*, pages 931-934, Banff, October 1992. ICSLP.

Michelle Q. Wang and J. Hirschberg. Automatic classification of intonational phrase boundaries. *Computer Speech and Language*, 6:175-196, 1992.

W. E. Cooper and J. M. Sorenson. Fundamental frequency contours at syntactic boundaries. *Journal of the Acoustical Society of America*, 62(3):683-692, September 1977.

J. P. Gee and F. Grosjean. Performance structure: A psycholinguistic and linguistic apprasial. *Cognitive Psychology*, 15:411-458, 1983.

E. O. Selkirk. Phonology and syntax: The relation between sound and structure. In T. Freyjeim, editor, *Nordic Prosody II: Proceedings of the Second Symposium on Prosody in the Nordic language*, pages 111-140, Trondheim, 1984. TAPIR.

K. Silverman, M. Beckman, J. Pierrehumbert, M. Ostendorf, C. Wightman, P. Price, and J. Hirschberg. TOBI: A standard scheme for labeling prosody. In *Proceedings of the Second International Conference on Spoken Language Processing*, Banff, October 1992. ICSLP.

John Pitrelli, Mary Beckman, and Julia Hirschberg. Evaluation of prosodic transcription labeling reliability in the tobi framework. In *Proceedings of the Third International Conference on Spoken Language Processing*, volume 2, pages 123-126, Yokohama, 1994. ICSLP.

B. Grosz and J. Hirschberg. Some intonational characteristics of discourse structure. In *Proceedings of the International Conference on Spoken Language Processing*, Banff, October 1992. ICSLP.

Barbara Grosz, Julia Hirschberg, and Christine Nakatani. A study of intonation and discourse structure in directions. In *Proceedings of the Workshop on the Integration of Natural Language and Speech Processing*. AAAI, August 1994.

Gayle M. Ayers. Discourse functions of pitch range in spontaneous and read speech. Presented at the Linguistic Society of America Annual Meeting, 1992.

THE RELATION BETWEEN VOWEL-TO-VOWEL COARTICULATION AND VOWEL HARMONY IN TURKISH

Patrice Speeter Beddor† and Handan Kopkalli Yavuz‡

†Program in Linguistics, University of Michigan, Ann Arbor, Michigan, U.S.A.
‡English Department, Anadolu University, Eskisehir, Turkiye

ABSTRACT

Does the phonological process of palatal vowel harmony in Turkish parallel its patterns of vowel-to-vowel coarticulation even in words not subject to harmony? Acoustic analysis of disharmonic roots shows mainly anticipatory effects of coarticulation, a different direction from the left-to-right process of palatal harmony. Given current theories of interactions between phonological and coarticulatory organization, this outcome is perhaps not surprising, but all the same points to the complex phonetic and phonological patterns which co-occur in a language.

1. INTRODUCTION

Articulatory and acoustic studies have shown that the coarticulatory effects of vocalic gestures extend beyond adjacent consonants to vowels in flanking syllables (e.g., [1], [2], [3], [4], [5], [6], [7]). These vowel-to-vowel coarticulatory effects are often systematic and can be sufficiently robust that listeners are able to use the information on the coarticulated vowel in identifying the vowel that triggered the effects (e.g., [8], [9], [10], [11]).

This study examines the relation between the phonetic phenomenon of vowel-to-vowel coarticulation and the phonological process of vowel harmony. In vowel harmony systems, only a restricted set of vowels sharing certain features can co-occur in a given string (usually a word). For example, most words of Turkish, Hungarian, and Finnish are subject to 'palatal' harmony whereby all vowels have the same value for the feature [back].

While the coarticulatory effects of one vowel on another are phonetic and local, the effects of vowel harmony are phonemic over a relatively broad domain. Despite these differences, a central assumption underlying the present study is that there is a cause-and-effect relation between vowel-to-vowel coarticulation and many vowel harmony processes in the world's languages. Under this view, the dynamic articulatory behavior becomes 'phonologized' as a language-specific grammatical process. We expect, with Ohala [12], [13] (see also Fowler [14]), that listener misperceptions contribute to phonologization. With respect to vowel harmony, if listeners (particularly language learners) fail to adjust for the coarticulatory effects of one vowel on another, then a coarticulatory effect might be misinterpreted as an inherent property of the coarticulated vowel and in turn be produced as such. In this way, intended non-harmonic VCV sequences could be interpreted as harmonic.

If at least some harmony processes evolve as phonologized coarticulation, what relation holds between coarticulatory organization and vowel harmony in a language with an established harmony system? This study asks whether phonological harmony in a language influences — or, more neutrally, parallels — the current patterns of coarticulation in that language even in words not subject to phonological harmony. In addressing the relation between current phonetic and historically linked phonological patterns in a language, we aim for a better understanding of what types of phonetic and phonological patterns can co-occur in the same language.

Our focus is palatal vowel harmony, in part because vowel-to-vowel coarticulatory effects are reportedly strongest along the front-back dimension [6]. Turkish is a particularly appropriate language for comparison of palatal harmony and vowel-to-vowel coarticulation. Turkish has eight phonemic vowels, front /i, y, e, ø/ and non-front /ɯ, u, a, o/. In most words, all vowels agree in backness, giving rise to front-back alternations in suffix vowels (Turkish being non-prefixing), these alternations conditioned by the backness of the root vowel(s):

	'loaf'	'bone'
Nom. sg.	somun	kemik
Nom. pl.	somunlar	kemikler
Gen. pl.	somunlarɯ	kemikleri

(Turkish also exhibits labial harmony for high vowels; see Boyce [15] for study of the coarticulatory consequences of this process.)

As exhibited by the suffixation patterns above, Turkish palatal harmony is a left-to-right process. There is some controversy as to whether palatal harmony is exclusively progressive or is bidirectional in Turkish. Anderson [16] argues for the former view; Clements and Sezer [17] propose that spreading is bidirectional, with a bias towards spreading from the left.

Although most Turkish words obey palatal harmony, the vowels of a large class of polysyllabic 'disharmonic' roots in Turkish violate this process. Importantly, as discussed in detail by Clements and Sezer [17], although these roots are exceptions to palatal harmony, they are not exceptions to other phonological rules of Turkish and are judged to be well-formed by native speakers in psycholinguistic tests.

The class of disharmonic roots is the testing ground for our study, as these roots provide a means for investigating front-back coarticulation in a language with palatal harmony. An acoustic analysis was conducted to determine the coarticulatory structure of Turkish disharmonic roots. If vowel-to-vowel coarticulation in disharmonic roots were to parallel the (exclusively or primarily) left-to-right direction of palatal harmony in Turkish, then carryover coarticulation should be the predominant pattern.

2. METHOD

Three native speakers of Turkish, 2 male and 1 female, were recorded reading multiple repetitions of a randomized list of CVCV(C) roots embedded in two carrier phrases, either [ʤem __ temizle] ('Cem clean __ ') or [ʤan __ tanɯmla] ('Can identify __ '). (Cem and Can are proper names. The vowels are all front in the first carrier and non-front in the second.) The target words in the reading list involved 58 real-word minimal pairs in which the vowels of one pair member differed in backness (i.e., were disharmonic) and the vowels of the other were identical (hence harmonic). (Sample disharmonic/ harmonic pairs are: [bɨra] 'beer'/ [bara] 'bar (dative)'; [misal] 'example'/ [misil] 'similar'; [deva] 'remedy'/ [deve] 'camel'.) To disguise the purpose of the experiment from the speakers, the reading list also included a large number of filler words.

Vowels in the minimal pair target words were /i, e, a/. Grouping these vowels into corresponding disharmonic - harmonic pairs yields 8 comparison types:

1. a. C<u>a</u>Ci – C<u>a</u>Ca	2. a. CiC<u>a</u> – CaC<u>a</u>
b. C<u>a</u>Ce – C<u>a</u>Ca	b. CeC<u>a</u> – CaC<u>a</u>
c. C<u>i</u>Ca – C<u>i</u>Ci	c. CaC<u>i</u> – CiC<u>i</u>
d. C<u>e</u>Ca – C<u>e</u>Ce	d. CaC<u>e</u> – CeC<u>e</u>

Each of these comparison types was represented in the total set of 58 pairs by 6-9 word pairs.

For each subject, 3 tokens (2 from the back-vowel carrier and 1 from the front-vowel carrier) of each target word were analyzed acoustically. Analysis was restricted to the underscored vowels (<u>i</u>, <u>e</u>, or <u>a</u>) in 1a-d and 2a-d above. Spectral measures of V_1 (in 1a-d) test anticipatory coarticulation and measures of V_2 (in 2a-d) test carryover coarticulation. Formant frequencies were measured at vowel midpoint for both V_1 and V_2, as well as vowel offset for V_1 and vowel onset for V_2 (i.e., offset and onset measures flanked the medial consonant). Measurements were based on superimposed FFT and LPC analyses. Prior to measurement, all vowels (plus a short portion of the adjacent consonants) were extracted from their original context and assigned an arbitrary label, so that experimenters were unaware of the harmonic or disharmonic context in which the vowels were produced.

F2 measurements are reported here, F2 being taken as a primary acoustic correlate of front-back tongue body constriction. Our primary measure of coarticulation is an F2 difference score: for each minimal pair, F2 of a vowel in the harmonic root was subtracted from F2 of the corresponding vowel in the disharmonic root (i.e., [$F2_{disharmonic}$ – $F2_{harmonic}$]). If /a/ were to exhibit

coarticulatory effects of a flanking front vowel, then the F2 difference score should be positive (i.e., F2 of /a/ in the context of disharmonic /i, e/ should be higher than F2 of /a/ in the context of /a/). If /i/ and /e/ were to show coarticulatory effects of flanking /a/, then the F2 difference score should be negative (i.e., F2 of front vowels should be relatively low in the context of non-front /a/).

3. RESULTS

3.1. Anticipatory Coarticulation

The average F2 difference scores for V_1 are shown in Fig. 1. (There was no apparent effect of front- vs. back-vowel carrier; results are averaged over carriers, word pairs, and subjects.) The F2 difference scores for both non-front /a/ (Fig. 1a) and front /i, e/ (Fig. 1b) are in the direction expected if horizontal tongue body constriction for V_2 influenced that for V_1. Specifically, at vowel midpoint and offset, F2 in disharmonic roots was relatively high when /a/ was followed by /i, e/ and was relatively low when /i, e/ were followed by /a/. As would also be expected if these measures reflect coarticulatory effects, the F2 difference scores were smaller (about half as large) at vowel midpoint than at vowel offset. Presumably, the diminishing acoustic effects of V_2 on V_1 reflect the decreasing overlap between gestures for V_2 and V_1 over time.

The data in Fig. 1 also suggest that the magnitude of the coarticulatory effects of V_2 on V_1 is comparable for non-front and front vowels. At both midpoint and offset, the increase in F2 of /a/ attributed to following /i, e/ was roughly the same size as the decrease in F2 of /i, e/ due to following /a/.

Fig. 1 sums across the data for /i/ and /e/ both as *triggers* (Fig. 1a) and *undergoers* (Fig. 1b) of anticipatory coarticulation. The averaged data are representative of the individual behavior of /i/ and /e/ as undergoers of anticipatory coarticulation. Specifically, at vowel offset, the mean F2 difference of -113 Hz is also the mean for each of /i/ and /e/; at vowel midpoint, the F2 difference of -53 Hz is the mean of -44 Hz for /i/ and -62 Hz for /e/.

In contrast, /i/ and /e/ exhibit systematic differences in their behavior as triggers of anticipatory coarticulation. These differences can be seen in Fig. 2, which separates the averaged results for /a/ in Fig. 1a into the results for /a/ followed by /i/ (Fig. 2a) and /a/ followed by /e/ (Fig. 2b), for the 3 speakers (male speakers S1 and S3; female speaker S2). Particularly noteworthy is that, for each of the three speakers, the F2 difference score at /a/ midpoint is considerably greater in the /e/ context than in the /i/ context, with the latter context exerting a negligible influence at midpoint. These results are consistent with a difference in gestural timing in /aCi/ and /aCe/ sequences, with the vowels in /aCe/ sequences having greater temporal overlap.

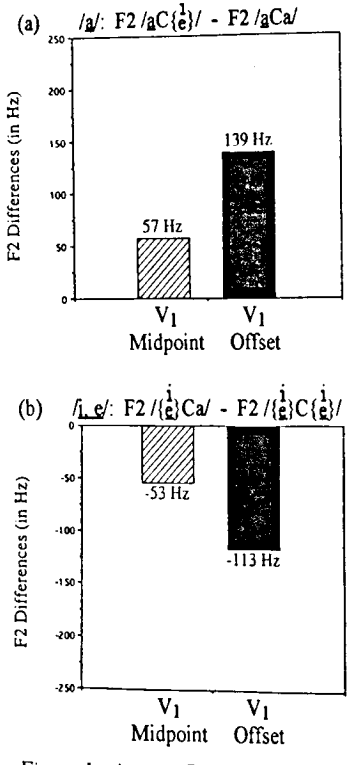


Figure 1. Average F2 difference scores for V_1 measurements at midpoint and offset. Top (a): V_1 = /a/; bottom (b): V_1 = /i, e/.

3.2. Carryover Coarticulation

Of central interest here is whether the carryover effects of vowel-to-vowel coarticulation are as great as or exceed the anticipatory effects. Fig. 3 gives the average F2 difference scores for the midpoint and onset of V_2. For /a/ (Fig. 3a), as expected if preceding /i, e/ exert a coarticulatory influence, F2 in disharmonic roots is relatively high. Comparison of the data in Fig. 3a and Fig. 1a suggests that the carryover and anticipatory effects, respectively, of a flanking front vowel on /a/ are similar at vowel midpoint and approaching the medial consonant (i.e., vowel onset/offset).

However, when V_2 is a front vowel, a quite different picture emerges. Recall that if front vowels show coarticulatory effects of a flanking non-front vowel, then [$F2_{disharmonic}$ – $F2_{harmonic}$] should yield a negative value. The positive values in Fig. 3b are clearly inconsistent with this expectation and suggest that /a/ does not exert a coarticulatory influence on following front vowels. (That the F2 difference scores here are positive rather than more nearly hovering around zero is due largely to Speaker 2's vowels. When $V_1 = V_2$ (i.e., /iCi/, /eCe/, /aCa/), Speaker 2 tended to centralize both vowels. Thus what might appear to be dissimilation -- i.e., a relatively high F2 for /i, e/ preceded by /a/ -- is rather lack of centralization when $V_1 \neq V_2$.)

Are the data in Fig. 3, which sum across /i/ and /e/, representative of these

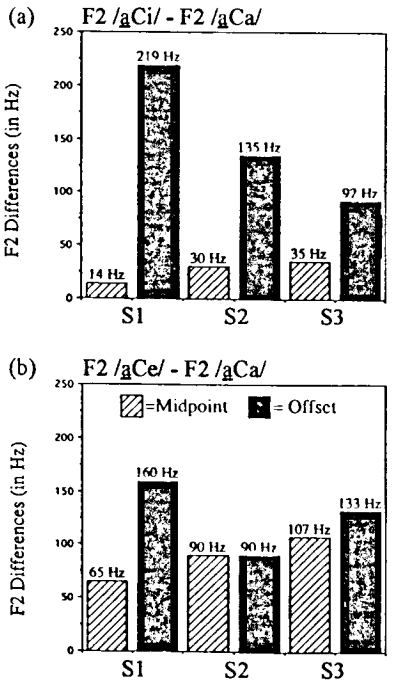


Figure 2. F2 difference scores for three speakers (S1, S2, S3) for V_1 = /a/. Top (a): V_2 of disharmonic root = /i/; bottom (b): V_2 of disharmonic root = /e/.

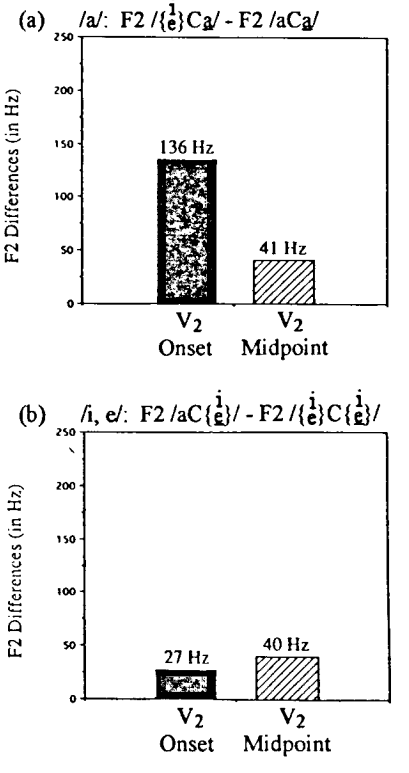


Figure 3. Average F2 difference scores for V_2 measurements at onset and midpoint. Top (a): V_2 = /a/; bottom (b): V_2 = /i, e/.

vowels both as triggers (Fig. 3a) and undergoers (Fig. 3b) of carryover coarticulation? Once again, the averaged data are representative of the individual behavior of the two front vowels as undergoers of coarticulation. (F2 difference scores for both /i/ and /e/ were slightly positive at vowel onset and midpoint.) However, also once again, the averaged data in Fig. 3 are not characteristic of both /i/ and /e/ as triggers of coarticulation. The different patterns of behavior are shown in Fig. 4, which gives the results for /a/ preceded by /i/ (Fig. 4a) and /a/ preceded by /e/ (Fig. 4b) for the three speakers. For all speakers, the F2 difference scores indicate that /i/ has a much greater coarticulatory influence than does /e/ on following /a/. Indeed, Speaker 3's

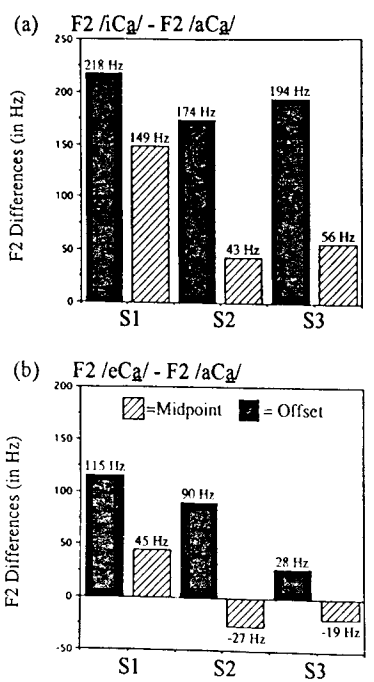


Figure 4. F2 difference scores for three speakers (S1, S2, S3) for V_2 = /a/. Top (a): V_1 of disharmonic root = /i/; bottom (b): V_1 of disharmonic root = /e/.

vowels show no carryover coarticulatory effects of /e/ on /a/.

4. DISCUSSION

Using [$F2_{disharmonic} - F2_{harmonic}$] scores as the measure of vowel-to-vowel coarticulation, the patterns of anticipatory and carryover coarticulation in Turkish VCV sequences were found to differ. All three vowels, /i, e, a/, exhibited effects of anticipatory coarticulation at V_1 offset and to a lesser extent at V_1 midpoint. Moreover, the magnitude of the anticipatory effects was roughly comparable for the three vowel categories. The effects of carryover coarticulation onto V_2 were much more restricted. Only /a/ was susceptible to carryover effects of a preceding vowel, with the strongest effects being due to a preceding /i/.

We evaluate these findings first relative to those of previous studies of vowel-to-vowel coarticulation. Many studies report differences in the coarticulatory patterns of /i/ and /a/, few coarticulatory studies having analyzed /e/. /i/ tends to be a stronger trigger of coarticulation and more resistant as an undergoer of coarticulation than does /a/. One or both of these patterns have been reported for Japanese [18], German [2], Italian [19], Spanish and Catalan [5], Swahili and Shona [20]. The Turkish data reported here corroborate these patterns for carryover coarticulation: /i/ conditioned large frequency shifts in F2 of following /a/, and /i/ (as well as /e/) showed no coarticulatory effects of preceding /a/. However, the pattern does not hold for anticipatory coarticulation: /i, e, a/ behaved similarly as undergoers of anticipatory coarticulation and there was no overall tendency for /i/ to be stronger trigger of anticipatory coarticulation.

Previous studies have also shown that the primary direction (anticipatory or carryover) of vowel-to-vowel coarticulation differs across languages. The Turkish data reported here exhibit a predominantly anticipatory pattern and in this respect are more like the relatively strong anticipatory effects reported for Japanese [18] and Swahili and Shona [20] than like the stronger carryover effects found for English [3] and Catalan [4].

Of particular interest here is assessment of Turkish coarticulatory patterns relative to Turkish phonological structure. Recall that if the primarily left-to-right phonological pattern of vowel harmony in Turkish were to reflect active phonetic patterns of vowel-to-vowel coarticulation, then carryover should exceed anticipatory coarticulation in Turkish disharmonic roots. However, the reverse phonetic pattern holds for the set of roots tested here in that anticipatory coarticulation was found to be a much more general phenomenon than carryover coarticulation.

Of course, it does not follow from the apparent mismatch between the directionality of palatal harmony and front-back coarticulation in Turkish that coarticulatory organization is not linked to phonological structure. It is tempting to speculate that the more consistent effects of anticipatory coarticulation may be linked to word-final stress in Turkish. We suggest this in light of results of previous coarticulatory studies which have systematically manipulated stress location. Although the effects of stress on coarticulatory structure are complex (e.g., [21]), findings generally indicate that stressed vowels exert more of a coarticulatory influence than do unstressed ones on flanking vowels (e.g., [3], [18], [19]).

In view of our speculation on the possible link between patterns of coarticulation and final stress in Turkish, and our assumption of a historical link between coarticulation patterns and vowel harmony, we note with interest that phonological studies have proposed a relation between location of word stress and the development/ demise of progressive vowel harmony [22], [23].

5. CONCLUSION

These acoustic findings suggest that current patterns of vowel-to-vowel coarticulation in Turkish, as measured in disharmonic roots, do not parallel that language's phonological process of palatal harmony. The implications of this outcome for our understanding of the relation between phonetics and phonology are of course subject to debate. One interpretation to be considered is that perhaps this outcome should cause us to question our underlying assumption of a cause-and-effect relation between vowel-to-vowel coarticulation and vowel harmony. That is, perhaps coarticulatory organization does not reflect the left-to-right phonological pattern simply because there is no historical link between the two. A well-argued rejection of this interpretation is beyond the scope of this paper, although we observe that vowel harmony systems tend to be phonetically motivated (e.g., [16], [17]) in that they are organized in terms of features such as [back], [high], [round], which in turn are based on articulatory organization.

The alternative interpretation that we offer of the directional difference between the coarticulatory data and Turkish palatal harmony is that, once a phonetic behavior is phonologized, it becomes a phenomenon largely distinct from the behavior which gave rise to it. This is not surprising. If current theories about the interactions of phonological and coarticulatory structure are correct (see, for example, Fowler [14], Keating [24], and Manuel [7], [21]), then changes in segment inventories, prosodic structure, and other aspects of phonological structure should lead to changes in coarticulatory structure. It does not follow, however, that all historically linked phonological phenomenon would in turn be affected. But while not surprising, this outcome points to the complexity of the phonetic and phonological patterns which can co-occur in a language.

ACKNOWLEDGMENTS

This research was supported in part by National Science Foundation Grant SBR 9319597 to the first author. We thank André Cooper and the phonetics-phonology group at Cornell University for insightful comments, and Alicia Beckford and James Harnsberger for help in data collection and manuscript preparation.

REFERENCES

[1] Öhman, S. E. G. (1966), "Coarticulation in VCV utterances: spectrographic measurements", *Journal of the Acoustical Society of America*, vol. 39, pp. 151-168.

[2] Butcher, A. and Weiher, E. (1976), "An electropalatographic investigation

of coarticulation in VCV sequences", *Journal of Phonetics*, vol. 4, pp. 59-74.
[3] Fowler, C. A. (1981), "Production and perception of coarticulation among stressed and unstressed vowels", *Journal of Speech and Hearing Research*, vol. 46, pp. 127-139.
[4] Recasens, D. (1984), "V-to-V coarticulation in Catalan VCV sequences", *Journal of the Acoustical Society of America*, vol. 76, pp. 1624-1635.
[5] Recasens, D. (1987), "An acoustic analysis of V-to-C and V-to-V coarticulatory effects in Catalan and Spanish VCV sequences", *Journal of Phonetics*, vol. 15, pp. 299-312.
[6] Farnetani, E. (1990), "V-C-V coarticulation and its spatiotemporal domain", in W. J. Hardcastle and A. Marchal (eds.) *Speech production and speech modelling*, pp. 93-130, Dordrecht: Kluwer.
[7] Manuel, S. Y. (1990), "The role of contrast in limiting vowel-to-vowel coarticulation in different languages", *Journal of the Acoustical Society of America*, vol. 88, pp. 1286-1298.
[8] Alfonso, P. J. and Baer, T. (1982), "Dynamics of vowel articulation", *Language and Speech*, vol. 25, pp. 151-173.
[9] Martin, J. G. and Bunnell, H. T. (1982), "Perception of anticipatory coarticulation effects in vowel-stop consonant-vowel sequences," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 8, pp. 473-488.
[10] Fowler, C. A. & Smith, M. R. (1986), "Speech perception as 'vector analysis': an approach to the problems of segmentation and invariance", in J. Perkell and D. Klatt (eds.), *Invariance and variability of speech processes*, pp. 123-139, Hillsdale, NJ: Erlbaum.
[11] van Heuven, V. J. and Dupuis, M. C. (1991), "Perception of anticipatory VCV-coarticulation: effects of vowel context and accent distribution", *Proceedings of the 12th International Congress of Phonetic Sciences*, vol.4, pp. 78-81, Université de Provence Aix-Marseille I.
[12] Ohala, J. J. (1981), "The listener as a source of sound change", in C. S. Masek, R. A. Hendrick, and M. F. Miller (eds.) *Papers from the Parasession on Language and Behavior*, pp. 178-203, Chicago: Chicago Linguistic Society.
[13] Ohala, J. J. (1993), "Coarticulation and phonology", *Language and Speech*, vol. 36, pp. 155-170.
[14] Fowler, C. A. (1990), "Some regularities in speech are not consequences of formal rules: comments on Keating's paper," in J. Kingston and M. E. Beckman (eds.) *Papers in laboratory phonology I: between the grammar and physics of speech*, pp. 476-489, Cambridge: Cambridge University Press.
[15] Boyce, S. E. (1990), "Coarticulatory organization for lip rounding in Turkish and English", *Journal of the Acoustical Society of America*, vol. 88, pp. 2584-2595.
[16] Anderson, S. R. (1980), "Problems and perspectives in the description of vowel harmony", in R. M. Vago (ed.) *Issues in vowel harmony*, pp. 1-48, Amsterdam: John Benjamins.
[17] Clements, G. N. & Sezer, E. (1982), "Vowel and consonant disharmony in Turkish", in H. van der Hulst and N. Smith (eds.) *The structure of phonological representations (Part II)*, pp. 213-255, Dordrecht: Foris.
[18] Magen, H. (1984), "Vowel-to-vowel coarticulation in English and Japanese", *Journal of the Acoustical Society of America*, Suppl. 1, vol. 75, p. S41.
[19] Farnetani, E., Vagges, K., and Magno-Caldognetto, E. (1985), "Coarticulation in Italian /VtV/ sequences: a palatographic study", *Phonetica*, vol. 42, pp. 78-99.
[20] Manuel, S. Y. and Krakow, R. A. (1984), "Universal and language particular aspects of vowel-to-vowel coarticulation", *Haskins Laboratories Status Reports on Speech Research*, vol. SR-77/78, pp. 69-78.
[21] Manuel, S. Y. (In press), "Cross-linguistic studies of coarticulation II: relating language- particular coarticulation patterns to other language-particular facts", in W. J. Hardcastle (ed.) Coarticulation, Cambridge: Cambridge University Press.
[22] Rédei, K. (1987), "The development of vowel harmony in Proto-Uralic and Proto-Finno-Ugrian", *Finnisch-Ugrische Forschungen: Zeitschrift fur Finnisch-Ugrische Sprach- und Volkskunde*, vol. 48, pp. 39-50.
[23] Binnick, R. I. (1980), "The underlying representation of harmonizing vowels: evidence from Modern Mongolian", in R. M. Vago (ed.) *Issues in vowel harmony*, pp. 113-132, Amsterdam: John Benjamins.
[24] Keating, P. A. (1990), "The window model of coarticulation: articulatory evidence", in J. Kingston and M. E. Beckman (eds.) *Papers in laboratory phonology I: between the grammar and physics of speech*, pp. 451-470, Cambridge: Cambridge University Press.

PHONETIC EXPLANATIONS FOR SOUND PATTERNS: IMPLICATIONS FOR GRAMMARS OF COMPETENCE

John J. Ohala
University of California, Berkeley

ABSTRACT

Phonological grammars try to represent speakers' knowledge so that the 'natural' behavior of speech sounds becomes self-evident. Phonetic models have the same goals but have no psychological pretensions. Phonetic models succeed in explaining the natural behavior of speech whereas phonological representations largely fail. The 'phonetic naturalness' requirement in phonological grammars should be re-examined and probably abandoned.

INTRODUCTION

The quest to find a representation of speech sounds that makes their behavior self-evident goes back at least 3 centuries [1, 10, 11] but has been most intense in the past 3 decades. Two approaches to "natural" representation have been developed in parallel, one, the "mainstream" phonological one which employs discrete linguistic primitives and at the same time purports to represent the knowledge of the native speaker [4, 5, 9, 13] and another, phonetic models which are expressed with continuous physical primitives [7, 8, 15, 26, 27, 28] but which do not pretend to reflect psychological structure. In this paper I review certain well-known cases of sound patterns which are better explained by phonetic rather than mainstream phonological representations and then discuss the relevance of this for phonological (mental) grammars.

CONSTRAINTS ON VOICING

There is a well known aerodynamic constraint on voicing in obstruents. Some languages, like Korean and Mandarin have only voiceless stop phonemes; in languages like English that possess both voiced and voiceless stops, the voiceless [p] [t] [k], tend to occur more often in connected speech than the voiced [b] [d] [g]. This constraint derives from the following: voicing (vocal cord vibration) requires sufficient air flow through the glottis; during an obstruent air accumulates in the oral cavity such that oral air pressure rises; if the oral pressure nears or equals the subglottal pressure, air flow will fall below the threshold necessary to maintain vocal vibration and voicing will be extinguished. This constraint can be overcome (within limits) by expanding the oral cavity volume to absorb the accumulating air. Such expansion may be done passively due to the natural compliance or "give" of the vocal tract walls to impinging pressure or actively by lowering the tongue and jaw, lowering the larynx, etc. But there are fewer options for vocal tract enlargement the further back the obstruent is articulated. Thus voiced velar stops are often missing in languages that use the voicing contrast in stops at other places of articulation; they may lose their voicing, their stop character or both. This is the reason why /g/ is missing (in native vocabulary) in, e.g., Dutch, Thai, Czech. See [12, 16, 20] for additional phonetic and phonological data reflecting this.

Additional considerations and variations on this constraint account for (a) the greater bias against voicing in fricatives than stops and in geminate (long) stops than in singletons.[16, 20]

If back-articulated stops such as [g] and [G] are threatened in voiced stop series, it seems that it is the front-articulated stop [p] that is threatened in the voiceless series. This not due as such to aerodynamic but rather to acoustic factors: an abrupt amplitude transient is one of the cues for a stop; the stop burst of a [p] is less intense and thus less noticeable than those for other, further back, places of articulation because a labially-released stop lacks any downstream resonator. [p] seems thus frequently to become a labial fricative (which happened in the history of Japanese). (Although the burst from a voiced [b] would be subject to the same factors, a rapid amplitude gradient on the voicing that follows it would still cue its stop character; with [p] and especially, [p^h], this additional stop cue would be weak.)

Thus for aerodynamic reasons place of articulation can influence what happens at the glottis and for acoustic reasons what happens at the glottis can influence the viability of place distinctions supraglottally.

How have these constraints been represented using conventional phonological notations? Although the phonetic reasons for the voicing constraint are clearly stated (in prose) by Chomsky & Halle in *SPE* [4] (p. 330-1) and they explicitly recognize that the formal representation of phonological rules fails to reflect the 'intrinsic content' of the features, their response is the marking convention (p. 406):

$$[u \text{ voice}] \rightarrow [\text{-voice}] \; / \; \underline{\quad\quad} \; [\text{-son}]$$

(read 'the *u*nmarked value of voice is minus voice in combination with minus sonorant') which is to say that the voicing constraint on obstruents is just stipulated -- it is not made self-evident, it is treated as a primitive. None of the newer formal notations in phonology have offered any improvement.

Feature geometry [5, 13] proposes to capture dependency relations between features using a simple, transitive, asymmetric relation "dominates". 'Simple' in that it is the same relation everywhere it is used; 'transitive' in that if F_a dominates F_b and F_b dominates F_c, then F_a also dominates F_c; 'asymmetric' in that if F_a dominates F_b, then F_b cannot dominate F_a. The relation 'dominate' can be a one-to-many relation, such that a given feature may dominate more than one other features but a given feature may itself be immediately dominated by only one other feature. It follows as a corollary of this that features at intermediate or terminal nodes in the resulting feature hierarchy may not dominate each other. A simplified version of this hierarchy is given in Fig. 1

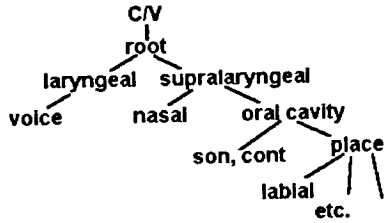


Fig. 1. Feature hierarchy proposed by Clements.

Such an arrangement makes it impossible to capture (other than by stipulation) the aerodynamic constraints between obstruency and voicing, between voicing and place or articulation, or the acoustic constraints between glottal state and supraglottal place of articulation. For the most part Fig. 1 loosely embodies the configuration of the vocal tract by virtue of the particular dependency relations proposed, i.e., separating the laryngeal mechanism from the supralaryngeal system which, in turn, is divided into nasal and oral cavities. The simple monolithic character of the relation 'dominates' prevents a separate encoding of dependency relations due to speech aerodynamics, acoustics, or perception. The asymmetric character of "dominates" prevents simultaneous dominance of place by laryngeal features and vice-versa and prevents dominance by features that are at terminal nodes or intermediate nodes of different branches of

the feature hierarchy. In addition, there is nothing in the feature geometry mechanism to allow only ***one*** value of a feature, e.g., [-voice], to dominate or be dominated by another feature without the other value of the feature ([+voice]) also sharing in the relation.

Thus, as with the formalism used in *SPE* [4], the modern phonological representations can state (stipulate) things known to be true about the behavior of speech sounds but they are inherently incapable of showing the "natural" or self-evident character of these relations. Phonetic models, however, which are also formal, succeed in deriving this behavior from primitives which are for the most part extra-linguistic, e.g., entities and relations from the physical universe [15, 26, 28].

OBSTRUENTS FROM NON-OBSTRUENTS

There are morphophonemic alternations in Welsh and Kwakiutl (and other American languages in the vicinity of British Columbia and Washington state) were the voiced lateral approximant [l] alternates with the voiceless lateral fricative [ɬ]. Although perhaps not obvious, I think a related phonological phenomenon, an extremely common one, is the affrication of stops before high close glides or vowels; see examples in Table 1 (refs in [21]). Both are cases of what I call emergent obstruents [21].

Proto-Bantu	Mvumbo	Translation
*-buma	bvumo	*fruit*
*-dib-	dʒiwo	*shut*
*-tiitu	t͡ʃir	*animal*
*-kiŋgo	t͡ʃiuŋ	*neck, nape*
*-kuba	p͡fuwo	*chicken*
BUT:		
*-bod	buo	*rot* (v)
*-dɪ	di	*eat*

Table 1. Stops become affricated before high, close vowels but not before lower vowels.

Does the obstruent character of the [ɬ] or the affricated release of the stops have to be introduced explicitly by a special rule? Not at all; I claim they are directly derivable from pre-existing elements. To see why, it is necessary to briefly review some of the aerodynamic factors giving rise to turbulence [20, 21, 26, 27].

Turbulence increases when the velocity, v, (so-called 'particle velocity') of the air increases. Particle velocity, in turn, varies as a function of volume velocity, U (how much air is moving past a given point per unit time), divided by the physical characteristics of the channel it moves through, simplified as d (= diameter), in (1).

(1) $$v = U/d$$

Volume velocity, in turn is determined by the area of the aperture, A, through which the air moves, and the pressure difference across that aperture (given as $P_{Oral} - P_{Atmospheric}$), as in (2) ($c$ is a constant and a varies between 1 and 0.5 depending on the nature of the flow).

(2) $$U = A\ (P_{ORAL} - P_{ATMOS})^a\ c$$

From these equations we see that turbulence can be increased by decreasing the cross-dimensional area of the channel. This is the usual view of how fricatives differ from approximants. But I don't think this is what is involved in the cases cited. Rather, another way to create increased turbulence is by increasing U, the volume velocity and this, in turn can be effected by increasing P_{Oral}. In the case of the [ɬ] the P_{Oral} is increased by virtue of its voicelessness: this reduces the resistance at the glottis to the expiratory air flow. The upstream pressure is then essentially the higher pulmonic pressure. Thus the fricative character of the [ɬ] need not result from its having a narrower channel than the approximant [l] but simply from being

[-voice]. In the case of the affrication developing on stops before high close vowels or glides the higher P_{Oral} occurs for different reasons: a stop generates a high upstream pressure; when the stop is released before a high close vowel or glide, some of the air must escape through the narrow channel present. It can take a few tens of milliseconds for the P_{Oral} to reach P_{Atmos} and during this time the air will be forced through the constriction at a higher rate. Hence the initial portions of the vowel or glide can be fricated, especially after a voiceless stop but also after a voiced stop.

To my knowledge there has been no attempt to use current phonological representations to capture the phonetic naturalness of such cases where [-son] elements emerge from [+son] segments simply by appearing simultaneously with [-voice] or sequentially after [-cont, -son]. But, again, the current models such as feature geometry would be inherently incapable of handling these cases because, first, they ignore the aerodynamic aspects of speech and, second, because of the prohibition on dependency relations between separate branches of the feature hierarchy (e.g., [voice] may not dominate [manner]).

EMERGENT STOPS

Occasionally one finds a stop consonant emerging between a nasal consonant and an oral consonant: *Thompson* (< *Thom* + *son*); *Alhambra* (< Arabic *al hamra*, "the red (edifice)"); *humble* (related to *humility*, < Latin *hŭmĭlis* "of the earth"); *empty* < Old English *amtig*; Sanskrit *viṣṇu* "Vishnu" > viṣṭṇu > Bengali *biṣṭu*.

To understand how these stops arise, it is necessary to view speech production (in part) as a process controlling the flow of expiratory air using certain anatomical structures as valves. A nasal consonant is made by channeling air through the nasal cavity: there must be a valvular closure in the mouth and a valvular opening into the nasal cavity (by a lowering of the soft palate). The nasal consonant [m], for example, has the lips closed while the passage between the oral and nasal cavities is open (represented schematically in Fig. 2a). An oral consonant like [s] on the other hand, requires a closure of the nasal valve (by an elevation of the soft palate); see Fig. 2c. If the oral consonant's soft palate closure is made prematurely during the latter portion of the nasal, i.e., undergoes anticipatory assimilation, then with both the oral and nasal valves closed (and there are no other outlet channels for the expiratory airflow) a complete stoppage of the air flow is produced; see Fig. 2b.

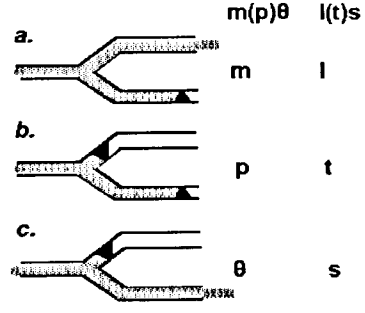


Fig. 2. Schematic representation of the vocal tract and the valves which regulated the flow of air; expiratory air symbolized by grey; valves by black triangles.

Fig. 2 also serves to show the basis for changes of the sort [ls] > [lts] except that in this case the upper branch represents the lateral air passage -- which is open for the lateral [l] and closed for the fricative [s] -- and the lower branch represents the midline passage -- which is closed for [l] and open for [s]. In the transition between these two sounds both air passages may be briefly closed, thus forming a stop. (See [21, 22] for more details, further data and references, and discussion of how the same principles

can account for some cases of emergent ejectives and clicks.)

Using autosegmental notation, Wetzels [29] and Clements [6] correctly characterize /mΘ/ > [mᵖΘ] as arising from the spreading of [-nasal] from the [Θ] into the [m] (although they incorrectly assume that such spreading could not occur from left-to-right as in Sanskrit *visnu*, cited above). But in the case of [ls] > [lts] or [sl] > [stl], they resort to a rule that simply inserts a consonant; it seems they are unable to generate a [-continuant] from the spreading of features from two [+continuants]. But as detailed above, and illustrated in Fig. 2, the overlap of gestures from two continuants *can* create a non-continuant or obstruent! The problem with the phonological representations here lies in taking [±continuant] or [±sonorant] as *primitives*; they are in fact *derived* from the states of the valves which control airflow.

THE STORY OF [w].

The labial velars [w] [kp] [gb] [ŋm]) are doubly-articulated consonants, having two simultaneous primary constrictions, labial and velar. In spite of their two constrictions in certain cases these sounds pattern with simple labials such as [p] [b] [m] and in other cases with simple velars [k] [g] [ŋ]. But their behavior as labial or velar depends on the nature of the particular contextual effect involved.

When generating noise, labial velars are labial.

When generating noise (frication or stop bursts) labial velars tend to behave as labials. Some examples: Brit. English [lɛf'tɛnənt] for *lieutenant*; in Tenango Otomi the /h/ before /w/ is realized as the voiceless labial fricative [Φ]. The probable reason for this is that since noise is inherently a relatively high frequency sound, even if noise were generated equally at both the velar and labial places, the noise at the velar constriction would be attenuated by the low-pass filtering effect of the downstream resonator. (See [7, 27].)

Nasals assimilating to [w] are velar

A nasal assimilating to the labial-velar [w], insofar as it shows any assimilatory change and shows only one place of articulation, becomes the velar nasal [ŋ], not the labial nasal [m]. Tswana /-roma/ "send" + /wa/ (pass. sfx.) = /-roŋwa/; Melanesian dialects show the variant pronunciation /mwala/ ~ /ŋwala/ for the name of *Mala Island*.

Some principles adduced by Fujimura [8] help to explain this pattern. (See also [24].) Fig. 3 gives a schematic representation of the air spaces that determine the resonances of nasal consonants. As shown in the figure, all nasal consonants have the pharyngeal-nasal air space in common (marked by a dashed line). What differentiates one nasal from another is the length of the air cavity (marked by a dotted line), a cul-de-sac, which branches off of this pharyngeal-nasal air space. Measured from the point where the two air cavities diverge, this branch is quite long in the case of the labial nasal [m] but is quite short in the case of the velar nasal [ŋ]. In the case of the labial-velar nasal there are two constrictions, one labial and one velar, but only the rearmost constriction defines the extent of the branch (measured from the point where it diverges from the pharyngeal-nasal cavity); the forwardmost (labial) constriction will be largely irrelevant in determining the characteristic resonances. Thus labial-velar nasals will tend to sound like simple velar nasals.

These labial velar sound patterns could not fall out from current phonological representations since they fail to incorporate aerodynamic and acoustic relations and do not allow for dependency relations between [nasal], [manner], [place] features.

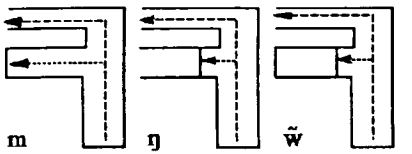


Fig. 3. Schematic representation of the air spaces creating the distinctive resonances for nasal consonants. From left to right: [m], [ŋ], and a labial-velar nasal. Dashed lines show the air spaces identical in all nasals; the dotted lines show those air spaces whose dimensions may differ between nasals.

CONCLUSION: EXPLANATIONS FOR SOUND PATTERNS IN LANGUAGE

Mainstream phonological representations purport to simultaneously (a) reflect speakers' knowledge of the sound patterns in their language and (b) represent this knowledge in a way that makes the the 'naturalness' of the sound patterns self-evident. I have tried to demonstrate in this paper that the second goal is not achieved. Could this goal be met by some appropriate modification of the representations used, e.g., by some new feature geometry having separate dependency relations for speech articulations, aerodynamics, acoustics, etc., including the inter-domain dependencies (aerodynamics and acoustics together determine why the noise generated by labial velars is predominantly labial)? I submit that if such a revised representation were constructed, one that would then be capable of embodying the 'intrinsic content' of the elements of speech, it would be identical to the phonetic models referred to above. But this solution would be unacceptable because such models use continuous physical parameters and physical relations between them such as Boyle-Mariotte's Law -- and with justification no one believes that a speaker's competence includes knowledge of physics.

I see no way out of this impasse except to abandon the requirement that phonological grammars reflect the phonetic naturalness of the sound patterns in language. Can we justify this step and, if so, what are its consequences?

A full justification would require more space than I am allotted here but a few comments are possible. In searching for the origin of the requirement that the rules in the speaker's mental grammar reflect phonetic naturalness, it seems that it came about in two steps. In *Syntactic structures* [3] (and earlier [2]) Chomsky proposed that simplicity be a criterion for evaluating competing grammars. By explicit projection the criterion used by the linguist to find a theory (grammar) of the language should also be the criterion the language learner uses to evaluate grammatical specifications of his language. Features, alpha variables, abstract underlying forms, ordered rules, the transformational cycle, etc. were subsequently shown to lead to simplifications of grammar. At this, the 'quantitative' phase, simple length was used to evaluate grammars and their parts. In *SPE* [4], chap. 9, the authors declared that a simple quantitative evaluation of rules was not sufficient and that further simplifications could be achieved if a qualitative differentiation could be made between common, widely-attested sound patterns and those less common -- a way that would reflect the 'intrinsic content' of features. Thus the burden was shifted to the representation of speech sounds. The marking conventions, autosegmental notation, feature geometry, etc. were designed to incorporate more of the inherent structure -- presumably their phonetic structure -- which is responsible for the 'natural' behavior of speech. But as far as I have been able to tell the proposals to make grammars quantitatively and qualitatively optimal were made without any serious consideration of psychological evidence or implications. The whole notion of 'simplicity' is quite elusive and what sorts of

optimization speakers impose on their mental representation of the phonological component of their language is largely unknown. The amount of psychological evidence on speakers' awareness of what is phonetically natural is in inverse relation to the impact that the issue of 'naturalness' has had on mainstream phonological theory. Moreover, there is some evidence that non-phonetic factors, e.g., morphology, semantics, play a much more important role in speakers' conception and manipulation of sound patterns [25].

The existence of phonetically natural processes in the sound patterns of languages needs no special or extravagant explanation. Universal, physical phonetic factors lead to a speech signal which obscures the speaker's intended pronunciation; listeners may misinterpret ambiguous phonetic elements in the signal and arrive at a pronunciation norm that differs from the speaker's. This is how sound change works [18, 19] and how natural sound patterns arise. Such changes will reflect phonetic constraints without speaker or listener having to "know" about them. Similarly, when we eat, walk, see, hear, etc. our behavior is subject to a myriad of universal physical constraints without the necessity of our knowing them either consciously or unconsciously. Even a rock obeys the laws of physics without having to know them.

There is, in sum, more than ample justification to abandon the 'phonetic naturalness' requirement for the representation of speakers' competence.

What would the consequences of this move be for current phonological practice? Historical grammars or any account of phonological universals would, as now, still have to meet the requirement of representing speech sounds in a way that would accurately predict their behavior. For this purpose existing phonetic models suffice, as illustrated in the body of this paper. Of course, there is now and always will be a need to elaborate and revise existing models and to introduce new ones as we seek to explain more sound patterns in language. Adequate representations of native speakers' competence could -- ironically -- be much simpler, possibly formulated with no more than unanalyzed phonemes [14]. There may be no need for features, underspecification, autosegmental notation, feature geometry or similar speculative devices. However, whatever is attributed to the speaker's mental grammar should be subject to the same standards of empirical verification as are elements in phonetic models. Such evidence would probably come from psycholinguistic experiments.

No matter what sort of account or model is given of speech sounds and their behavior it would be beneficial if they were preceded by an explicit statement regarding what part of the universe the model represented, whether the speaker's vocal tract, the speaker's mind, or the speaker's DNA. That would determine the part of the universe where empirical verification of the model would be sought.[17, 23]

ACKNOWLEDGEMENT

Parts of this paper recapitulate arguments given in [21] and were inspired in part by [14].

REFERENCES

[1] Amman, J. C. (1694), *The talking deaf man*, London: Tho. Hawkins.
[2] Chomsky, N. (1956), *The logical structure of linguistic theory*. [Ms.] Cambridge: Harvard College Library.
[3] Chomsky, N. (1957), *Syntactic structures*. The Hague: Mouton.
[4] Chomsky, N. & Halle, M. (1968), *Sound pattern of English*, New York: Harper & Row.
[5] Clements, G. N. (1985), "The geometry of phonological features." *Phonol. Yrbk.*, vol. 2, pp. 225-252.
[6] Clements, N. S. (1987), "Phonological feature representation and the description of intrusive stops", *CLS Parasession*, vol.23, pp. 29-50.
[7] Fant, G. (1960), *Acoustic theory of speech production*, The Hague: Mouton.
[8] Fujimura, O. (1962), "Analysis of nasal consonants", *J. Acoust. Soc. Am.* vol. 34, pp. 1865-1875.
[9] Goldsmith, J. (1979), *Autosegmental phonology*. New York: Garland Press.
[10] Jespersen, O. (1889), *Articulation of speech sounds, represented by means of analphabetic symbols*. Marburg in Hesse: N. G. Elwert.
[11] Key, T. H. (1855), "On vowel-assimilation, especially in relation to Professor Willis's experiment on vowel-sounds." *Trans. Philological Society* [London], vol. 5, pp. 191-204.
[12] Maddieson, I. (1984), *Patterns of sounds*, Cambridge: Cambridge University Press.
[13] McCarthy, J. J. (1988), "Feature geometry and dependency: a review." *Phonetica* vol. 45, pp. 84-108.
[14] Myers, J. (1994), "Autosegmental notation and the phonetics-phonology interface",. Unpub. Ms. Buffalo.
[15] Ohala, J. (1976), "A model of speech aerodynamics", *Rep. Phonology Lab.* (Berkeley) vol. 1, pp. 93-107.
[16] Ohala, J. (1983), "The origin of sound patterns in vocal tract constraints." P. F. MacNeilage (ed.), *The production of speech*. New York: Springer-Verlag. 189-216.
[17] Ohala, J. (ed.), (1986), "The validation of phonological theories", *Phonol. Yrbk.* vol. 3.
[18] Ohala, J. (1993), "The phonetics of sound change", C. Jones (ed.), *Historical Linguistics: Problems and Perspectives*. London: Longman. pp. 237-278.
[19] Ohala, J. (1993), "Sound change as nature's speech perception experiment", *Speech Comm.*, vol. 13, pp. 155-161.
[20] Ohala, J. (1994), "Speech aerodynamics", R. E. Asher and J. M. Y. Simpson (eds), *The Ency. Lang. & Ling.* Oxford: Pergamon, pp. 4144-4148.
[21] Ohala, J. (In press), "Emergent obstruents", D. Demolin & M. Dominicy (eds.), *Studies in sound change*. Amsterdam: J. Benjamins.
[22] Ohala, J. (In press), "A probable case of clicks influencing the sound patterns of some European languages. *Phonetica*.
[23] Ohala, J. & Jaeger, J. (eds.), (1986), *Experimental phonology*. Orlando, FL: Academic Press
[24] Ohala, J. & Ohala, M. (1993), "The phonetics of nasal phonology: theorems and data", M. Huffman & R. Krakow (eds.), *Nasals, nasalization, and the velum*. San Diego, CA: Academic Press. pp. 225-249.
[25] Ohala, M. and Ohala, J. (1987), "Psycholinguistic probes of native speakers' phonological knowledge", W. U. Dressler, et al. (eds.), *Phonologica 1984*. Cambridge University Press. pp. 227-233.
[26] Scully, C. (1990), "Articulatory synthesis", W. J. Hardcastle & A. Marchal (eds.), *Speech production and speech modelling*. Dordrecht: Kluwer. pp. 151-186.
[27] Stevens, K. N. (1971), "Airflow and turbulence noise for fricative and stop consonants: Static considerations", *J. Acoust. Soc. Am.* vol. 50, pp. 1180-1192.
[28] Westbury, J. & Keating, P. (1985), "On the naturalness of stop consonant voicing", *Working Papers in Phonetics* (UCLA), vol. 60, pp. 1-19.
[29] Wetzels, W. L. (1985), "The historical phonology of intrusive stops. A nonlinear description", *Canadian J. Ling.* vol.30, pp. 285-33.

A MODEL OF HUMAN JAW AND HYOID MOTION AND ITS IMPLICATIONS FOR SPEECH PRODUCTION

R. Laboissière†, D.J. Ostry, P. Perrier†*

†Institut de la Communication Parlée
URA CNRS 368 / INPG / Univ. Stendhal
46, av. Félix Viallet 38031 Grenoble CEDEX 1 France

**Dept. of Psychology, McGill University*
1205 Dr. Penfield Avenue, Montreal QC Canada H3A 1B1

E-mail: rafael@icp.grenet.fr

ABSTRACT

We present a model of sagittal plane jaw and hyoid motion based on the λ version of the equilibrium point hypothesis. This hypothesis suggests that movements arise from shifts in the equilibrium position of the speech articulator. The equilibrium is described as a consequence of the interaction of central neural commands, reflex mechanisms, muscle properties and external loads, but it is under the control of central neural commands. These commands act to shift the equilibrium via centrally specified signals acting at the level of the motoneurone (MN) pool. In the context of the model, we focus on a number of issues in speech research. We consider the implications of the model for the notion of articulatory targets. We suggest that simple linear control signals may underlie smooth articulatory trajectories. We explore as well the phenomenon of intra-articulator coarticulation in jaw movement.

INTRODUCTION

A major difficulty in inferring control strategies in speech from the kinematic data characterizing human orofacial motion is the lack of physiologically based models of the underlying control. Without such models, it is difficult to assess the adequacy of descriptive accounts in which the underlying control is contaminated by factors such as dynamics and muscle mechanical properties. Physiological models permit the separation of aspects of the kinematics due to neural control from those due to the biomechanical properties of the system. In this paper, we present one such physiological model, the equilibrium point hypothesis of motor control, and consider its specific application to issues in speech control.

The model is introduced in detail below. However, very briefly, the hypothesis suggests that movements arise from shifts in the equilibrium position of the speech articulator. The equilibrium is a consequence of the interaction of central neural commands, reflex mechanisms, muscle properties and external loads, but it is under the control of central neural commands. These commands act to shift the equilibrium via centrally specified signals acting at the level of the motoneuron (MN) pool.

In the sections which follow, we introduce the basic concepts of the EP hypothesis. We describe its application to the development of a seven muscle, four degree of freedom model of jaw and hyoid motion (Laboissière, Ostry, and Feldman submitted). With the aid of simulations we show (1) how the concept of equilibrium position provides insights into the concept of articulatory targets and (2) how at least some of the observed variability associated with intra-articulator coarticulation can be attributed to dynamics rather than central control.

THE JAW MODEL

Biomechanical Structure One of the main goals in modelling orofacial function is to study the form of CNS commands which underlie the kinematic observables. Consequently, in order to understand control on the basis of kinematics, it is necessary to be able to separate control signals from the system's biomechanics. Modelling the elaborate geometry and dynamics of the orofacial articulators is a necessity, not because of a specific interest in their characteristics, but because it is otherwise exceedingly difficult to relate control signals to the resulting kinematics which may be measured empirically.

With this aim, we have recently developed a model of the jaw and hyoid system (see Fig. 1) (Laboissière, Ostry, and Feldman submitted). The model, which is implemented as a computer simulation, has seven muscles (or muscle groups) and four kinematic degrees of freedom. Movements are not controlled directly in terms of commands to individual muscles. Rather, consistent with empirical evidence (Bothorel 1975; Ostry and Munhall 1994), control signals, which are based on different combinations of commands to muscles, are organized at the level of the system kinematic degrees of freedom. This enables independent production of jaw rotation, horizontal jaw translation, vertical hyoid translation and horizontal hyoid translation. The level of cocontraction is also controlled. These control signals may act alone or in combination.

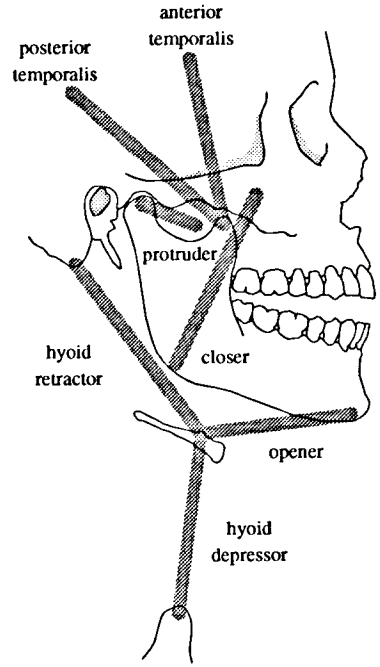


Figure 1: Schematic representation of modelled muscles with their attachments to the jaw and hyoid bone.

Control of Jaw Motion: The Equilibrium Point Hypothesis Motor innervation to muscles arises from α MNs which innervate the main body of the muscle and from γ MNs which contribute to a MN excitation through reflexes. The basis of the model is the suggestion that movement arises from changes to neural control variables which shift the equilibrium state of the motor system. The essential control variables are independent changes in the membrane potentials of α and γ motoneurones (MNs) which establish a threshold muscle length (λ) at which the recruitment of MNs begins. Muscle activation and hence force vary in relation to the difference between the actual and the threshold muscle lengths and the rate of muscle length change. Thus, by shifting λ through changes to the central facilitation of MNs, the system can produce movement to a new equilibrium position.

The central notions associated with the ba-

sic control mechanism are shown in Fig. 2 in the context of a single jaw muscle and load. For simplicity, we will focus on the jaw closer muscle masseter (shown in black) and on the load due to the gravitational force. The panel on the left shows a number of different jaw configurations. The corresponding depolarization of α MNs is shown at the top right. The horizontal line gives the threshold for MN recruitment. The descending input to the MN provides the level of central facilitation which may be specified independent of muscle length. Afferent facilitation also contributes to the depolarization of the MN and varies directly with muscle length. Thus, while the equilibrium position is essentially under central control, the activation level of the MN reflects a net contribution which includes both the direct descending input to the MN and indirect input due to afferent pathways.

In (a), we see, in the left hand panel, the subject resting horizontally with the system in equilibrium at a position near to occlusion. The level of total depolarization, as seen in the top right hand panel, exceeds the threshold level and the load due to gravity is supported by an overall level of central and afferent activity. When the subject changes to a vertical position (b), the load acting to extend the masseter increases. This increases the level of muscle-length dependent afferent facilitation of the MN pool which in turn acts to establish a new equilibrium position. Note, that the level of central facilitation is unchanged by these changes in load even though the total level of MN depolarization is changed.

The lower right hand panel of Fig. 2 demonstrates these characteristics in terms of the muscle's force-length curve. The variable λ gives the muscle length at which motoneuron recruitment begins. The exponential shape of the force-length relation reflects the well known size-principle for MN recruitment such that as the difference between the actual and threshold muscle length progressively increases, progressively larger motor units with larger force outputs are recruited. At muscle length l, a force equal to F is generated which balances the load (a). A change in the position of the head relative to the gravitational force loads the jaw and stretches the muscle to length l'. The length dependent afferent facilitation results in the recruitment of new motor units which increases force to F' (b). At this point, the muscle force balances the load force. To summarize, changes in load which result in muscle stretch (or unloading) lead to the recruitment (or derecruitment) of motor units as a result of changes in length dependent facilitation to the MN pool. The measure of independent central control, λ, is unaffected even though both force and muscle length are changed.

As shown by comparing (b) and (c), the model suggests that voluntary movement arises as a consequence of increases in the level of central facilitation to the MN pool. Increases in facilitation depolarize MNs and result in the recruitment of additional motor units. This increases total force and results in muscle shortening. As the muscle shortens, the facilitation to the MN pool due to length dependent afferent input decreases and a new equilibrium is established. Voluntary movements are depicted in the lower right hand panel in terms of the muscle's force-length relation. At (b), where threshold muscle length is λ, the weight of the jaw is supported by muscle force F' at muscle length l'. Increasing central facilitation serves to reduce the threshold length for MN recruitment from λ to λ'. As λ shifts, the difference between the actual and threshold muscle length increases, more MNs are recruited and the muscle begins to actively shorten. As the threshold length reaches λ', the jaw achieves a new equilibrium state in which muscle force is F' and muscle length is l. The movements which arise from changes to the independently specified parameter λ thus depend on both direct central facilitation to the MN pool and facilitation arising from afferent input to the MN.

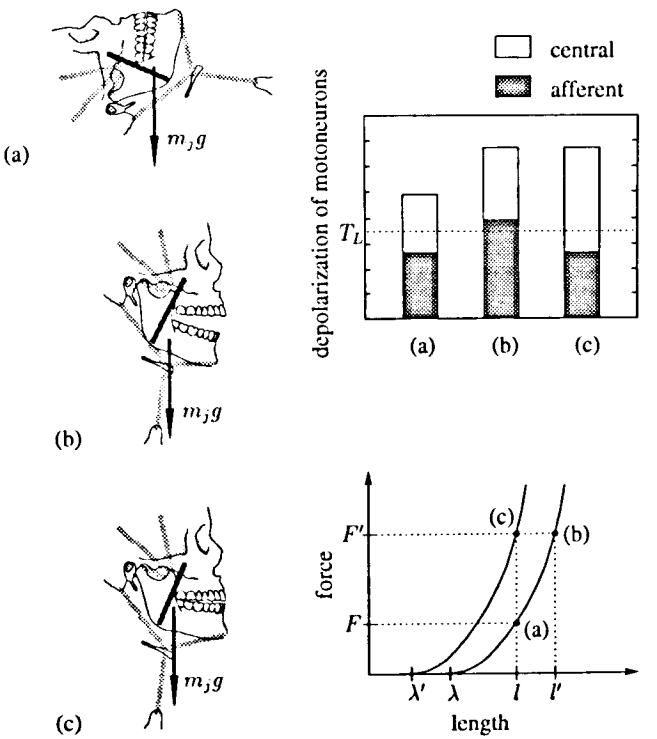


Figure 2: Jaw positions (left hand panel), levels of depolarization of MNs (upper right), and force-length curves (lower right) are shown (see text for details).

Control signals which are associated with changes in values of the system's four degrees of freedom are each mapped onto control signals at the level of individual muscles (λ). This corresponds to a time varying facilitation to the MNs of each muscle. As described above, the muscle activation and force depend on the difference between the threshold muscle length and the current length as well as the rate of muscle length change. Fig. 3 shows the layout of the model and its control mechanisms. Muscle length and velocity information are provided by muscle spindle afferent input to MNs following a reflex delay. The elements in each of the muscle blocks in the figure correspond to individually modelled muscle mechanical properties. These include muscle properties such as the dependence of force on muscle length and on passive elastic properties and the graded development of muscle force due to calcium dependent muscle kinetics. Mechanical damping is provided by velocity dependent reflex inputs and muscle intrinsic properties. The force arising in each muscle contributes to the production of jaw and hyoid forces and torques. These act through the system's equations of motion to produce changes in jaw and hyoid position and orien-

namics and realistic musculo-skeletal geometry.

The model thus provides the means to study human jaw motions in speech in a manner which explicitly integrates biomechanical characteristics and the underlying control. In the following section we will explain the significance of a number of properties of the λ model and then, using simulations with the jaw model, show their implications for the control of speech

THE FORM OF CENTRAL COMMANDS, INVARIANCE AND VARIABILITY IN SPEECH

The Form of Central Commands One of the fundamental problems in motor control is obtaining an understanding of the form of the neural signals which underlie movements. In speech, the form of the control signal is of particular importance as it contributes to our understanding of the relationship between the phonological level and the corresponding organization at the level of the vocal tract. By exploring speech at the level of control, we can assess extent to which the regularities observed kinematically and hence acoustically correspond to invariances postulated at the linguistic level. The EP hypothesis, by its very nature, allows us to address in what ways the equilibrium as specified at the level of motor system might correspond to aspects of spatial targets, which serve as landmarks in the control of the speech sequence.

According to the model, movements are effectively changes in posture, that is, shifts in the equilibrium state of the system. We suggest that the control of speech may be related to specific postures of articulators and hence that posture and successive changes in posture correspond to a representation of the articulatory task at the level of control. This idea that articulatory movements are intended towards spatial positions is related to MacNeilage's (1970) proposals of spatial articulatory targets.

Fig. 4 lets us explore the concept of speech targets in the context of the jaw model. The figure shows empirical and simulated jaw motions during repetitions of /isisa/. The empirical data are shown with solid lines, the predicted jaw kinematics with dots and the presumed underlying equilibrium shifts with alternating dots and dashes. The jaw orientation angle is shown in the upper panel and horizontal jaw position is shown below.

In fitting the data, we have assumed that the jaw equilibrium angle and equillibrium horizontal position both shift at a constant rate. Changes in the rate and duration of the equilibrium shift are the two controlled variables. Examination of the data shows that the correspondence between empirical and model data is generally good. Note that constant rate equilibrium shifts in the model produce the smooth movements which are observed kinematically. This suggests that smoothness of movements may not be explicitly planned or controlled but rather may arise from the dynamics.

The equilibrium shifts, particularly in the case of horizontal jaw tranlsation, are often observed to extend beyond the kinematic endpoints of the movement. The overshoot of the actual trajectory by the equilibrium arises in the model from the need to produce the sufficiently large accelerations which are required to move the jaw in a continuous fashion at speech rates. The need to have the equilibrium position overshoot the actual spatial goal to produce rapid movement has also been demonstrated in simulations of multi-joint arm movements (Hogan 1984).

The idea that articulatory movements are intended towards spatial positions has been proposed by MacNeilage (1970). Our simulations suggest that the literal interpretation of spatial targets as actually achieved positions within the vocal tract may be incorrect in the case of continuous speech. Nevertheless, we think it is reasonable to assume that regularities relating speech as a linguistic task to speech at the motor level may be found in terms of the control signals underlying movement. This will, of necessity, entail a comparison of empirical and model data.

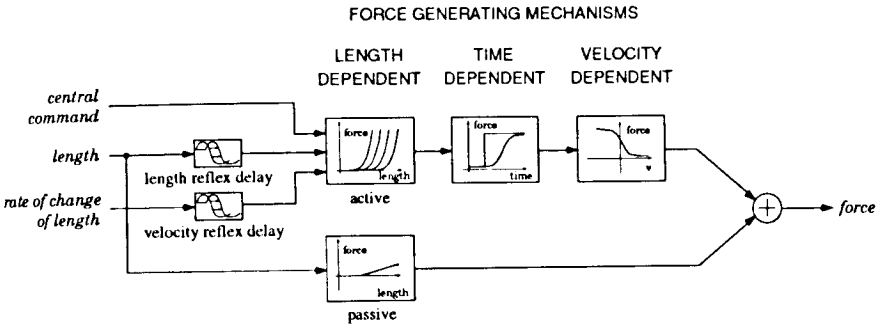


Figure 3: Schematic representation of the muscle model

Regularities relating the units of linguistic description to the control signals of speech motions might be sought in terms of correspondences related to both equilbrium position and rates of equilibrium shift.

Speech Variability Variation in articulatory movement is clearly one of the most pervasive characteristics of speech. Some of the aspects of speech movement variability are almost certainly planned and relate to specific changes in control signals. Others may not be planned but may arise from factors such as muscle mechanics, musculo-skeletal geometry and the dynamics of the physical system. Evidence that inter-articulator variations in speech are planned is supported by the findings of Abry and Lallouache (in press, also see Perkell and Matthies 1992). These authors analyzed anticipatory lip protrusion in [iCy] sequences, in which C represents consonant clusters of 0 to 5 consonants, none of which involved lip protrusion. They showed that the onset time of the protrusion movement increased linearly with the size of the consonant cluster. The fact that lip protrusion to produce the same final vowel begins earlier in some contexts than in others supports the idea that anticipatory patterns are the result of a process which takes account of upcoming phonetic context when planning successive speech movements.

The kinematic patterns of intra-articulator coarticulation which are readily measurable in empirical studies may also appear to be centrally controlled on the basis of kinematic changes which arise in response to upcoming phonetic segments (e.g., Lindblom 1983). However, without explicit models of speech articulators, measured kinematic effects correctly attributable to central planning cannot be distinguished from the kinematic patterns which are due to dynamics and are not represented in the underlying control. To address this possibility, we will show how kinematic variability may arise even when the underlying control signals related to the specification of articulatory position remain fixed. The main conclusion we will wish to draw is that unplanned effects due to physical sources must be accounted for before drawing conclusions about central control or inferring planning mechanisms.

Using the jaw model, we have studied the predicted kinematic patterns in simulated V_1CV_2 transitions. In these simulations, the equilibrium shifts associated with the V_1C movement remain constant while the equilibrium shifts associated with the CV_2 movement amplitude are systematically varied. Thus, at the level of central control, no account was taken of upcoming phonetic context in the specification of the CV_2 transition. However, when one examines the predicted kinematic patterns (Fig. 5), we see that the V_1C amplitude and duration are systematically affected by the identity of the

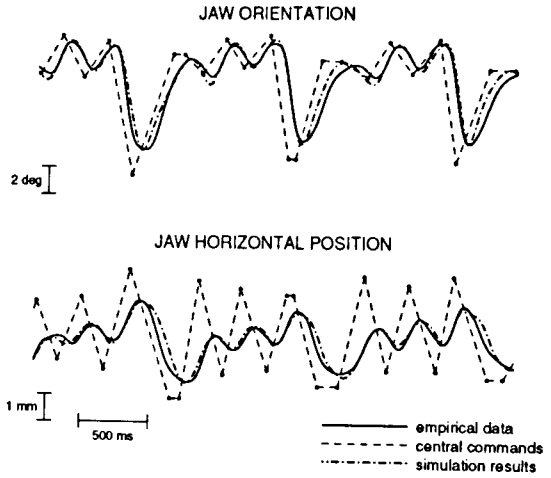


Figure 4: Empirical and model data during repetitions of /isisa/. The hyoid bone is at the assumed rest position for occlusion through the simulated movement.

final vowel. As movement amplitude for the final vowel decreases, the simulated amplitude and duration of the initial transition increase. Comparable patterns of intra-articulator coarticulation have been reported in empirical studies of jaw, tongue dorsum, velar, and lower pharyngeal wall coarticulation (Ostry and Gracco 1995; Parush, Ostry, and Munhall 1983; Parush and Ostry 1986; Parush and Ostry 1993). Thus, while on the basis of kinematic evidence alone, it could be concluded that intra-articulator coarticulation is consistent with the notion of planned coarticulation, our present simulations suggest that this possibility should be evaluated with care: unplanned effects due to articulator dynamics must be accounted for before drawing conclusions about the role of central control in intra-articulator coarticulation.

CONCLUSION

In this paper, we have described a model of jaw and hyoid motion based on the EP hypothesis. We have described simulations which examine the form of the central control signal. We have shown that smoothness in movement may arise from dynamics and need not be planned. We have suggested that regularities relating speech as a linguistic task to speech at the motor level may be found in the control signals underlying movement. We have examined sources of articulatory variability. We have shown that kinematic patterns comparable to those reported in intra-articulator coarticulation may arise as a result of dynamics rather than central planning.

ACKNOWLEDGEMENTS

This research was supported by NIH grant DC-00594 from the National Institute on Deafness and Other Communication Disorders, Advanced Telecommunications Research (ATR), Kyoto, Japan, European Community ESPRIT - BR grant # 6975 Speech Maps, Natural Sciences and Engineering Research Council of Canada, and France-Quebec project # 070192. We also acknowledge the Rhône-Alpes Region for support.

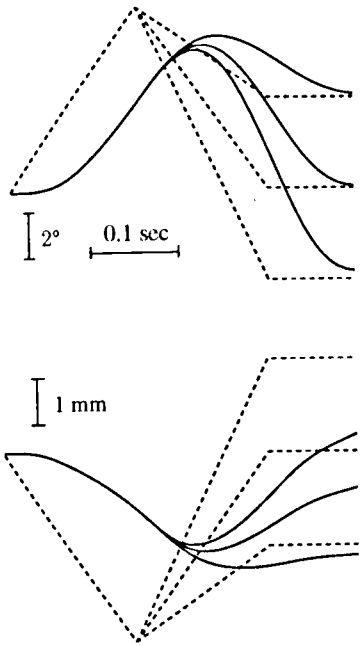


Figure 5: Predicted kinematic patterns of jaw rotation (above) and horizontal translation (below) and the presumed underlying control signals during a V_1CV_2 utterance. Note that the whereas the magnitude and the duration of the equilibrium shifts associated with V_1C transition are fixed, the predicted V_1C duration and movement amplitude vary with V_2.

REFERENCES

Abry, C. and T. Lallouache (in press). Propositions pour un modèle d'anticipation dependant du locuteur. données sur l'arrondissement du français. *Bulletin de la Communication Parlée*. Grenoble, France: Institut de la Communication Parlée, University of Grenoble.

Bothorel, A. (1975). Positions et mouvements de l'os hyoïde dans la chaîne parlée. *Travaux de l'Institut de Phonétique de Strasbourg 7*, 80–132.

Hogan, N. (1984). An organizing principle for a class of voluntary movements. *J. Neurosci. 4*(11), 2745–2754.

Laboissière, R., D. Ostry, and A. Feldman (submitted). The control of human jaw and hyoid movement. *J. of Neurophysiol.*.

Lindblom, B. (1983). Economy of speech gestures. In P. MacNeilage (Ed.), *The Production of Speech*, pp. 217–245. New-York, USA: Springer Verlag.

MacNeilage, P. (1970). Motor control of serial ordering of speech. *Psychol. Rev. 77*, 182–196.

Ostry, D. and V. Gracco (1995, June). Is intra-articulator speech coarticulation planned? *Paper presented at the 129th Meeting of the Acoust. Soc. Am.*. Washington, DC.

Ostry, D. J. and K. G. Munhall (1994). Control of jaw orientation and position in mastication and speech. *Journal of Neurophysiology 71*, 1528–1545.

Parush, A. and D. Ostry (1986). Superior lateral pharyngeal wall movements in speech. *J. Acoust. Soc. Am. 80*, 749–756.

Parush, A. and D. Ostry (1993). Lower pharyngeal wall movement in speech. *J. Acoust. Soc. Am. 94*, 715–722.

Parush, A., D. J. Ostry, and K. G. Munhall (1983). A kinematic study of lingual coarticulation in VCV sequences. *J. Acoust. Soc. Am. 74*(4), 1115–1125.

Perkell, J. and M. Matthies (1992). Temporal measures of anticipatory labial coarticulation for the vowel [u]: Within- and cross-subject variability. *J. Acoust. Soc. Am. 91*, 2911–2925.

TOWARDS A PHYSIOLOGICAL MODEL OF SPEECH PRODUCTION

R. Wilhelms-Tricarico and J. S. Perkell
M.I.T. - RLE 36-581, 50 Vassar Street, Cambridge MA 02139, U.S.A.

ABSTRACT

At the periphery, speech production is a biomechanical process that has acoustic effects. To help understand this process and its control, we propose a model of the vocal tract that is based on biomechanics and adaptable to the individual speaker's anatomy; we report about initial and planned efforts to realize such a model; and we outline a hierarchical, modular control structure that transforms a stream of articulatory and acoustic goals into physiological signals that drive the vocal tract model.

INTRODUCTION

Just as the legs are not primarily "designed" for ball dribbling in soccer, the body structures that participate in the physical execution of speech processes, such as the lungs, larynx, tongue, jaw, and lips, did not evolve primarily for speech production. Speech as a physical process is based on audible effects of the movements of a poorly-understood, complicated biomechanical structure, which performs many non-speech purposes and functions. As with control of the many different uses of the hands, the speech process may require a special control structure in the human brain that manages to make appropriate use of the intricate biomechanics of the speech production apparatus.

Physiologically based speech production research usually investigates the hypothesized control structures and attempts to explain observations and measurements of speech production processes, in terms of functional constructs such as compensatory articulation, coarticulation, motor equivalence and articulatory invariance. Unfortunately, the complexity of the biomechanical and neurological system that makes up the production apparatus has often been characterized by oversimplified models which make unsupported or unrealistic assumptions about the nature of articulatory dynamics and kinematics.

We argue that without taking biomechanics into account, many empirical observations of speech production can only be explained with ambiguity, because no objective criteria exist to establish a distinction between what are mechanical effects and what are effects of the control system. In order to overcome the difficulty of establishing more clearly the domain of the different system interactions that influence speech motor output, a physiologically- based vocal tract model will be built. The methods and the current state of the modeling are reported in this article. In addition, we continue a discussion about the structure of a controller, in order to be able to express current ideas about speech motor control in a computational model.

TASK DYNAMICS

The most advanced, comprehensive production model to date has been developed at Haskins Laboratories (see [14] for a review). The Haskins model forms the core of a "dynamical approach to gestural patterning in speech production", which attempts "to reconcile the linguistic hypothesis that speech involves an underlying sequencing of abstract, discrete, context-independent units, with the empirical observation of continuous, context-dependent interleaving of articulatory movements" [18]. The fundamental invariant unit is the abstract gesture. Combinations of abstract gestures underlie phonetic segments.

In this approach, a "task dynamic" model is the controller for an articulatory synthesizer, which is a geometrical model in the midsagittal plane. Computations of area functions are based on the articulator configurations, in combination with formulae derived from static three-dimensional data. An utterance-specific "gestural score" provides the input to the task-dynamic model in the form of sequences of activation pulses for the abstract gestures, following Browman and Goldstein [3].

In the task dynamic model, the formation and release of linguistically-significant vocal-tract constrictions are specified in a "tract variable" coordinate system. Articulatory movement is generated by modeling the influence of each discrete abstract gesture in this coordinate system as a time-invariant linear second order system. Since all dynamical properties of this system reside in the controller, the biomechanical properties of the vocal tract are not represented explicitly. The model accounts for coarticulation (as "coproduction" of sequences of (partly) overlapping abstract gesture complexes), and overlapping influences of multiple abstract gestures and tract variables on movements of individual articulators (as a result of "blending" of abstract gestures).

Limitations and Improvements

Perhaps the most important limitation of the Task-Dynamic Model is the concentration of its dynamical properties entirely in the abstract task space, when it is likely that a significant proportion of the kinematics of speech are determined by the anatomical and biomechanical properties of the peripheral production mechanism. An inadequacy of such simple task-space dynamics has also been shown in characterizing arm movements [9]. Many types of actual movement appear to be characterized by more gradually-increasing accelerations, and depending on the movement objectives, movements may be programmed (in part) according to optimization principles such as minimization of expended effort (cf. [13],[12]). If the non-linear biomechanical (dynamical) properties of the vocal tract were included in the model of the physical plant, then physically-based movement optimization criteria could be explored, simulations should be more accurate, and the form of the actual underlying control might be investigated with more revealing results.

Another limitation, less fundamental than the first, is seen in the almost axiomatic assumption of linear second order dynamics in the task space. Originally, it was proposed that in the task dynamic model, the programs of inter-articulatory coordination are "task-specific autonomous (time invariant) dynamic systems that underlie an action's form as well as its stability properties" (See Salzman and Munhall, [18] p. 337). This sufficiently general definition can, in principle, cover a large class of dynamic systems that may be needed to describe the dynamics of underlying coordinative structures. According to task dynamics, the movement from one segment to the next can be understood as a transition from the influence of one dynamic regime to the next. However, in the further development of the model, the understanding of task dynamics has been reduced by some to thinking of *second order* dynamic systems as the only possible model, resulting in a tendency towards over-simplification in which "everything" is to be accomplished by point attractors and limit-cycles. This limitation can be overcome by considering more general control systems, which by themselves can be dynamical systems, namely motor pattern generators. For modeling motor synergies, motor pattern generators have been proposed previously to simulate reflex behavior in animals (see [10]).

The task dynamics model also assumes generally that the movement goals of the abstract gestures are defined in terms of vocal-tract constrictions. However, recent motor equivalence studies of the vowel /u/ indicate that its goal may be defined more appropriately in terms of the acoustic transfer function [17]. Other sounds may also have goals that are defined primarily in acoustic terms [6].

Finally, it has been suggested that the establishment of sound categories is influenced partly by anatomy [14, 16].

Further, individual morphological differences between speakers may have influences on the motor-planning of individual speakers. To investigate such hypotheses, the current restriction to two-dimensional geometric vocal tract models needs to be overcome.

A BIOMECHANICAL MODEL

Our first step towards a biomechanical vocal tract model is a three- dimensional model of the human tongue (cf. [20]). Compared to previous work on 3-D finite element tongue models, the current work in progress entails a more accurate description of motion, by using large-strain finite elements and accounting for inertia of the moving structures.

So far, a simplified tongue model has been implemented and tested; the model consists of 42 elements, and contains eight tongue muscles. Fig. 1 shows the shape of the fixed reference configuration of the model tongue and the muscle fiber directions of the styloglossus muscle. Because of the lack of an accurate model of tongue tissue, a pragmatic phenomenological muscle model was adopted. The stress in the muscle tissue has an active and passive component. The passive stress is modeled as a nonlinear- elastic, linear-viscoelastic response of the tissue to deformation. The active component is computed by a stress production model that takes into account the elongation and the rate of elongation or shortening of the muscle fibers.

Computational methods

The application of the finite element method to the discretization of the equations of motion results in a system of differential equations which have to be solved. The system of equations relates the forces, displacements and accelerations at each node. The complete system has the following form:

$$M\ddot{u} + J(u, \dot{u}, \Pi) = B + T(u, \dot{u}) \quad (1)$$

In Eq. 1 the global node displacement vector u contains the displacement vectors of each node. The internal force vector J depends in a nonlinear manner on the global node displacement vector u, the global node velocity vector $\dot{u}$, and further, upon a multi-tuple of parameters Π which influence the constitutive behavior of the matter (strain–stress relation). The parameters Π are activation levels of the muscles in the model. Virtual forces are computed to maintain the incompressibility of the tissue, which is essentially a geometric constraint on the movement. This is done by computing a pressure field that varies over time but is spatially constant or varies linearly within each element. The pressure field is contained in the internal forces in equation (1). The right hand side of the equation is the system of external forces. Forces such as gravity, which act upon the whole body, are included in the vector B, and surface forces are represented by $T(u, \dot{u})$. The surface forces are responsible for constraining the model's movements geometrically. They include the forces resulting from intra-oral air pressure during closures and those forces acting on the tongue when it contacts and slides along surfaces such as the hard palate. In the current model surface forces have not yet been implemented. The details of the derivation and implementation of the computational methods are described in [20].

Since the time-dependent muscular activation levels modify the constitutive equations of the muscle tissue, they influence the stress field in the continuum, which is computed based on the instantaneous strain and rate of strain in the tissue. The computed stresses give rise to node forces. Thus, the node forces are a function of the deformation and of the muscle activation levels. The varying muscle activity levels constitute a multidimensional parametric control of the system.

The model has been used mainly to achieve an operational state of the computer code and to show the feasibility of the proposed methods, in that some typical movements of the tongue could be realized in simulation experiments [20].

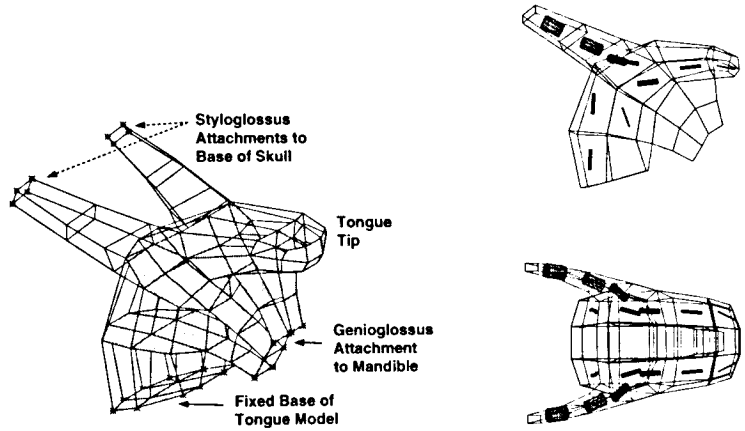


Figure 1: **Left:** *The tongue reference configuration with the initial placements of nodes. The nodes indicated with stars are fixed in the current implementation. They represent the connection of styloglossus muscles with the skull, the genioglossus with the mandible, and the base of the tongue in the plane of the hyoid bone. In the computational simulations, deformations of the tongue are computed as displacements of the nodes relative to their position in the fixed reference configuration.* **Right:** *Fiber directions of the styloglossus in lateral projection (above) and top view (below). Styloglossus fibers are specified on both sides in lateral elements. The long axes of the small cylinders represent the directions of the stress production by muscle fibers. The cylinders' diameter represent the relative amounts of maximal generated stress and can be interpreted as fiber density in the elements.*

CONTROLLER SCHEMES

Even at the current stage of development of the vocal tract model, it is possible to think about the general structure of a control mechanism that will compute the muscular activation time functions to steer the biomechanical model. This structure is not understood in terms of actual neurological functionality but rather as "software" which should simulate plausible functional components of the neurological controller. The controller transforms input signals that are described in terms of desired acoustical and/or articulatory goals into muscular activations which cause movement of the biomechanical plant.

Evidence for hierarchical organization

Perturbation experiments (cf. [2]) have provided evidence for the existence of mechanisms in speech production that use orosensory and internal feedback to control synergistic actions of multiple articulators for achieving functionally-specific goals. For example, if the lower-lip is unexpectedly impeded in its upward movement toward a bilabial closure for a /p/, the upper lip may move further downward than planned, with increased velocity to complete the closure (cf. [1]).

For articulators that are *not* biomechanically coupled, it has been shown [11] that the laryngeal articulation could be influenced by perturbing lip movements, further supporting the idea

that during speech production "coordinative structures" are programmed, which temporarily link articulators to achieve task specific goals. Weismer [19] describes this concept as follows: "Various articulatory gestures may be transiently linked to accomplish an articulatory goal, then unlinked as this goal expires and the next one arises." Such linking of articulators may involve the online use of both internal feedback and peripheral sensory information in the form of a "sensory template" that is specific for a motor control task. A sensory template is defined, according to Burgess [4] as "a central representation of the sensory receptor discharge that would be expected to occur during a movement if the movement is executed according to the plan".

Evidence of articulatory synergisms, *i.e.*, coordinative structures, also comes from motor equivalence experiments, which involve observation of many repetitions of the same behavior without the use of external perturbations. The term "motor equivalence" refers to the finding that the same goal is reached in more than one way (cf. [5]). Theoretically, across multiple repetitions, there can be trading relations (complementary covariation) in the relative contributions of: (1) multiple muscles to the same movement, (2) multiple movements to the same acoustically-critical vocal-tract cross- sectional area and (3) two area-function constrictions to the same acoustic transfer function.

Hypothetic controller structure

The movements to achieve sequences of articulatory and acoustic goals may be controlled by a hierarchical system that reduces the number of controlled degrees of freedom at each successively higher level.

The purpose of the hierarchical, modular controller is to control: multiple constrictions to determine the aerodynamics and acoustics of the vocal tract, multiple articulator movements for each constriction, and multiple muscles for each articulatory movement (see [15]). This hierarchy can be expressed by making the assumption that the controller has three hypothetical levels: The lowest level incorporates control structures, which generate synergistic muscle actions that result in simple gestural movements, such as raising the tongue blade or rounding the lips. The next level orchestrates the "elemental gestures" of the lower level to perform articulatory tasks that can be best described as creating vocal tract constrictions with certain characteristics (manners), for example creating an appropriate constriction for a vowel or producing a bilabial stop closure or a dento-alveolar constriction for a fricative. The third level, which orchestrates both lower levels, receives input signals that are described in terms of both desired acoustical consequences of articulation and/or as articulatory goals directly. The selection of acoustic goals and articulatory goals at the highest level comprises the translation of a hypothetical symbolic representation of speech into control actions.

Jordan and Rumelhart's distal learning strategy will be used as a paradigm for the implementation of each level of the controller, starting at the lowest level. The psychological and neuro-physiological idea of the internal model, or efferent copy, appears in this strategy as a forward model. The tentative general structure, shown in Fig. 2, has two components. One component (the controller C) maps from "intentions" (i) to motor commands (u), and the other component, called the forward model (FM), from motor commands to predicted sensations ($\hat{s}$). The forward model is trained using the difference between predicted sensation and actual sensation (s), which arise in the plant (P) as the result of the controlled actions. The composite system (C and FM) is trained using the difference between the desired sensations (d) and the actual sensation. See Jordan and Rumelhart [8] for further details.

In this context, Fig. 2 is a sketch of the first level controller. Since the biomechanical plant (P) is a dynamic system, the internal model of the lowest control level will also be a dynamic system (but without neural transmission and biomechanical response delays). The plant transforms the motor control input u and its current state x into two types of sensory results, s, and

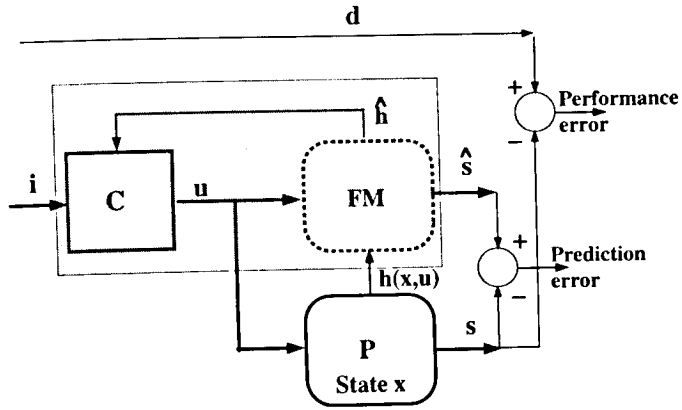


Figure 2: *The composite learning strategy (adopted and modified from Jordan, Flash and Arnon, 1993). The controller (C) transforms intentions (i) into motor commands (u). The forward model (FM) predicts $\hat{s}$ as the sensory result of the action u on the plant. The difference between the predicted sensation $\hat{s}$ and the actual sensation s, the prediction error, is used to optimize the forward model during training. The forward model is sensitive to the state of the plant; it receives input $h(x,u)$ that contains information about the state (for example muscle length information). Feedback delays of plant outputs are not shown.*

h(x,u) which makes the plant's state x (consisting of all displacements and velocities) partially observable by the forward model. The signal h(x,u) may contain measures of the length and rate of change of length of the muscles. The forward model learns to map motor commands (u) and current observables of the state (h(x,u)) into estimated sensory output ($\hat{s}$). It further learns to predict the development of relevant observables that are related to the plant state, shown as $\hat{h}$. Once the forward model is trained, the actual controller (C) relies on the estimated information ($\hat{h}$) about the state of the plant. This amounts to "internalizing" a feedback loop in the controller-forward model composite system. The purpose of the internal model is to "mimic" aspects of the plant, and to represent sufficient information to allow predicting the result of a control action on the plant. This information is represented in the forward model as a (learned) mapping from the current input u and the current state of the forward model $\bar{x}$ to the delayed sensory output $\hat{s}$ and the delayed estimated observables $\hat{h}$. The state $\bar{x}$ of the forward model is not necessarily an estimate of the actual state of the plant. For instance, if the forward model is implemented as a neural network structure, its state variables do not correspond to the state variables in the biomechanical model. However, the network learning should result in an input-output behavior that resembles the input-output behavior of the biomechanical model.

In view of the complexity of the biomechanical model and the speech motor control task, it will certainly be necessary to subdivide the overall control problem on each level into smaller ones. Subdivision on the lowest level is particularly sensible because the biomechanical plant consists of parts (e.g. the tongue body, tongue blade, mandible, lips, velum) that can act quasi-independently. Subdivision at the next level is motivated by the possibility for control of constrictions at different locations along the vocal tract with different articulators and manners. Another motivation for sub-

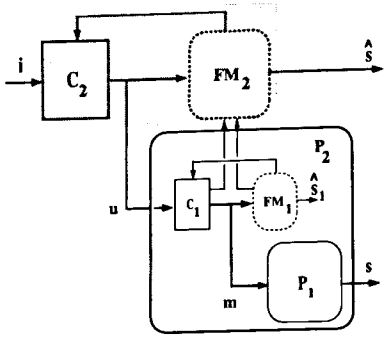


Figure 3: *The second level of the controller, consisting of the controller C_2, generates higher level motor commands u and controls the plant P_2 which is an encapsulated system consisting of the biomechanical model and the lowest level controller C_1. The biomechanical plant P_1 is driven by muscle activation functions m which are generated by the lower level controller C_1.*

division into subcontrollers for special tasks, in particular on the second level, is seen in recognized structures for interarticulatory coordination, such as between jaw and lips. Jordan and Jacobs [7] have extended their previously proposed competing experts paradigm to constitute a hierarchical architecture [the "hierarchical mixture of experts" (HME) paradigm] that allows such a subdivision of control problems.

By starting to build the controller from the bottom to the top, from simple movement models to complex movement models, each level of the controller is designed to achieve a reduction of complexity and degrees of freedom for the higher level controllers. For example, (sub-)controllers in the second level do not have to care about and operate without knowledge of individual muscle states, since that knowledge is incorporated into the lowest level of control. The internal model built into the lowest level of the control includes a partial (but sufficient) representation of the biomechanical model. The next higher level operates on an "encapsulated" lower level and its internal model includes a representation of the effects of combined generalized motor commands (such as "tongue tip raising" and "tongue back lowering") on sensory output, but only to the extent as it is relevant for the second level. Stated differently, the higher levels of the controller receive more general intentions, issue more global motor commands, and result in more abstract sensations. Fig. 3 sketches this arrangement for the two lower levels. The controller of the second level C_2 controls the augmented plant P_2 which consists of the lower level controller C_1 and the biomechanical model. Before controller C_2 can be trained properly, the forward model FM_2 of the second level needs to be trained to include a partial representation of the augmented system P_2, including the prediction of state-related information of the controller in P_2. The third level of the proposed hierarchy will operate on a plant formed by encapsulating the presented structure, and augmenting it further by adding another level of acoustical ouput resulting from computations in an extended biomechanical and acoustical model.

CONCLUSIONS

We have outlined a physiologically based speech production model and have reported the first steps taken in its implementation. Development of the biomechanical and control models will be coupled closely to the morphological (MRI), kinematic and acoustic data from individual speakers. A reasonably faithful model of the vocal tract biomechanics coupled with a pragmatically motivated, hierarchical and modular control structure, should permit investigations that allow greater insight into the actual underlying control strategies.

REFERENCES

[1] Abbs, J. H., Gracco, V. L., and Cole, K. J. (1984) Control of multimovement coordination: Sensorimotor mechanisms in speech motor programming. *J. Motor Behavior*, 16(2):195–231.

[2] Abbs, J.H. and Gracco, V.L. (1984) Control of complex motor gestures: orofacial muscle responses to load pertubations of lip during speech. *Jounal of Neurophysiology*, 51:705–723.

[3] Browman, C.P. and Goldstein, L. (1992) Articulatory phonology: An overview. *Phonetica*, 49:155- 180.

[4] Burgess, P.R. (1992) Equilibrium points and sensory templates. *Behavioral and Brain Sciences*, 15(4). Commentary to Bizzi et al. (1992), Does the nervous system use equilibrium-point control to guide single and multiple joint movements?

[5] Hughes, O.M. and Abbs, J.H. (1976) Labial- mandibular coordination in the production of speech: Implications for the operation of motor equivalence. *Phonetica*, 33:199–121.

[6] Johnson, K., Ladefoged, P., and Lindau, M. (1993) Individual differences in vowel production. *J. Am. Soc. Acoust.*, 94(2):701–714.

[7] Jordan, M. I. and Jacobs, R. A. (1993) Hierarchical mixtures of experts and the EM algorithm. Computational Cognitive Tech. Rep. 9301, MIT.

[8] Jordan, M.I. and Rumelhart, D. E. (1992) Forward models: Supervised learning with a distal teacher. *Cognitive Science*, 16:307–354.

[9] Katayama, M., and Kawato, M. (1992) Visual trajectory and stiffness ellipse during multi-joint arm movement predicted by neural inverse models, *Technical Report*, ATR Laboratories, Kyoto, Japan.

[10] Liaw, J-S., Weerasuriya, A., and M.A., Arbib. (1994) Snapping: A paradigm for modeling coordination of motor synergies. *Neural Networks*, 7(6/7):1137–1152.

[11] Munhall, K.G., Löfqvist, A., and Kelso, J. A. S. (1994) Lip-larynx coordination in speech: Effects of mechanical pertubations to the lower lip. *J. Acoust. Soc. Am.*, 95(6):3605-3616.

[12] Munhall, K.G., Ostry, D.J., and Parush, A. (1985) Characteristics of velocity profiles of speech movements *J. Exp. Psych.*, 11:457–474.

[13] Nelson, W.L. (1983) Physical principles for economies of skilled movement, *Biological Cybernetics*, 46:135–147.

[14] Perkell, J.S. (1991) Models, theory and data in speech production. In *XIIth International Congress of Phonetic Sciences*, volume 1, pages 182–191, Aix-en-Provence, France.

[15] Perkell, J.S. (1994) Articulatory processes. To appear in J. Hardcastle and J. Laver (eds.), A Handbook of Phonetic Science.

[16] Perkell, J.S. (1995) Properties of the tongue help to define vowel categories: Hypotheses based on physiologically-oriented modeling. *Journal of Phonetics.* in press.

[17] Perkell, J.S., Matthies, M.L., Svirsky, M.A., and Jordan, M.I. (1992; in press) Goal-based speech motor control: A theoretical framework and some preliminary data. J. Phonetics.

[18] Saltzman, E.L. and Munhall, K.G. (1989) A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, 1:333–382.

[19] Weismer, G. (1988) Speech production. In *Handbook of Speech-Language Pathology and Audiology*, chapter 8. B.C. Decker Inc.

[20] Wilhelms-Tricarico, R. (1995) Physiological modeling of speech production: methods for modeling of soft-tissue articulators. *J. Acoust. Soc. Am.*, in press.

ARTICULATORY CO-ORDINATION AND ITS NEUROBIOLOGICAL ASPECTS: AN ESSAY

Shinji MAEDA
Centre National de la Recherche Scientifique, URA 820
and Ecole Nationale Supérieure des Télécommunications, Département SIGNAL
(46, rue Barrault, 75634 Paris, France)

Kiyoshi HONDA
ATR Human Information Processing Research Laboratories
(2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02 Japan)

ABSTRACT

The aim of this talk is to describe a jaw-tongue articulatory compensation during unrounded vowels. Simulation experiments using an articulatory model indicated that an optimum compensation can be directly specified by linear relationships between these two articulators' positions. As if a speaker knows how to compensate the position of the two articulators, the linear relationships were very closed to those actually observed in the X-ray film data of the speaker. As implications of this finding, we propose a feedforward control model for compensatory articulation and then discuss on the organizations of motor control at different levels.

INTRODUCTION

In our previous factor analysis of X-ray film data showed that the tongue contours in the mid-sagittal plane can be described by four orthogonal components, *i.e.*, "articulators" [1]. Since the contribution of each component upon the tongue contours is specified by a single parameter, the observed contours are determined by the following four parameters; an extrinsic parameter, the jaw position (**jw**), and three intrinsic ones, tongue-body position (**tp**), tongue-body shape (**ts**), and tongue-tip position (**tt**). The statistics indicated that these four parameters explain more than 90% of the variance of the observed tongue contours. In this method, the set of parameter values can be calculated from an observed tongue shape. Thus, the variation of the tongue shapes along sentences is described by the corresponding frame-by-frame variations, *i.e.*, temporal patterns, of these articulatory parameters.

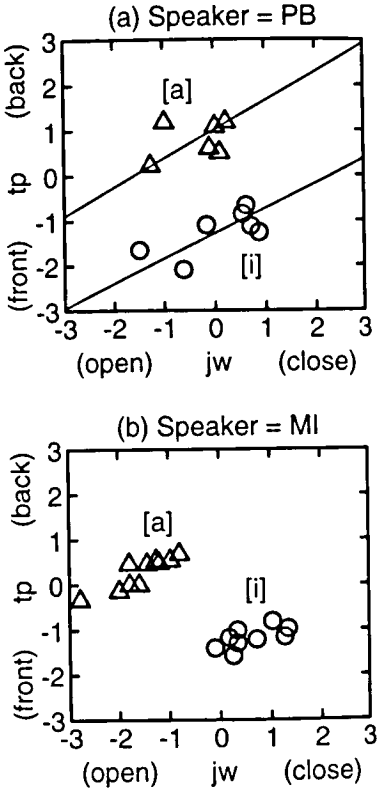


Figure 1. *Scattergrams of jaw (jw) vs. tongue-body positions (tp) at the "targets" for /i/ (circles) and /a/ (triangles) determined from the measured movements of the articulators in sentences read by PB in (a) and in nonsense words read by MI in (b).*

One of findings was that the temporal patterns exhibited a large contextual variability for the same vowel. In fact, this contextual variability seemed to be far greater than the inter-speaker variability for the same vowels. Figure 1a shows scattergrams on the **jw-tp** space concerning with two extreme unrounded vowels, /i/ and /a/, extracted from the different phonetic contexts in the 10 short French sentences uttered by a speaker (PB) [2]. Similar scattergrams are shown in Figure 1b where the data points were derived from X-ray film data for /pV1CV2/ nonsense words read by another speaker (MI). Those data points in the figures correspond to "targets" defined as turning points along the jaw-tongue trajectories during the vowels. The unit of both **jw** and **tp** is standard deviations. The value of these articulatory parameters rarely exceeds plus and minus three standard deviations. It is noticed that the points corresponding to each of the two vowels covers a large range. The dispersions are so large that jaw position or tongue position alone cannot distinguish these two extreme vowels, in particular for the speaker PB, exhibiting a significant degree of the context dependent variability in jaw and tongue positions for the same vowel. Nevertheless, the scattered points for the two vowels in both figures are well separated from each other.

When the first and second formant frequencies (F1 and F2, respectively) were calculated from those parameters, the corresponding F1-F2 scattergrams indicated a tighter pattern in comparison with the articulatory scatters [2]. We, therefore, recognize this large variability as a manifestation of the compensatory maneuver by the speakers. As the consequence of this, an unrounded vowel is not specified by a single set of parameter values, **jw** and **tp**, but rather by a compensation rule which determined (infinite) possible combinations of parameter values appropriate to produce that vowel. So, the question we ask is how we can describe this inter-articulator compensatory co-ordination.

We note in passing that for rounded vowels, such as /u/ and /o/, the compensatory maneuvers between the jaw and the intrinsic lip aperture (not tongue-body position as in the case of unrounded vowels) are predicted from simulation experiments with the articulatory model. We don't have sufficient number of rounded vowel tokens to validate the prediction, however. For this reason, we deal only with unrounded vowels in this paper. Moreover, we deal with only two parameters, **jw** and **tp**, since they are dominant in the determination of the tongue contours, especially for unrounded vowels (see the paper by Bouabana & Maeda in this congress.) Also we deal with only the speaker PB, since the complete articulatory model, which enables us to calculate acoustic characteristics, is available only for that speaker.

AN OPTIMUM JAW-TONGUE COMPENSATORY RULE

The target points associated with each vowel seem to exhibit a linear relation. Notice that the target points for each vowel in Figure 1a are scattered around the straight line which corresponds to the first principal axes. This means that it is possible to predict, although approximately, the tongue position for a vowel by a linear function of the actual jaw position or *vice versa*. The straight lines in Figure 1a were determined by the principal component analysis on a small number of points, for example only seven points for /i/ and six points for /a/ in the case of PB. These lines, therefore, are not necessarily "optimum" as compensatory rules. The term "optimum" means that a rule defining the jaw-tongue co-ordination results in the minimum of F1-F2 variability, when **tp** and **jw** are varied according to the rule. We thus carried out a simulation experiment, independently of the data, to determine optimum rules, *i.e.*, the linear relationships, for these two vowels and plus the posterior vowel /ɑ/.

Let us empirically assume that **tp** is linearly related to **jw** as follows:

$$\mathbf{tp} = a\,\mathbf{jw} + b, \qquad (1)$$

where "a" is a slope coefficient and "b" is the intercept. There is no mathematical reason to assume **jw** as a function of **tp**. It appears to us more reasonable to assume that tongue position is a function

of jaw and not *vice versa*. Let us denote the intercept as **tp0**. Then, the linear equation becomes

$$\mathbf{tp} = \mathbf{a}\,\mathbf{jw} + \mathbf{tp0}. \quad (2)$$

For given values of "a", a change in **tp0** (front/back tongue position) would result in a change in the phonetic value of vowel. It is easy to see, for example in Figure 1a, that varying **tp0** would vertically slide the straight line up and down, resulting in a change in the phonetic value of vowel, such as between /i/ and /a/.

It seems then reasonable to explore the variation of the acoustic variability index in function of the slope "a" for a fixed **tp0** value appropriate for each of the vowels /i/, /a/, and /ɑ/. Since the optimal rules should remain at the vicinity of the data derived principal axis, as already shown in Figure 1a, we used the following somewhat *ad hoc* search scheme to find out optimum rules. For a specified intercept **tp0**, we calculate variability index varying the value of "a" to determine its optimum value that results in the minimum index value. We then verify whether the determine value of "a" is also optimum along that line. To do so, we select anchor points, (**jw1**, **tp1**) along the determined line, such that $\mathbf{tp1} = \mathbf{a}\,\mathbf{jw1} + \mathbf{tp0}$, where **jw1** = -2, -1, 0, 1, and 2. The index values, therefore, are calculated varying the slope "a" for the five different straight lines defined as

$$\mathbf{tp} = \mathbf{a}\,\mathbf{jw} - (\mathbf{a}\,\mathbf{jw1} - \mathbf{tp1}). \quad (3)$$

If the acoustic variability index determined with **jw1** = 0 is always smallest, we could conclude that the determine rule is indeed optimum and that the straight line is a reasonable specification of the compensatory rule.

Now let us define the acoustic variability index. We use an averaged normalized variance of F1 and F2 frequencies as described as follows:

$$v = 100\sqrt{\frac{1}{2}\left(\frac{\sigma_1^2}{\sigma_{1_{max}}^2} + \frac{\sigma_2^2}{\sigma_{2_{max}}^2}\right)} \quad (\%) \quad (4)$$

where σ_i^2 (i = 1, or 2) is the variance of F1 or F2, respectively. The first and second formant frequencies are calculated using the articulatory model. The model specifies the vocal tract shapes with seven parameters including **jw** and **tp**. The value of **jw** is always varied from -3 to +3 with 1 standard deviation as the step size and the corresponding value of **tp** is determined by Eq. (3). Values of the remaining parameters, **ts**, **tt**, two lip parameters, and larynx position, are determined from the X-ray data frames. The area function and then formant frequencies were computed from the model-specified vocal-tract shape. The acoustic calculations take into account the effects of the non-rigidity of the tract walls and of sound radiation from the lips. The F1 (and F2) variance, finally, is computed from the seven different combinations of **jw** and **tp** associated with the straight-line rule. An example of such calculations is illustrated in Figure 2 for vowel [i].

The maximum possible variances of F1 and F2, $\sigma_{1_{max}}^2$ and $\sigma_{2_{max}}^2$ respectively, needed for the normalization are difficult to estimate. We, therefore, employed the values of half range of F1 and of F2 frequencies derived from the articulatory model. The half ranges, 300 Hz for F1 and 1000 Hz for F2, were determined from F1-F2 projections obtained by systematically varying the values of seven model parameters [3]. The result of the index calculations for the three vowels are shown in Table 1.

For the vowel /i/ shown at Table 1a, we choose the value of **tp0** as -1.3 at **jw0** = 0 based on the observation of the scattered points in Figure 1a. As described before, we calculated, first, variability index varying the slop from 0.3 to 0.8. The results are shown at the column, **jw1** = 0 and **tp1** = -1.3, framed vertically. The minimum variability occurs at slope ("a") equal to 0.6. Second, we selected four anchor points along the determined straight line rule, *i.e.*, **tp** = 0.6 **jw** - 1.3, which are listed at the top two rows. Note that variability indexes for a = 0.6, framed horizontally, have the identical value, since they are derived with the same straight line rule. It

can be seen that acoustic variability is always minimum at "a" = 0.6 regardless of anchor points. Although the variability never becomes zero, in other words, the compensation isn't perfect, the linear relation appears adequate to specify the compensatory rule. It is noted, moreover, that the determined optimum value of slope, 0.6, favorably compares with the slope value calculated from the data points (shown in Figure 1a), 0.56. This good agreement suggests that the speaker knows an acoustically optimum way of co-ordinating the jaw and tongue-body movements in the production of this vowel in different phonetic context.

Table 1. *F1-F2 variability index (in %) for three vowels as a function of the slope coefficient "a" calculated at different anchor points (**jw1**, **tp1**) using Eq. 3.*

(a) Vowel /i/, where **tp1** = 0.6**jw1** - 1.3

jw1	-2	-1	0	1	2
tp1	-2.5	-1.9	-1.3	-0.7	-0.1
a					
0.3	127.5	96.6	23.1	19.3	16.9
0.4	95.4	15.7	14.4	12.8	11.9
0.5	7.5	6.9	6.8	6.7	6.4
0.6	4.6	4.6	4.6	4.6	4.6
0.7	9.8	10.7	10.4	10.6	11.4
0.8	15.8	16.7	17.9	23.5	100.8

(b) Vowel /a/, where **tp1** = 0.5**jw1**+ 0.9

jw1	-2	-1	0	1	2
tp1	-0.1	0.4	0.9	1.3	1.9
a					
0.3	16.2	14.7	14.1	14.0	12.6
0.4	11.4	11.3	10.9	10.9	10.6
0.5	9.4	9.4	9.4	9.4	9.4
0.6	9.8	9.9	9.6	9.5	9.5
0.7	12.4	12.1	11.8	11.3	11.2
0.8	15.9	15.3	15.0	14.4	14.3

(c) Vowel /ɑ/, where **tp1** = 0.4**jw1**+ 2.6

jw1	-2	-1	0	1	2
tp1	1.8	2.2	2.6	3.0	3.4
a					
0.2	12.2	11.9	11.4	10.9	10.3
0.3	10.0	10.0	9.7	9.8	9.8
0.4	9.5	9.5	9.5	9.5	9.5
0.5	11.6	11.0	10.9	10.5	10.3
0.6	16.1	14.4	11.3	12.4	11.5
0.7	24.9	20.4	16.5	14.8	13.8

The calculations concerned with the vowel /a/, shown in Table 1b, indicate that variability index becomes minimum at the slope value equal to 0.5. This is always the case regardless of anchor points. However the optimum slope, 0.5, compares less favorably with that determined from the data (shown in Figure 1a), 0.63. This discrepancy might be due to, in part, the small number of the data points with a relatively large dispersion of the measured data points. In order to assess a general trend of the optimum slope in a function of **tp0**, *i.e.*, depending on the vowel identity, we calculated variabilites for the posterior vowel /ɑ/. The result is shown in Table 1c which indicates the optimum slope value of 0.4. From these calculations, the general trend appears to be a regular decreasing of the slope value from the front to back tongue-dorsum position.

Figure 2 shows the vocal-tract profiles and frontal lip shapes at the left and the corresponding transfer functions at the right for the vowel /i/. In this calculation, the value of jaw position was also varied from -3 to +3 with the step size of one standard deviation, as described before. The corresponding variations in the profile and vocal-tract transfer functions are indicated by the arrows. We used the jaw-tongue compensatory rule with the optimum slope that equals to 0.6. Since only the tongue-body position is compensated against a change in the jaw position, the dimensions of the lip tube, modelled by a uniform elliptic cylinder, vary significantly depending on the jaw position. Acoustically speaking, a decrease in the pharyngeal cavity volume is not important for F1 frequency, if that variation is compensated by the narrowing of the mouth channel (F1 of /i/ corresponds to the Helmholtz resonance.) Such an F1 compensation just occurs with "a" = 0.6. The F1 deviation due to the closing jaw is "corrected" automatically, but the pharyngeal cavity length is not. The non-corrected length change is the cause of the F2 drift toward lower frequencies, as seen in Figure 2.

When the vocal-tract transfers of the vowel /a/ were calculated with the optimum slope value of 0.5 and the intercept of 0.9, relatively large F2

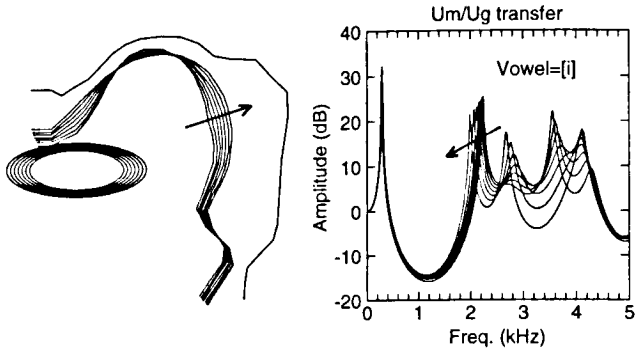


*Figure 2. Calculated VT profiles and frontal lip shapes at the left, the corresponding VT transfer functions at the right for the vowel /i/: (a)The tongue-body position is varied according to the compensatory rule, **tp** = 0.6 **jw** - 1.3. The arrows indicate the direction of variations from **jw** = -3 (close) to +3 (open).*

variation and, to a lesser extent, F1 variation occurs for this vowel. These variations are due to the lengthening of the front and back cavities combined with the narrowing of the lip aperture as the jaw position varies from low to high position. As consequence of this, the tongue-body compensation maneuver is less effective than in the case of the previous vowel /i/. This fact is manifested in the variability index value of 4.6% for /i/ and that of 9.4 % for /a/. The posterior vowel /ɑ/ exhibited the F1 and F2 variations similar to the vowel /ɑ/. The variability index at the optimum condition (a = 0.4 and **tp0** = 2.6) was 9.5%.

Speech synthesis experiments confirmed, in particular for the vowel /i/, a stability of the perceptual phonetic value under the jaw-tongue compensation specified by the linear relation, even when jaw position was varied covering its maximum range. The vowel /a/ also exhibits the stability, but as expected, within a narrower range of the jaw position variation than for the /i/ vowel.

VOWEL SPECIFICATION AND MOVEMENT CONTROL: DISCUSSIONS

It is well known in the domain of motor control that there are two kinds of movements, fast and slow. The duration of fast movements is 200 ms or less. Articulatory movements, therefore, belong to fast movements. Since the latency in the neural transmission of sensory information can be more than 200 ms or longer, there is no way to adaptively control articulatory movement by means of sensory feedback. The present compensatory rule enables us to postulate a simple feedforward "compute-and-control" scheme without feedback. The term "compute" means here a simple arithmetics such as Eq.2 defined before. A feedforward scheme for tongue-body position control is conceptually illustrated in Figure 3. The input to the control "box" is assumed to be the value of invariant target or command associated with a vowel and the current jaw position as an "afferent" information. We don't think there exists an invariant jaw position for a given vowel. Jaw position is influenced by stress and suprasegmentals as well. Here we assume that the position is provided somehow. Recalling the compensatory rule Eq.2, the input to the tongue position would be the values of the intercept and the slope, *i.e.*, **tp0** and "a", for unrounded vowels. The actual tongue-body position is calculated with rule. If the slope value can be approximated by a single value for all unrounded vowels or predicted from the intercept itself, only the specification of **tp0** where its value depends on the vowel identity is necessary. It is noted that the outlined control scheme should be recognized as a feedforward or an open-loop control,

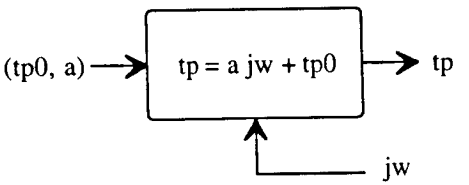


*Figure 3. A conceptual model of feedforward control for tongue-dorsum position **tp**. The box can be thought of as a "motor program". The program receives two inputs; "command" (**tp0** and the slope 'a') and an "afferent" information about jaw position, **jw**.*

since the sensory information comes from the jaw and not from the tongue itself.

The compensatory rule was derived from position data and naturally it relates between two positions. An obvious question is whether or not the positional variables have neurobiological reality as a control parameter of the muscles. There seems to be no definitive answer yet in the literature about the neural coding of positions. It should be mentioned, however, that in the tongue system, the extrinsic muscles seem to be organized in a clean way for speech production: the genioglossus posterior and hyoglossus form an antagonistic pair controlling high-front to low-back orientation of the tongue movements. The styloglossus (SG) and genioglossus anterior (GGa) are also considered antagonistic; the SG pulls the tongue body toward the high-back direction, whereas the GGa compresses the tongue in the opposite direction [4]. A vowel, then, is produced by selecting and activating one of the paired muscles, which is termed as "muscle group selection (MGS)" [5 & 6]. This kind of control organization can correspond to an elementary motor program.

In the muscular system, moreover, the contractile force and position (displacement) are related to each other, as a gross approximation, by a proportional principle (Hooke's law) [7]. In fact, our simulation experiments have shown that the values of positional articulatory parameters, **tp** and **ts**, can be determined, by the linear law, from the EMG activity patterns of these paired muscles. The F1-F2 patterns calculated from EMG data formed a reasonable English vowel pattern [8]. It is then safe to state that the output of the arrow-and-box control model depicted in Figure 3 can be understood as a net force to be generated by the selected group of the muscles in order to produce an intended vowel. Whereas the input can be regarded as the corresponding invariant specification of tongue position in terms of force. It follows that the control model situates between the motor organisation at the low-level, *i.e.*, the level of MGS or of elementary motor programs, and that at the high-level, where mappings between invariant auditory and articulatory representations of speech occur [6]. Thus the proposed control model could be considered as a motor program operating at the intermediate level, interfacing the high and low level motor organizations.

Although many details must be worked out, such a control model has the following three attractive characteristics:

(1) The model can explain the mechanisms of the bite-block vowels in straight forward manner, say, "compute and control" instead of "simulate and control" proposed by Lindblom [9]. It is noted that a series of papers has been published concerning with the "motor equivalence" that explains the jaw and lip co-ordination to maintain relatively invariant lip aperture regardless of a relatively large individual variations of these two articulators ([10] & [11] to mention a few). Motor equivalence was postulated from observations such that when the jaw contributed a large displacement to the opening or closing of the oral cavity, the upper lip and lower lip contributed proportionally less and conversely. It assumes the existence of afferent pathways, from the jaw to the lips, that provide information to adjust lip position by means of an open-loop control. Neurological evidence of such

open-loop control mechanism is found in the vestibulo-ocular reflex (VOR) [12]. VOR contributes to automatically stabilize retinal image against fast head movements. Its reaction time can be as fast as 10 ms. It should be noticed that there is a distinctive difference between the motor equivalence and our proposed feedforward control. The motor equivalence automatically operates as reflex and its control is not intentional. It contributes to the "correction" against local perturbation. Our feedforward scheme is to control an intended articulation. Thus we think that our proposed feedforward control operates at a higher level, presumably, at the level of the motor programming in the motor cortex and in the cerebellum.

(2) The model can handle the reflex reversal in speech production. Kelso *et al.* [13] reported that when jaw position had been unexpectedly perturbed during the vowel in /baeb/, the subject adjusted the lips, whereas during /baez/, the tongue was adjusted. This phenomenon can be explained by assuming that the selected motor program, such as depicted in Figure 3, activates the jaw-tongue sensory pathway, for example, in /baez/ but not in /baeb/.

(3) Since the final (and observable) tongue position is calculated as a function of the actual jaw position in the down stream using the explicit compensatory rule, the specification of tongue position for a given vowel in the motor program can be invariant as described above.

It is quite natural for one to raise a question about the neurobiological validity of such control model. To our knowledge there is no neurobiological evidence directly supporting the proposed control model. It is only possible to infer the model reality from the known functions of the human neurobiological system. The jaw-tongue and jaw-lip co-ordination including sensory pathways should functionally exist, at least, for vegetative functions such as food intake. Chewing food requires an intricate co-ordination between jaw and tongue movements. However, it is not clear whether or not the same co-ordinative mechanism is employed for speech. If the jaw-tongue co-ordination is acquired during learning speech, it might be the case that the basic (and innate) co-ordinative mechanisms including the selective sensory activation are specifically tuned for speech production, creating a new speech motor program. Learning the speech skill can be viewed as the process of such a "turning". It is known that the cerebellum plays a crucial role in learning of motor skill. It is suspected then that speech motor programs, at least part of them, are memorized in the cerebellum and they contribute to the formation of speech motor commands. A motor program is an abstract and functional concept, however, and consequently it could be wrong to consider that the speech motor programs and their functions are concentrated only in the cerebellum.

In summary, with the feedforward control of tongue movement, vowels can be specified by invariant muscular activity patterns (the input to the control system) which undergo lawful modifications according to the compensatory rule. Then a plausible scenario would be that in the stage of learning speech, the control process including feedback control play an important role in the speech production. But once the skill is acquired, the mode is transferred to the feedforward control discussed here. Or rather, when such change in the mode of control occurs, we recognize that a speaker has mastered how to speak.

ACKNOWLEDGEMENT

The first author acknowledges the support of the European project Esprit/BRA, N° 6975, SPEECH MAPS.

REFERENCES

[1] Maeda, S. (1990). Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal tract shapes using an articulatory model. In *Speech Production and Speech Modeling* (W.J. Hardcastle & A. Marchal, editors), 131-149. Kluwer Academic Publishers.

[2] Maeda, S. (1991). On articulatory and acoustic variabilities. *Journal of Phonetics*, **19**, 321-331.

[3] Boë, L-J. (1993). Speech Maps Interactive Plant "SMIPS". *Deliverable 6* (pp.3-20) of SPEECH MAPS (ESPRIT/BR N° 6975).

[4] Kusakawa. N., Honda, K. & Kakita, Y.(1993). Construction of articulatory trajectories in the space of tongue muscle contraction force. *ATR Technical Report*, TR-A-0171 (in Japanese).

[5] Honda, K., Kurita, T., Kakita, Y. & Maeda, S. (in press). Physiology of the lips and modeling of lip gestures. *Journal of Phonetics*.

[6] Honda, K. (in press). Organization of tongue articulation for vowels. *Journal of Phonetics*.

[7] Bouabana, S. & Maeda, S. (1994). Modélisation des mouvements articulatoires par la méthode de la LPC multi-impulsionnelle. presented at Troisième congrés français d'acoustique (Toulouse), in *Journal de Physique*, **4** (Colloque N° 5, Suppl. JP III, N° 5), 449-452.

[8] Maeda, S. & Honda, K. (1994). From EMG to formant patterns of vowels: the implication of vowel spaces. *Phonetica*, **51**, 17-29.

[9] Lindblom, B., Lubker, J., & Gay, T. (1979). Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation. *Journal of Phonetics*, **7**, 147-161.

[10] Folkins, J.W. & Abbs, J.H. (1975). Lip and jaw motor control during speech: Responses to resistive loading of the jaw. *Journal of Speech and Hearing Research*, **18**, 207-220.

[11] Gracco, V.L. & Abbs, J.H. (1988). Variant and invariant characteristics of speech movements. *Experimental Brain Research*, **65**, 156-166.

[12] Ito, M. (1974). The control mechanisms of cerebellar control systems. In F. O. Schmitt & F.G. Warden (Eds.) *The Neuroscinces Third World Study Program*, MIT Press, 293-303.

[13] Kelso, J.A.S., Tuller, B., Vatikiotis-Bateson, E., & Fowler, C.A. (1984). Functionally specific articulatory cooperation following jaw perturbations during speech: Evidence for coordinative structures. *Journal of Experimental Psychology: Human Perception and Performance*, **10**, 812-832.

INTERGESTURAL TIMING IN SPEECH PRODUCTION: DATA AND MODELING

Elliot Saltzman

Haskins Laboratories, 270 Crown Street, New Haven, CT USA; Center for the Ecological Study of Perception and Action, University of Connecticut; Storrs, CT USA

ABSTRACT

The task-dynamic model of speech production is reviewed, and current developments that incorporate a recurrent network dynamics of intergestural timing are described. Four sets of data are considered that provide a set of temporal patterning benchmarks for the evaluation of any model of intergestural dynamics. These benchmarks are being used to evaluate and to guide the development of the extended task-dynamic model.

INTRODUCTION

This paper examines the nature of intergestural timing in speech production from both a theoretical and an empirical perspective. The theoretical focus is on how a model developed previously in our laboratories, the task-dynamic model, is being extended to incorporate a dynamics of intergestural timing. In the extended model, a recurrent, connectionist network is used to simulate these dynamics. This network plays the role of a utterance-specific central timing network for speech that is bidirectionally coupled to a simplified set of speech articulators. The outputs of the network serve to parameterize the dynamics of the articulators, whose ongoing state is fed back to the network in a manner that modulates its timing. As this work is in its preliminary stages (Saltzman, Mitra, Levy, & Hogden, in progress), discussion will be limited to the directions being taken in these developments.

The empirical focus of this paper is a review of data from four classes of intergestural timing phenomena that provide a significant set of constraints that must be satisfied by this or any model of speech intergestural timing. These data thus serve to establish a set of benchmark criteria in the temporal domain, according to which different models can be evaluated.

THE TASK-DYNAMIC MODEL

Work on the task-dynamic model of speech production has been conducted in collaboration with several of our colleagues at Haskins Laboratories as part of an ongoing project focused on the development of a gesturally-based, computational model of linguistic structures (e.g., [1], [2], [3], [4], [5], [6]). The focus of this model is on the control schemes that underlie gestural patterning, and represents an attempt to reconcile the linguistic hypothesis that speech involves an underlying sequencing of abstract, context-independent units, with the empirical observation of context-dependent interleaving of articulatory movements.

The central thesis of the model is that the spatiotemporal patterns of speech are shaped by a dynamical system with two functionally distinct but interacting levels. The *interarticulator* level is defined according to both *model articulator* (e.g., lips and jaw) and *tract-variable* (e.g., lip aperture and protrusion) coordinates; the *intergestural* level is defined according to a set of *activation* coordinates. Invariant gestural units are posited in the form of context-independent sets of dynamical parameters (e.g., lip protrusion target, stiffness, and damping coefficients), and are associated with corresponding subsets of these coordinates. Each unit's activation coordinate reflects the strength with which the associated gesture "attempts" to shape vocal tract movements at any given point in time. The tract-variable and model articulator coordinates of each unit specify, respectively, the particular vocal-tract constriction (e.g., bilabial) and articulatory synergy whose behaviors are affected directly by the associated unit's activation. The formation and release of such constrictions are governed according to an overall control regime defined as a damped, second-order dynamical system that spans tract-variable and model articulator coordinates, and whose parameters are functions of the current set of gestural activations.

The interarticulator level accounts for the coordination among articulators at a given point in time due to the currently active gesture set. At present, the dynamics of this level are sufficiently well developed to offer promising accounts of spatially goal-directed, multiarticulator movement patterns observed during unperturbed and mechanically perturbed gestures, and during periods of coproduction. In all cases, the evolving configuration of the articulators during a given utterance results from the gesturally- and posturally-specific way that driving influences generated in the tract-variable space are distributed across the associated articulatory synergies.

Intergestural Timing and Sequential Networks

The intergestural level accounts for the relative timing of activation intervals for the gestural units participating in a given utterance, e.g., for vocalic and bilabial gestures in a vowel-bilabial-vowel sequence. At present, this relative timing is accomplished with reference to a *gestural score* that specifies the activation of gestural units over time across parallel tract-variable output channels. Currently, gestural scores are not derived from an underlying implicit dynamics. Rather, they are derived explicitly either "by hand" or according to the rules of Browman and Goldstein's *articulatory phonology* ([1], [2], [3]).

We have adopted the recurrent, sequential network architecture of Jordan ([7], [8], [9], [10]; see also [11], [12], [13], [14]) at the intergestural level, as a means of patterning the gestural activation trajectories for the task-dynamic model. In the network, a separate output unit exists for each distinct gesture in a sequence, with the values of the output units corresponding to the activation values of the associated gestures. The network additionally includes a set of *state* units that determine among themselves a dynamical flow with an intrinsic time scale specific to the intended sequence. The dynamics of activity among the state units are defined by weighted recurrent connections: a) among the state units themselves, and b) from the output units to the state units. A set of intermediate *hidden* units are connected to both the state and output units by two respective layers of weighted paths. Temporal ordering among the output units is an implicit consequence of the network architecture and the sequence-specific set of weight values associated with each connection path.

Using *back-propagation* training methods (e.g., [15]), sequential networks can learn *single* sets of weights that will allow them to produce *several* different sequences. The information that is used to define the *teaching vector* sequences applied to the network during training will be obtained from the same articulatory phonology rules that are presently used to generate gestural scores. Such information will include gestural targets stiffness and damping coefficients, and the required times or phases of target attainment. Each simulated sequence is produced (and learned) in the presence of a corresponding constant activation pattern in a set of *plan* units. These units provide a second, external set of inputs to the network's hidden layer, in addition to the internal inputs provided by the state units.

A simplified pilot network

Pilot investigation now under way is focused on the behavior of a simplified model system. This system consists of a sequential network whose outputs represent the activations of gestures that are defined in a small set of "tract variables," whose dynamics are first order. The activation nodes act only to insert their associated target values into the tract-variable dynamics; the tract-variable stiffness coefficients are fixed. There are no model articulators at this point, equivalent to the simplifying assumption that tract variables and model articulators are defined in a one-to-one manner. These simplifications were adopted in order to gain familiarity and to sharpen intuitions regarding the types of intergestural timing phenomena that are intrinsic to such coupled dynamical systems, while keeping the

"controlled" articulatory system as simple as possible.

FOUR BENCHMARK CLASSES OF TEMPORAL PHENOMENA IN SPEECH

This section focuses on four empirical classes of phenomena related to intergestural temporal patterns: 1) the temporal extent of anticipatory coarticulation; 2) the durational changes that result when mechanical perturbations are introduced to the articulators during speaking; 3) the effects of linguistic boundaries on the relative timing of gestures associated with the same phoneme; 4) the continuous sliding and aggregation effects, and discontinuous phase transition effects, that accompany increases of speaking rate.

Emphasis is placed on the first two classes of phenomena, since they are being addressed currently in our pilot modeling efforts. The remaining two classes will be reviewed more briefly, since they are not being addressed presently in our model, although they will provide significant constraints in guiding its future development.

Temporal Extents of Anticipatory Coarticulatory Fields

The definition adopted in this paper of a gesture's anticipatory field of coarticulation is the time from gesture onset to the time of target attainment. In a review of the literature on anticipatory lip rounding (e.g., [17], [18], [19]) and velum lowering (e.g., [20]), and on transconsonantal vowel-to-vowel coarticulation (e.g., [21]), Fowler and Saltzman [22] concluded that anticipatory coarticulation fields are temporally constrained, and that they do not typically extend very far backward in time from the time of target attainment. This interpretation is consistent with that provided by Bell-Berti & Harris' [23] frame model of coarticulation, and contrasts with the extensive degrees of anticipation that are possible in *look-ahead* models (e.g., [24]; in fairness, however, it should be noted that the anticipatory feature-spreading used in Henke's [24] model looked ahead only to the immediately following segment, although unlimited anticipation was allowed in principle).

In the task-dynamic model, this definition corresponds (roughly) to the interval from the time of gestural activation onset to the time of target attainment. Currently, the activation trajectories are specified in the gestural score as simple step functions, whose values change discretely between zero (the gesture is inactive) and one (the gesture is maximally active). Consequently, a gesture's anticipatory field is simply proportional to the gesture's intrinsic time constant, which is itself a function of the gesture's fixed stiffness and damping parameters. The onset of a given activation wave is specified according to the rules of articulatory phonology so that the gesture attains its target at the appropriate point in the simulated utterance (see also [16]).

Anticipatory behavior in the pilot network

In the pilot network, each first-order tract-variable is represented by a linear unit whose inputs are current target value and tract-variable position. The unit's output is current tract-variable velocity, which is fed into a linear, self-recurrent unit that provides a discrete time, Euler integration to generate the next tract-variable position. Taken together, therefore, the tract-variable and integrator units represent a model of the *forward dynamics* from current tract-variable state and target inputs to the next tract-variable state. Additionally, a delay line from the integrator unit is used to feed back current tract-variable state to the hidden units of the sequential network (see [8], [10], [25], [26] for related treatments).

Teaching vectors are applied intermittently at the tract-variable integrator units, and output errors are measured. At all other times, "don't care" conditions exist and no errors are defined. When errors are defined, however, they are propagated backward through the fixed tract-variable forward dynamics, and applied to the sequential net's output units. From this point on, these backpropagated errors are used in the usual manner to train the weights inside the sequential net. Because of the downstream recurrence implicit in the tract-variable dynamics, and the delayed tract-variable feedback into the sequential network, the *back-propagation through time* (e.g., [15]) training procedure is used.

It is known from previous work [7] that sequential nets whose outputs directly represent the distinctive features of speech will display relatively unconstrained temporal fields of anticipatory behavior. Hence, we hypothesize that "frame-model-like" behavior will require the addition of appropriate sets of *side constraints*, perhaps to the sequential net's output units, during the network's training phase (see [8] for detailed discussion of such constrained optimization methods). Under such training conditions, the anticipatory fields of activation waves should be limited to durations proportional to the tract-variable time constants of the activated gestures.

Effects of Mechanical Perturbations: Phase Resetting

Transient mechanical perturbations delivered to the speech articulators during repetitive and discrete speech sequences can alter the underlying timing structure of the ongoing sequence, and induce systematic shifts in the timing of subsequent movement elements ([27], [28], [29], [30], [31], [32], [33]). We have recently used such methods to examine the sequential dynamics governing bilabial and laryngeal gestures, both within segments and between successive syllables, in repetitive [...pæpæ...] and discrete [pə'sæpæpl] utterances ([34], [35], [36]).

Analyses of the repetitive utterances indicated that steady-state shifts occurred for both lips and larynx, but that no steady-state shifts occurred in the relative phasing of these articulators. These effects occurred only when the perturbation was delivered within a "sensitive phase" of the cycle. During this period, the downwardly directed lower lip perturbation opposed the just-initiated, actively controlled bilabial closing gesture for /p/. Thus, the sensitive period corresponded (roughly) to the acceleration portion of the closing gesture (Kawato, personal communication). When perturbations were delivered at the system's sensitive phase, the bilabial and laryngeal gestures were phase-advanced as a relatively coherent unit, maintaining their relative phasing as they were advanced in absolute time. Durational changes observed in other phases were systematic yet transient peripheral responses to the perturbation, and did not indicate central phase-resetting. Finally, the patterns observed in the discrete utterances resembled the transient effects seen in the repetitive utterances, indicating that similar dynamics underlie the production of both types of utterance.

Phase resetting: Implications for the pilot network

The fact that steady-state phase shifts exist supports the hypothesis that central intergestural dynamics can be "permanently" reset by peripheral articulatory events. Thus, any hypothesized central timing network for speech cannot not unidirectionally drive the articulatory periphery; rather, central and peripheral dynamics must be coupled bidirectionally, so that feedback information from the biomechanical periphery also can influence the state of the central "clock." In this regard, it is interesting to recall that feedback connections from the interarticulator level (tract-variable state) to the intergestural level (sequential network hidden units) were required in designing the pilot network used to produce unperturbed gestural sequences (see above section *Anticipatory behavior in the pilot network*).

It is unlikely, however, that this architecture will be sufficient to simulate the phase-resetting results. One reason is that such resetting occurred only during the syllable's sensitive phase, when the lip-opening perturbations opposed the actions of the just-initiated bilabial closing gesture. Additionally, although changes in syllable duration were found for other perturbed phases, these changes were simply transient effects, and did not indicate resetting of the central "clock". If one hypothesizes that efferent commands to the periphery are strongest near gestural initiation, then this pattern of results implies that the timecourse of afferent sensitivity mirrors that of efferent strength. In turn, such an hypothesis suggests a modeling possibility that we will investigate, namely, that "sensory" information regarding a perturbed tract-variable trajectory,

perhaps in the form of a tract-variable error signal (i.e., current target – current tract-variable position), might be multiplicatively gated into the sequential network as a function of the activation state of gestures associated with the tract-variable.

Effects of Linguistic Boundaries

Many so-called phonemes are produced as coordinated *constellations* [1] of gestures, with characteristically different patterns of intergestural phasing depending on the constellation's location relative to a linguistic (e.g., syllable) boundary ([3], [37], [38], [39], [40]). For example, /l/ is produced with two tongue gestures — tongue tip raising and tongue body retraction. In certain dialects of English, word initial /l/'s ("light" /l/'s) are produced using roughly synchronous gestures; in word final /l/'s ("dark" /l/'s), the gestures are produced asynchronously, with gestural onset for the tongue tip aligned roughly with time of target attainment for the tongue body. Similary, the nasal consonant /m/ is produced word-initially with roughly synchronous velic lowering and bilabial closing gestures; in word final position, these gestures are asynchronous, with gestural onset for velic lowering aligned roughly with the bilabial closing gesture's time of target attainment.

Linguistic boundaries: Constraints on modeling

It is possible that these timing differences are incorporated explicitly into the phonological rules that, in our modeling scheme, are used to genererate the set of tract-variable-target teaching vectors for a given sequence. It is also possible, however, that the boundaries themselves are represented as non-tract-variable elements in the teaching vectors. For example, one possibility is that such boundary elements might serve to dampen the activations (i.e., drive the output units of the sequential network toward zero) of all the gestures in the simulated sequence in proportion to the strength (e.g., [41], [40]) of the associated boundary. Such a boundary element would serve to reduce the magnitudes of tract-variable gestures within the boundary "gesture's" temporal field. If the anticipatory field is much larger than the carryover field, this could produce pre-boundary gestural reduction similar to that reported in the literature. Whether such boundary elements could also alter the relative timing of nearby gestures is an intriguing possibility that will be explored in future developments of our model.

Effects of Increased Speaking Rate

A striking phenomenon that accompanies increases in speaking rate is that the gestures associated with temporally adjacent phonemes tend to "slide" relatively continuously into one another with a resultant increase in temporal overlap (e.g., [42], [3], [43], [44], [45], [46], [47], [48]). For example, Hardcastle [43] has shown with electropalatographic data that the tongue gestures associated with producing the (British English) consonant sequence /kl/ tend to slide into one another with experimentally manipulated increases in speaking rate.

Continuous increases in speaking rate can also produce discontinuous transitions of intergestural phasing ([49], [50], [51]). In these studies, when subjects spoke the syllable /pi/ repetitively at increasing rates, the relative phasing of the bilabial and laryngeal gestures associated with the /p/ did not change from the pattern observed at a self-elected, comfortable rate. However, when the repeated syllable /ip/ was similarly increased in rate, its relative phasing pattern switched relatively abruptly at a critical speed, from that observed for a self-elected, comfortable rate to the pattern observed for the /pi/ sequences.

Speaking rate: Constraints on modeling

At this point, we can only speculate as to the dynamical underpinnings of these continuous and discontinuous changes as a function of speaking rate. For example, it is possible that the continuous patterns of intergestural sliding might result from relatively simple changes in the values of a control parameter or parameter set at the intergestural level; e.g., the increased parallelism might be simulated by having rate increases serve to decrease the effective inhibition among the sequential net's output nodes.

The discontinuous, intergestural phase transitions may be viewed as nonequilibrium phase transitions of the type seen in other physical and biological systems (e.g., [52]). In such a framework, the rhythmic units in question are characterized as nonlinearly coupled, limit cycle oscillators; the transition is characterized as a bifurcation from a modal pattern that becomes unstable with increasing rate to another modal pattern that retains its stability. The obvious implications for model development are that when the model system is trained to perform extended repetitive /pi/ or /ip/ sequences, the rhythmically active gestures should be similarly characterizable, and should display corresponding intergestural phase transitions.

ACKNOWLEDGEMENTS

This work was supported by NIH Grant DC-00121 and NSF Grant DBS-9112198 to Haskins Laboratories. I am grateful to Dani Byrd and Philip Rubin for discussion and helpful comments on an earlier draft of this paper.

REFERENCES

[1] Browman, C., & Goldstein, L. (1986). Towards an articulatory phonology. In C. Ewan, & J. Anderson (Eds.), *Phonology yearbook 3* (pp. 219-252). Cambridge: Cambridge University Press.

[2] Browman, C. P., & Goldstein, L. (1991). Tiers in articulatory phonology, with some implications for casual speech. In J. Kingston & M. E. Beckman (Eds.), *Papers in Laboratory Phonology: I. Between the Grammar and the Physics of Speech*. (Pp. 341-338) Cambridge, England: Cambridge University Press.

[3] Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica, 49*, 155-180.

[4] McGowan, R.S. & Saltzman, E. (1995). Incorporating aerodynamic and laryngeal components into task dynamics. *Journal of Phonetics, 23*, 255-269.

[5] Saltzman, E. (1986). Task dynamic coordination of the speech articulators: A preliminary model. *Experimental Brain Research, Series 15*, 129-144.

[6] Saltzman, E. L., & Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology, 1*, 333-382.

[7] Jordan, M. I. (1986). *Serial order in behavior: A parallel distributed processing approach* (Tech. Rep. No. 8604). San Diego: University of California, Institute for Cognitive Science.

[8] Jordan, M. I. (1990). Motor learning and the degrees of freedom problem. In M. Jeannerod, (Ed.). *Attention and Performance XIII*. Hillsdale, NJ: Erlbaum.

[9] Jordan, M. I. (1992). Constrained supervised learning. *Jounal of Mathematical Psychology, 36*, 396-425.

[10] Jordan, M. I., & Rumelhart, D. E. (1992). Forward models: Supervised learning with a distal teacher. *Cognitive Science, 16*, 307-354.

[11] Bailly, G., Laboissière, R., & Schwartz, J. L. (1991). Formant trajectories as audible gestures: An alternative for speech synthesis. *Journal of Phonetics, 19*, 9-23.

[12] Grossberg, S. (1986). The adaptive self-organization of serial order in behavior: Speech, language, and motor control. In E. C. Schwab & H. C. Nusbaum (Eds.), *Pattern recognition by humans and machines (Vol. 1)*. New York: Academic Press.

[13] Guenther, F. H. (1994). *Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production*. (Technical Report CAS/CNS-94-012). Boston University Center for Adaptive Systems and Department of Cognitive and Neural Systems, Boston, MA.

[14] Kawato, M. (1989). Motor theory of speech perception revisited from minimum torque-change neural network model. *Proceedings of the 8th Symposium on Future Electron Devices, October 30-31, Tokyo, Japan* (Pp. 141-150).

[15] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McLelland, (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition, Vol 1. Foundations*. Cambridge, MA: MIT Press.

[16] Coker, C. H. (1976). A model of articulatory dynamics and control. *Proceedings of the IEEE, 64*, 452-460.

[17] Boyce, S. (1990). Coarticulatory organization for lip rounding in Turkish and English. *Journal of the Acoustical Society of America, 88*, 2584-2595.
[18] Gelfer, C., Bell-Berti, F., & Harris, K. (1989). Determining the extent of coarticulation: Effects of experimental design. *Journal of the Acoustical Society of America, 86*, 2443-2445.
[19] Perkell, J. S., & Matthies, M. L. (1992). Temporal measures of labial coarticulation for the vowel /u/: Within- and cross-subject variability. *Journal of the Acoustical Society of America, 91*, 2911-2925.
[20] Bell-Berti, F., & Krakow, R. (1991). Anticipatory velar lowering: A coproduction account. *Journal of the Acoustical Society of America, 90*, 112-123.
[21] Magen, H. S. (1989). *An acoustic study of vowel-to-vowel coarticulation in English*. Unpublished Ph.D. dissertation, Yale University, New Haven CT.
[22] Fowler, C. A., & Saltzman, E. (1993). Coordination and coarticulation in speech production. *Language and Speech, 36*, 171-195.
[23] Bell-Berti, F., & Harris, K. (1981). A temporal model of speech production. *Phonetica, 38*, 9-20.
[24] Henke, W. (1966). *Dynamic articulatory models of speech production using computer simulation*. Unpublished Ph.D. dissertation, MIT, Cambridge, MA.
[25] Kawato, M. (1990). Computational schemes and neural network models for formation and control of multijoint arm trajectory. In W. T. Miller, III, R. S. Sutton, & P. J. Werbos, (Eds.). *Neural networks for control*. Cambridge, MA: MIT Press.
[26] Massone, L. L. E., & Myers, J. (1995). The role of plant properties in point-to-point arm movements. Poster presented at *5th Annual Conference on Neural Control of Movement*, Key West, FL.
[27] Gracco, V. L., & Abbs, J. H. (1989). Sensorimotor characteristics of speech motor sequences. *Experimental Brain Research, 75*, 586-598.
[28] Kollia, H. (1994). *Functional organization of velar movements following jaw perturbation*. Unpublished Ph.D. dissertation, Department of Speech and Hearing Sciences, City Universtiy of New York, New York, NY.
[29] Löfqvist, A., & Gracco, V. L. (1991). Discrete and continuous modes in speech motor control, *Perilus XIV*, Institute of Linguistics, University of Stockholm, Stockholm, Sweden, 27-34.
[30] Munhall, K., Löfqvist, A., & Kelso, J. A. S. (1994). Lip-larynx coordination in speech: Effects of mechanical perturbations to the lower lip. *Journal of the Acoustical Society of America, 95*, 3605-3616.
[31] Saltzman, E. (1992). Biomechanical and haptic factors in the temporal patterning of limb and speech activity. *Human Movement Science, 11*, 239-251.
[32] Saltzman, E., Kay, B., Rubin, P., & Kinsella-Shaw, J. (1991). Dynamics of intergestural timing. *Perilus XIV*, 47-56, Institute of Linguistics, University of Stockholm, Stockholm, Sweden.
[33] Saltzman, E., Löfqvist, A., Kinsella-Shaw, J., Rubin, P., & Kay, B. (1992). A perturbation study of lip-larynx coordination. In J. J. Ohala, T. M. Nearey, B. L. Derwing, M. M. Hodge, & G. E. Wiebe, (Eds.). *Proceedings of the International Conference on Spoken Language Processing (ICSLP '92): Addendum*. Edmonton, Canada: Priority Printing. Pp. 19-22.
[34] Löfqvist, A., Saltzman, E., Kinsella-Shaw, J., Rubin, P., & Kay, B. (1994). Phase resetting in speech. II. Discrete utterances. *Journal of the Acoustical Society of America, 95*, (Abstract).
[35] Saltzman, E., Löfqvist, A., Kinsella-Shaw, J.,Rubin, P., & Kay, B. (1994). Phase resetting in speech. I. Repetitive utterances. *Journal of the Acoustical Society of America, 95*, 2823, (Abstract).
[36] Saltzman, E., Löfqvist, A., Kinsella-Shaw, J., Kay, B., & Rubin, P. (1995). On the dynamics of temporal patterning in speech. In F. Bell-Berti, & L. Raphael, (Eds.). *Studies in speech production: A Festschrift for Katherine Safford Harris*. Woodbury, New York: American Institute of Physics.
[37] Browman, C. P., & Goldstein, L. (1995). Gestural syllable position effects in American English. In F. Bell-Berti & L. Raphael, (Eds.). *Studies in speech production: A Festschrift for Katherine Safford Harris*. Woodbury, New York: American Institute of Physics.

[38] Delattre, P. (1971). Consonant gemination in four languages: An acoustic, perceptual, and radiographic study, Part I. *IRAL*, 31-52.
[39] Krakow, R. A. (1989). *The articulatory organization of syllables: A kinematic analysis of labial and velar gestures*. Unpublished Ph.D. dissertation, Yale University.
[40] Sproat, R., & Fujumura, O. (1993). Allophonic variation in English /l/ and its implications for phonetic implementation. *Journal of Phonetics, 21*, 291-311.
[41] Lehiste, I. (1980). Phonetic manifestation of syntactic structure in English. *Annual Bulletin of the Research Institute of Logopaedics and Phoniatrics, 14*, 1-27.
[42] Barry, M. C. (1991). Temporal modelling of gestures in articulatory assimilation. In *Proceedings of the XIIth International Congress of Phonetic Sciences*. (Pp. 7-8). Aix-en-Provence: University of Provence.
[43] Hardcastle, W. J. (1985). Some phonetic and syntactic constraints on lingual coarticulation during /kl/ sequences. *Speech Communication, 4*, 247-263.
[44] Munhall, K., & Löfqvist, A. (1992). Gestural aggregation in speech: Laryngeal gestures. *Journal of Phonetics, 20*, 111-126.
[45] Nittrouer, S. (1991). Phase relations of jaw and tongue tip movements in the production of VCV utterances. *Journal of the Acoustical Society of America, 90*, 1806-1815.
[46] Nittrouer, S., Munhall, K., Kelso, J. A. S., Tuller, B., & Harris, K. S. (1988). Patterns of interarticulator phasing and their relation to linguistic structure. *Journal of the Acoustical Society of America, 84*, 1653-1661.
[47] Shaiman, S., & Porter, R. J. Jr. (1991). Different phase-stable relationships of the upper lip and jaw for production of vowels and diphthongs. *Journal of the Acoustical Society of America, 90*, 3000-3007.
[48] Zsiga, E. C. (1994). Acoustic evidence for gestural overlap in consonant sequences. *Journal of Phonetics, 22*, 121-140.
[49] Kelso, J. A. S., Saltzman, E. L., & Tuller, B. (1986a). The dynamical perspective on speech production: Data and theory. *Journal of Phonetics, 14*, 29-60.
[50] Kelso, J. A. S., Saltzman, E. L., & Tuller, B. (1986b). Intentional contents, communicative context, and task dynamics: A reply to the commentators. *Journal of Phonetics, 14*, 171-196.
[51] Tuller, B. & Kelso, J. A. S. (1991). The production and perception of syllable structure. *Journal of Speech and Hearing Research, 34*, 501-508.
[52] Haken, H., Kelso, J. A. S., & Bunz, H. (1985). A theoretical model of phase transitions in human hand movements. *Biological Cybernetics, 51*, 347-356.

A MODELING FRAMEWORK FOR SPEECH MOTOR DEVELOPMENT AND KINEMATIC ARTICULATOR CONTROL

Frank H. Guenther
Center for Adaptive Systems and Department of Cognitive and Neural Systems
Boston University, Boston, MA 02215

ABSTRACT

This paper presents three hypotheses that are central to a computational model of speech production: (1) Sound targets take the form of regions, rather than points, in a planning reference frame. (2) The planning frame is more acoustic-like than the frames used in most recent models. (3) A direction-to-direction mapping transforms planned trajectories into articulator movements. These hypotheses are supported by experimental data and simulation results.

1. INTRODUCTION: REFERENCE FRAMES AND MAPPINGS

It is useful to think of speech production as the process of formulating a trajectory within a planning reference frame to pass through a sequence of targets, each corresponding to a different phoneme in the string being produced. This trajectory can then be mapped into a set of articulator movements that carry out the planned trajectory. The articulator movements are defined within an articulatory reference frame that relates closely to the musculature or primary movement degrees of freedom of the speech articulators. The process of mapping from the planning frame to the articulator frame need not wait until the entire trajectory has been planned, but instead may be carried out in concurrence with trajectory planning.

This paper addresses three important questions that arise within this view of the speech production process. First, what is the nature of the phonemic targets? Second, what is the nature of the planning reference frame? Finally, what is the nature of the mapping from the planning frame to the articulator frame?

The answers provided in this paper arise from a computational model of speech production called DIVA. This model is briefly introduced in Section 2. Sections 3 through 5 then address the three questions posited above. Simulation results verifying the model's ability to produce vowels are presented in Section 6. (More thorough simulations of an earlier version of the model are described elsewhere [1], [2].) Finally, Section 7 shows how the model's answers to the questions posed above lead to a simple explanation for the anomalous observation that the same speaker will often use entirely different vocal tract configurations to produce the sound /r/ in different contexts.

2. MODEL DESCRIPTION

An overview of the DIVA model is shown in Figure 1. The model is formulated as a self-organizing neural network that undergoes a babbling phase, during which synaptic weights in the adaptive neural mappings (shown as filled semicircles in Figure 1) are tuned, and a performance phase, during which arbitrary phoneme strings specified by the modeler are produced as continuous movements of the speech articulators. The main components of the model are briefly described in the following paragraphs; more complete descriptions are given elsewhere [1], [2].

Each cell in the Speech Sound Map in Figure 1 corresponds to a different phoneme. The cell corresponding to the phoneme to be produced has an activity level of 1; all other cells in the map have zero activity. During babbling, the Speech Recognition System monitors the acoustic signal produced by the model (after an "auditory processing" stage that extracts formant values) and activates the appropriate cells in the Speech Sound Map when phonemes are detected. This allows learning in the weights projecting from the active Speech Sound Map cell to the cells in the Planning Direction Vector. These weights encode a target for the phoneme in planning coordinates; this target can later be used to produce the sound. The nature of these targets is the subject of Section 3.

The Planning Position Vector stage represents the current state of the vocal

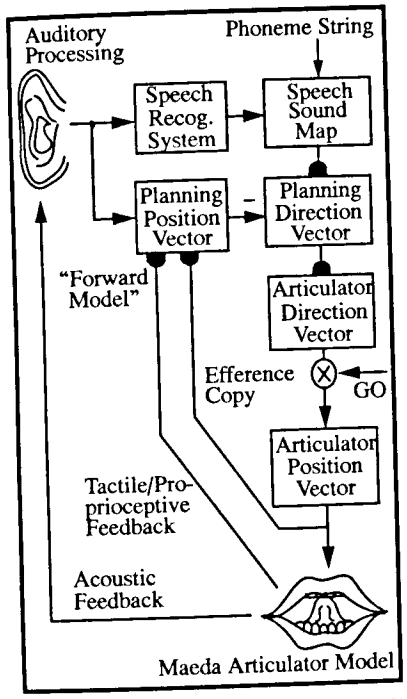


FIGURE 1. Overview of the DIVA model. Learned mappings are indicated by filled semicircles.

tract within the planning reference frame. This is needed to calculate the desired movement direction, which is formed at the Planning Direction Vector stage by subtracting the Planning Position Vector from the current sound's target. The time course of activity of the Planning Direction Vector constitutes the planned trajectory, specified within the planning reference frame described in Section 4.

The desired trajectory in planning coordinates is mapped into a set of articulator movements via the adaptive weights projecting from the Planning Direction Vector to the Articulator Direction Vector. This mapping from the planning frame to the articulator frame is the subject of Section 5.

The Articulator Direction Vector represents the desired movement direction in articulator space. Cells in the Articulator Position Vector integrate these activities (after multiplicative gating by a GO signal that controls movement speed) to produce articulator position commands. The current version of DIVA uses an articulator model created by Maeda [3]. This model consists of seven variables (jaw height, tongue body position, tongue body shape, tongue tip position, lip height, lip protrusion, larynx height) that make up the seven dimensions of the articulator reference frame.

The current version of the model includes two refinements to the version described in [1], [2]. First, the Maeda articulator set replaces the simplified articulatory structure of the earlier version, allowing synthesis of an acoustic signal based on vocal tract shape. Second, the current version uses a more acoustic-like planning reference frame. This is discussed further in Section 4.

3. PHONEMIC TARGETS

To explain how infants learn language-specific and phoneme-specific limits on variability in the production of speech sounds, Guenther [1], [2] posited a convex region theory for the targets of speech. Within this theory, the target for a speech sound consists of a range of acceptable values for each dimension of the planning frame. For example, the target for a vowel consists of a small range of acceptable values along the acoustically important dimension of tongue body constriction degree, and a much larger range along the less important dimension of velum height. This is in contrast to traditional targets, which are "all or nothing" in nature; that is, a sound's target either specifies a single target value along a dimension, or the dimension is completely unspecified.

As discussed in [2], convex region targets provide an intuitive explanation for contextual variability. The specific position within the target region that will be reached by the model during production depends on factors such as context and speaking rate; therefore, the model may end up on any point within the target region. Experimental results indicate that speakers tend to produce more variation along acoustically unimportant dimensions than they do along acoustically important dimensions [4]. This is explained by the model since the target ranges along important dimensions are much smaller than the ranges along unimportant dimensions.

Further implications of the convex region theory on several long-studied speech production phenomena were also investigated in [2]. First, it was shown that this theory provides a parsimonious explanation for a collection of speaking rate effects not previously treated by a single model. This treatment included an explanation for the experimental observation that increased speaking rate can lead to increased velocities for consonant movements but decreased velocities for vowel movements [5]. The model explains this behavior even though both sound types are produced by the same control scheme, with the effect arising from differences in the shapes of target regions for vowels and consonants.

Next, it was shown how the convex region theory provides insight into the mechanisms of carryover coarticulation. The dynamics of the neural network produce movement trajectories from the current position of the vocal tract to the closest point on the convex region target. For example, when producing the word "luke", the vocal tract configuration will move from the configuration for the back vowel /u/ to the nearest point of the target for /k/. Because the target for /k/ allows for a large amount of anterior-posterior variation in tongue body position, this will lead to a /k/ configuration that has the tongue in a relatively posterior position. For "leak", however, the tongue starts out further forward for the front vowel /i/, leading to a /k/ configuration that has a more anterior position. Thus the configuration used to produce /k/ reflects aspects of the configuration used to produce the preceding vowel, which is carryover coarticulation.

Finally, a preliminary study of anticipatory coarticulation was carried out within the framework of convex region targets. A generalization of the well-known look-ahead model of anticipatory coarticulation (e.g., [6]) was defined to allow for convex region targets and was shown to account for data not treated by the traditional look-ahead model. Because the generalized look-ahead approach posits that the amount of coarticulation is limited by the size of the convex region targets, it accounts for experimental results showing decreased coarticulation in cases where smaller targets are necessitated, including speech in languages with more crowded vowel spaces [7] and hyperarticulated speech for stress [8]. The model can also account for vowel-to-vowel anticipatory coarticulation, which is not explained by traditional look-ahead models.

4. MOVEMENT PLANNING SPACE

There are many different forms that the planning reference frame might take. For example, MacNeilage hypothesized that the target for a speech sound is a set of muscle lengths that place the speech articulators in a particular configuration that results in the target sound [9]. Producing a string of speech sounds then involves the formation of a trajectory in muscle length space that sequentially passes through the muscle length targets corresponding to the string of sounds. However, this theory runs into problems when considering speech motor equivalence; that is, speakers can use many different muscle length trajectories to produce the same phoneme. For example, speakers can immediately produce a vowel with a bite block that holds the jaw at an unnatural position, even though proper production of the vowel in this case requires a completely different set of muscle lengths than are needed under normal conditions [10]. This suggests that the target for the vowel is specified in a frame that relates more closely to the acoustic signal than to a particular configuration of the articulators.

More recently, modelers have utilized planning frames that relate more directly to the acoustic signal and thus overcome shortcomings of MacNeilage's proposal. The most common choice has been frames that describe the locations and degrees of key constrictions in the vocal tract (e.g., [1], [2], [11]). Such constriction-based frames are typically assumed to have fewer degrees of freedom than the articulator frame, so that any target in the constriction frame can be produced by one of infinitely many different configurations in articulator coordinates. In the case of a vowel, for example, the same target tongue body constriction could be reached with the jaw high and the tongue body low under normal conditions, or with the jaw lower and the tongue body higher if a bite block is present. Models that transform movement trajectories planned in a constriction frame into articulator movements in a way that automatically compensates for articulator constraints have proven capable of explaining much of the motor equivalence seen in speech [1], [11].

This discussion touches upon an important concept of biological movement control: maximally flexible performance is achieved if movements are planned in a reference frame that relates as closely as possible to the task space for the movement (e.g., acoustic space for speech), rather than a frame that relates closely to the articulators. The difficulty in explaining motor equivalence with MacNeilage's theory occurs because muscle length targets overspecify the shape of the vocal tract; targets that require all of the articulators to be in specific positions are not flexible enough to deal with things like bite blocks that prevent some of the articulators from reaching their commanded positions. A better speech production system will instead plan trajectories in an acoustic-like space, then map these trajectories into articulator movements. If the mapping process automatically compensates for externally imposed constraints on the articulators while nearly invariantly producing the planned trajectory (as the mapping described in Section 5 does), then the planning process is greatly simplified because it does not need to account for such constraints.

Although constriction frames are more directly related to the task space for speech production than articulator frames, they are still only approximations to the real task space for speech production. The true goal of the speech production mechanism is to produce an acoustic signal that conveys linguistic units to listeners, not to produce particular vocal tract constrictions. Once we consider that the end goals of speech production are acoustic, it becomes clear that constriction planning spaces can also overspecify the shape of the vocal tract. To see this, consider vowel production and note that it is possible to produce the same acoustic result (e.g., formant values) with different configurations in constriction space. For example, when producing the vowel /u/, lip rounding and tongue body raising have similar acoustic results: they both mainly act to lower F2 [12]. Thus, changes in the tongue body constriction can be compensated for acoustically by complementary changes in lip rounding. Early experimental results appear to support the idea that speakers utilize such trading relations when producing vowels [12], and analogous results have recently been observed for consonant productions [13]. If targets are specified as constriction locations and degrees, then this type of trading relation between constrictions could not be used because the target specifies a location and degree for both constrictions. This is analogous to the problem mentioned earlier for the model of MacNeilage [9]. There, a target specifying the positions of all articulators indeed leads to the correct acoustic signal, but such a target overspecifies the shape of the vocal tract and thus eliminates the possibility for automatic motor equivalent compensation.

Overspecifying the shape of the vocal tract not only reduces the ability to compensate for perturbations or constraints on the articulators, but it can also lead to inefficient movement sequences during normal speech. To see this, consider the extreme case where each phoneme's target specifies a position for every articulator. Moving from phoneme to phoneme then requires movement of many articulators that are not acoustically important.

Further evidence against constriction targets comes from studies of the American English phoneme /r/, which is a rare example of a phoneme that has at least two very different articulations that produce nearly identical acoustic patterns [14], [15]. Figure 4 shows two such configurations for /r/, known generally as "bunched" (Figure 4a) and "retroflex" (Figure 4c). The existence of two completely different configurations for producing the same phoneme is difficult for theories that hypothesize phonemic targets and movement planning in a constriction-based reference frame rather than a more acoustic-like frame. This is because the constriction locations and degrees used to produce the two /r/'s in Figure 4 are completely different; therefore the corresponding targets must also be completely different. This leads to an unparsimonious explanation in which an individual chooses one or the other target depending on context. Although not completely unreasonable, this explanation is not particularly elegant. A more

parsimonious explanation utilizing a single target specified within an acoustic-like planning frame is given in Section 7.

In summary, from a modeling standpoint it makes great sense for the speech production system to utilize an acoustic-like space for target specification and movement planning rather than a constriction space or articulator space, and experimental evidence that human production systems indeed use such a frame is starting to accumulate. In keeping with this, the model of Figure 1 currently utilizes a planning frame whose dimensions correspond to formant values (see also [12], [16]). That is, the model plans formant trajectories to reach formant targets and maps these trajectories into articulator movements as described below. This formant frame replaces the constriction planning frame used in the earlier version of the model [1], [2].

5. MAPPING FROM PLANNING FRAME TO ARTICULATOR FRAME

Trajectories planned in an acoustic-like reference frame must be carried out by articulator movements. One possibility is to use a position-to-position mapping from planning space to articulator space; i.e., map each point in formant space directly into an articulator configuration that produces the desired formants. Another possibility is to use a direction-to-direction mapping from desired movement directions in planning space into movement directions of the articulators. With this kind of mapping, the configuration used to produce a desired set of formants will depend on factors such as starting configuration and externally imposed constraints on the articulators. It has been shown elsewhere [1], [16] that direction-to-direction mappings are capable of explaining motor equivalence data that position-to-position mappings cannot explain. Therefore, the DIVA model utilizes a direction-to-direction mapping to transform the desired formant changes represented at the Planning Direction Vector stage into articulator movements at the Articulator Direction Vector stage.

Earlier work demonstrated the model's ability to reach phoneme targets even in the presence of external perturbations or constraints applied to the articulators (e.g., complete blockage of jaw movement) [1]. As in humans, compensation in the model is automatic; i.e., no new learning is required under the constraining conditions, and compensation occurs without invoking special strategies to deal with the constraints. Simulations reported in the next section verify the ability of the current version of the model to compensate for bite blocks during vowel production, and simulations reported in Section 7 show how the direction-to-direction mapping helps explain the variability in /r/ production described above.

6. VOWEL SIMULATIONS

The earlier version of the model [1], [2] produced arbitrary combinations of a set of 29 phonemes, including both vowels and consonants, using its simplified articulatory structure and the constriction-based planning frame. The current version, which utilizes an acoustic-like planning frame, does not yet produce consonants.

Simulations of the model were carried out on a Sparc-10 workstation. Ten English vowels were learned during babbling. After learning, synthesis of the the model's vocal tract configurations while producing each vowel in isolation resulted in recognizable vowel sounds. Each vowel could be produced by the model from any starting configuration of the vocal tract. As illustrated in Figure 2, the resulting vocal tract shapes correspond roughly to shapes seen in humans producing the same vowels, even though no vocal tract shape information is encoded in the targets learned by the model.

Each of the ten vowels were also successfully produced with the jaw blocked at various positions, demonstrating the model's motor equivalence capabilities. With the jaw blocked, other articulators such as the tongue compensated, allowing the vocal tract to assume an overall shape that reached the acoustic target for the vowel. Phonemes produced with the jaw blocked were perceptually indistinguishable from phonemes produced with an unconstrained jaw.

7. /r/ SIMULATIONS

Section 4 discussed how the use of two completely different articulator configurations for /r/ by the same speaker is troublesome for models using a constriction-based planning frame. This section describes how the use of an acoustic-like planning space and a direction-to-direction mapping from the planning frame to the articulator frame provides a simple explanation for this observation.

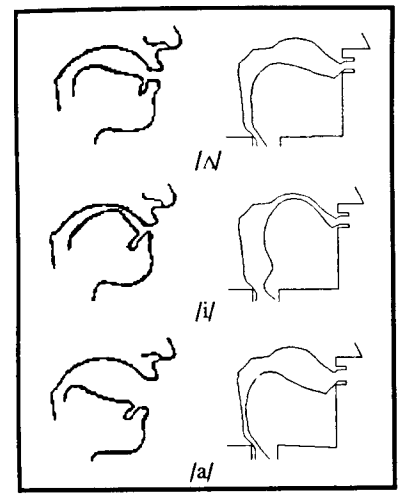


*FIGURE 2. Vocal tract configurations corresponding to different vowels. The left side shows schematics of configurations used by humans (adapted from [17]) and the right side shows the configurations produced by the model. **Top row.** The central vowel /ʌ/ as in "up". **Middle row.** The high front vowel /i/ as in "beet". **Bottom row.** The low back vowel /ɑ/ as in "father". Model configurations approximate human configurations even though no articulatory or vocal tract shape information is used for target specification or movement planning in the model.*

It is important to note that simple target regions in acoustic space often correspond to complex regions in articulator space. The top half of Figure 3 shows a simple convex region in formant space that approximates the ranges of F1 and F2 for the phoneme /r/. The bottom half of the figure shows the corresponding region in two dimensions of articulator space. This figure was produced by fixing five of the Maeda articulators and varying the remaining two through their entire ranges to determine which configurations result in formants in the ranges specified in the top half of the figure. Note that the articulator space region is non-convex; in fact, it is broken into two distinct sub-regions. The top sub-region corresponds to a flattened tongue tip as in retroflex /r/, and the bottom sub-region corresponds to a bunched tongue configuration as in bunched /r/.

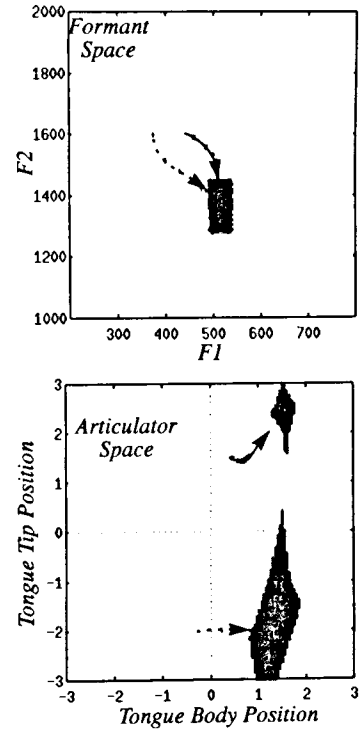


FIGURE 3. Relationship between a simple convex region corresponding to /r/ in formant space (top) and the corresponding region in articulator space (bottom). Arrows indicate model trajectories when producing /r/ starting from a /d/ configuration (solid lines) and from a /g/ configuration (dashed lines).

Assume that the same speaker is to produce the words "grab" and "drag". Often the same speaker will use a bunched /r/ following /g/ as in "grab" and a retroflex /r/ following /d/ as in "drag" [15]. The same target for /r/, consisting of the formant ranges in the top half of Figure 3, is specified to the production mechanism regardless of phonetic context. This target is manifested

by the weights between the Speech Sound Map and the Planning Direction Vector stages of the model. The Planning Direction Vector activity then represents the desired movement direction in formant space. These desired formant changes are transformed into articulator movements through the direction-to-direction mapping manifested by the weights projecting from the Planning Direction Vector to the Articulator Direction Vector. These movements result in the formant trajectories shown in the top half of Figure 3. The solid arrow is the trajectory produced when /d/ precedes /ɹ/, and the dashed arrow is the trajectory produced when /g/ precedes /ɹ/.

The important thing to note is that the direction-to-direction mapping transforms these formant trajectories, which go to a single target region in formant space, into articulator trajectories that end up at *different* sub-regions in articulator space. The articulator space trajectories are indicated by the arrows in the bottom half of Figure 3. (Because this plot represents just two of the seven articulator dimensions, the trajectories are only approximate.) Roughly speaking, the direction-to-direction mapping causes the model to automatically move to the closest sub-region in articulator space. When /g/ precedes /ɹ/ the bottom sub-region corresponding to bunched /ɹ/ is closest (dashed arrow), and when /d/ precedes /ɹ/ the upper sub-region corresponding to retroflex /ɹ/ is closest (solid arrow). This behavior is only possible because: (i) the target is acoustic-like and does not include articulatory or constriction specifications, and (ii) formant positions in the planned trajectory are not mapped directly into articulator positions, but instead formant *changes* are mapped into articulator position *changes* by the direction-to-direction map.

The articulator configurations produced by the model in the two contexts are shown in Figure 4b,d. Although the model captures the major aspects of /ɹ/ articulation of interest here (i.e., different sub-regions in articulator space, corresponding to a flattened tongue tip and a bunched tongue tip, are used to produce /ɹ/ in different contexts), the model's configurations only roughly correspond to human configurations. In particular, the model's tongue tip during retroflex /ɹ/ is not as retroflexed as a human's tongue tip. There are two reasons for this. First, the limited degrees of freedom of the Maeda articulators do not allow for much retroflexion of the tongue. Second, an important acoustic cue for /ɹ/ is a very low F3. Because the model does not currently include F3 in the planning space, this aspect is not captured here. The sublingual cavity formed with retroflex tongue shapes is partly responsible for lowering F3. It is therefore anticipated that incorporating F3 in the planning space and using a better model of the tongue and sublingual cavity will result in /ɹ/ productions that are more retroflexed than in Figure 4d.

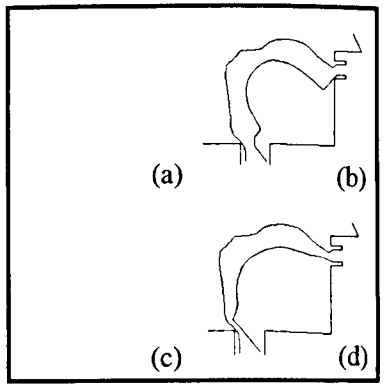


FIGURE 4. (a) Typical bunched /r/ configuration (adapted from [14]). (b) Configuration reached by the model when producing /r/ after /g/. (c) Typical retroflex /r/ configuration (adapted from [14]). (d) Configuration reached by the model when producing /r/ after /d/.

It should also be noted that the preceding phoneme is only one of the factors that affect the choice of /ɹ/ configuration, so the model does not yet account for all aspects of /ɹ/ variability.

8. CONCLUDING REMARKS

This paper has presented three simple hypotheses that explain a wide range of experimental data on speech articulation. These hypotheses are implemented in a computational model of speech acquisition and production. Simulation results verified the model's ability to produce vowels with or without a bite block and to explain the anomolous observation that the same speaker will often use widely different articulator configurations to produce /ɹ/ in different contexts.

Ongoing research is addressing the addition of F3 to the planning space and the addition of consonants to the model's phonemic repertoire. It is believed that consonants, unlike vowels and /ɹ/, may require the incorporation of constriction information in the target specification. Therefore, a hybrid planning space including both formants and constrictions will be investigated.

9. REFERENCES

[1] Guenther, F.H. (1994), "A neural network model of speech acquisition and motor equivalent speech production." *Biological Cybernetics*, vol. 72, pp. 43-53.

[2] Guenther, F.H. (in press), "Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production." *Psychological Review.*

[3] Maeda, S. (1990), "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model." In Hardcastle W.J. and Marchal, A. (eds), *Speech Production and Speech Modelling*, Boston: Kluwer Academic Publishers, pp. 131-149.

[4] Perkell, J.S., & Nelson, W.L. (1985), "Variability in production of the vowels /i/ and /a/." *Journal of the Acoustical Society of America*, vol. 77, pp. 1889-1895.

[5] Gay, T., Ushijima, T., Hirose, H., & Cooper, F.S. (1974), "Effects of speaking rate on labial consonant-vowel articulation." *Journal of Phonetics*, vol. 2, pp. 47-63.

[6] Kozhevnikov, V.A., & Chistovich, L.A. (1965), *Speech: Articulation and perception.* Translation by Joint Publications Research Service, Washington DC, JPRS 30543.

[7] Manuel, S.Y. (1990), "The role of contrast in limiting vowel-to-vowel coarticulation in different languages." *Journal of the Acoustical Society of America*, vol. 88, pp. 1286-1298.

[8] De Jong, K., Beckman, M.E., & Edwards, J. (1993), "The interplay between prosodic structure and coarticulation." *Language and Speech*, vol. 36, pp. 197-212.

[9] MacNeilage, P.F. (1970), "Motor control of serial ordering in speech." *Psychological Review*, vol. 77, pp. 182-196.

[10] Lindblom, B., Lubker, J., & Gay, T. (1979), "Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation." *Journal of Phonetics*, vol. 7, pp. 147-161.

[11] Saltzman, E.L., and Munhall, K.G. (1989), "A dynamical approach to gestural patterning in speech production." *Ecological Psychology*, vol. 1, pp. 333-382.

[12] Perkell, J.S., Matthies, M.L., Svirsky, M.A., and Jordan, M.I. (1993), "Trading relations between tongue-body raising and lip rounding in production of the vowel /u/: A pilot 'motor equivalence' study." *Journal of the Acoustical Society of America*, vol. 93, pp. 2948-2961.

[13] Perkell, J.S., Matthies, M.L., and Svirsky, M.A. (1994), "Articulatory evidence for acoustic goals for consonants." *Journal of the Acoustical Society of America*, vol. 96(5), Pt. 2, p. 3326.

[14] Delattre, P., and Freeman, D.C. (1968), "A dialect study of American r's by x-ray motion picture." *Linguistics*, vol. 44, pp. 29-68.

[15] Espy-Wilson, C., & Boyce, S. (1994), "Acoustic differences between 'bunched' and 'retroflex' variants of American English /ɹ/." Journal of the Acoustical Society of America, vol. 95(5), Pt. 2, p. 2823.

[16] Bailly, G., Laoissière, R., & Schwartz, J.L. (1991), "Formant trajectories as audible gestures: an alternative for speech synthesis." *Journal of Phonetics*, vol. 19, pp. 9-23.

[17] Bullock, D., Grossberg, S., & Guenther, F.H. (1993), "A self-organizing neural model of motor equivalent reaching and tool use by a multijoint arm." *Journal of Cognitive Neuroscience*, vol. 5, pp. 408-435.

[18] Flanagan, J.L. (1972), *Speech analysis, synthesis, and perception*, New York: Springer-Verlag.

10. ACKNOWLEDGEMENTS

This research was supported in part by the Air Force Office of Scientific Research (AFOSR F49620-92-J-0499). The author would like to thank Dave Johnson for his contribution to this work.

LOCAL SHAPES AND GLOBAL TRENDS

Mary E. Beckman
Department of Linguistics, Ohio State University, Columbus, OH, USA
and ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan

ABSTRACT

The topic of this symposium assumes answers to two more basic questions: What is intonation? How should we understand phonological structure in general? Several phenomena basic to the debate over superpositionality (e.g., variation in overall pitch range and accent realization in the context of variation in prominence and downtrend) are reviewed in the context of these questions.

INITIAL DEFINITIONS

It will be difficult to discuss whether the structure of intonation is linear or superpositional unless we lay out our background assumptions on two more general questions: *What do we mean by "structure"? What is "intonation"?* So let me begin here. I think we all agree that the "structure" is phonological or phonetic. We are talking about the structure of **sound**, and not syntactic structure or discourse topic structure or any of the other kinds of structure relevant to understanding the sound structure of intonation. On "intonation", though, we are in less agreement. Where Grønnum [1] proposes to exclude, for example, microprosody, my prejudice is to make the definition as inclusive as possible. In modeling intonation, I want to understand **all** aspects of the perceived pitch pattern that the speaker intends for the hearer to use in understanding the utterance, or that the hearer does use whether intentionally controlled by the speaker or not.

ON REPRESENTATION

With these definitions, then, the topic of this symposium is: *What kinds of structures should we use to represent the speaker's and hearer's knowledge of these relevant aspects of the pitch pattern?* In other words, the question becomes one of the representation of knowledge about sound. And here again, I think we will find points of agreement and points of disagreement among the participants in this symposium.

One salient point of agreement is that we have chosen as our primary phonetic representation that most convenient measure, fundamental frequency. All of us on this panel have worked almost exclusively with this representation — because by comparison to perceptual or physiological records, F0 is extremely easy to get. Also, (although we explicitly invoke this reason less often) the F0 computed from a complex harmonic signal is known to be psychoacoustically close to the pitch perceived for it.

Note that in the above restatement of the title of this symposium I have deliberately rephrased it in the plural. I have said "kinds of structures" rather than "the structure", because I am prejudiced to think that useful representation is not monolithic. Every representation is a model, a hypothesis about some aspect of the speaker/hearer's competence. There are many different dimensions along which we can study speech, and a representation that is useful for modeling competence in one dimension is not necessarily useful for any other; nor does it necessarily map easily to useful representations for other dimensions. For example, there is no representation of tongue shape that is useful for modeling the speaker's control of vowel constrictions and which also maps simply onto any useful representation of the hearer's knowledge of the resulting variations in timbre.

To justify this prejudice specifically for intonation, let me entertain a proposal that is not as silly as it may sound. We all use the F0 contour as a convenient phonetic representation. But suppose we were to take it not as a representation, but as the representation of intonation. Then the answer to our debate would be trivial: the structure of intonation is linear, because it is the linear sequence of F0 values calculated for the succession of sampling intervals in the utterance. I think the other participants in this symposium would join me in quickly rejecting this trivial solution. And this

immediate response would be at least in part because of a set of phenomena that I think we all would agree requires a superpositional representation, involving what I will call "overall pitch range".

ON PITCH RANGE

There are several kinds of variation that fit into this set. First, different speakers have different ranges of F0 values that they can produce comfortably. This is largely an artifact of laryngeal anatomy. However, hearers do identify a common intonation pattern in the sequence of relatively low F0 values on the phrase *Thank you* in an adult male speaker's production of *Marco, say "Thank you."* and in the sequence of much higher F0 values when 4-year old Marco obeys his father's injunction. So hearers clearly can factor out the speaker differences in parsing the intonation pattern — just as they normalize to different vocal tracts in perceiving the timbre of vowels produced by different speakers — and all of our theories of intonational structure include at least an implicit representation of the speaker's overall pitch range in our models of the hearer's competence. (I would argue for including it in our models of the speaker's control as well, since any phonetically competent speaker knows how to speak in a higher than comfortable pitch range in order to sound like a younger cuter speaker, or in a lower than comfortable pitch range to sound like an older more authoritative speaker.)

A second kind of variation in overall pitch range involves global variation in vocal effort. Unless trained to do otherwise, speakers use higher F0s when speaking up to be heard over distance or above ambient noise. Two recent studies suggest that this increase is caused by the higher subglottal pressure of the louder voice [2, 3], so we may not want to include it in the same way as other aspects of F0 variation in our models of the speaker's control. However, it has been convenient to model the hearer's competence at parsing it using the same representation as for inter-speaker variation (e.g. [4]).

A third seemingly related but functionally different phenomenon is variation in pitch range to reflect different emotional states. For example, a happy voice might involve an upward shift of the pitch range, whereas an angry voice involves a "pressed" voice quality often associated with a lower pitch range. Trained actors can produce this sort of variation at will in order to simulate emotional states that they do not feel, but comparisons between productions of "aprosodic" patients and of untrained controls (e.g. [5]) suggest that this professional ability is a honing of a skill within any normal speaker's competence.

A final related variation is the use of increased overall pitch range to signal greater emotional involvement with the content of an utterance. The expanded pitch range seems to be associated typically with a louder voice as well, but perhaps context helps hearers distinguish this use from increased vocal effort to project above ambient noise. This is an aspect of pitch range variation that can be quasi-conventionalized to distinguish alternative interpretations of some intonation patterns. For example, there is an intonation pattern in Korean that involves a sharp pitch rise localized to the last syllable in the utterance (i.e. a H% edge tone), which can be interpreted either as a straightforward yes-no question or as an incredulous echo question. Jun & Oh [6] have shown that the expanded pitch range of emotional involvement is a strong cue biasing the hearer toward the latter interpretation. Since Korean morphology does not distinguish WH-question words from corresponding indefinite pronouns, the expanded pitch range can be the salient cue distinguishing the two grammatical categories in utterances with this tune. Thus, the sentence [nuka wajo] produced with a H% and in a neutral pitch range might be interpreted as 'Is anyone coming?' but in an expanded pitch range as 'WHO did you say is coming?!' A relatively nonlocal increase in F0 also distinguishes the literal question interpretation from the rhetorical question interpretation of the analogous tone pattern in Osaka Japanese [7], the "incredulity" from the "uncertainty" reading of the scooped rise-fall-rise pattern in English [8], and the incredulous rhetorical question interpretation from the declarative interpretation of utterances ending in a L% boundary tone in Kipare [9]. In all

these cases, it seems possible to model the phenomenon as an expansion of pitch range over the whole utterance — in other words, as affecting the same dimension of intonational structure that we posit to account for how hearers accommodate to different speakers and to the effects on F0 of increased overall vocal effort in background noise.

I think even the most radically linear of us would agree that we should model the way in which this dimension interacts with other dimensions of the hearer's competence by adopting a superpositional representation of pitch range. We model the hearer as deriving some kind of backdrop graph paper abstracted away from the actual pitch contour in order to represent more local events (such as the high- versus mid- versus low-level lexical tones of Taiwanese or the components contrasting different pitch accents in English) as invariant across productions by different speakers or by the same speaker in different "voices". Our points of disagreement, rather, involve the ways in which the clearly local events interact with kinds of variability other than variation in overall pitch range. In particular, we have published rather divergent ideas about the best way to represent certain patterns of variability associated with such functions as signaling local prominence or discourse topic structure. Before I discuss these phenomena, however, I must admit to one last prejudice about phonological structure in general.

ON ABSTRACTION

In the preceding section, I used the notion "abstraction" to describe what the hearer does in factoring out certain effects of change in speaker, in vocal effort, and in type and degree of emotional involvement in the utterance. The hearer abstracts away from the actual pitches, a representation of the utterance's overall pitch range. Degree of abstraction has sometimes been equated with different levels in a model of grammar whereby different rule modules apply in sequence to generate from the speaker's intentions some sound signal. That is, a more abstract representation has been taken to mean something further upstream in the generative process, and phonological representation, in particular, has been contrasted with phonetic representation as being more abstract. By this usage, the representation of overall pitch range abstracted away from the F0 contour might be considered "phonological".

I am not sure enough of how the grammars of real speakers and hearers work to say confidently what is the difference between "phonological" and "phonetic" representation, but I am convinced that abstractness is irrelevant. The representation at the heart of many applications of the acoustic theory of speech production — that of the vocal tract as a lossless uniform tube — is an extremely abstract idealization, but who would argue that this is nearer to the speaker's intent than a representation of a lossy bent tube that better approximates the physical reality? To be sure, by abstracting away from the F0 contour to a graph-paper representation of overall pitch range, I think we do get closer to the phonological in the sense that we are making a plausible hypothesis about what it is the hearer does to parse the signal for the speaker's intent to produce phonologically contrasting tone levels. However, I do not think this makes the representation of pitch range itself phonological, because the dimensions of the "graph paper" are still continuously variable. I think it is useful for other reasons (e.g., for understanding how sociolinguistic variation develops into sound change) to say that any such continuously variable representation derived from a physical dimension is "phonetic" and only becomes "phonological" when it is reanalyzed in terms of arbitrarily discrete categories. Therefore, I would like to reserve "phonological" to refer to categories of conventional paradigmatic contrast that arbitrarily discretize relations of difference along some (possibly very abstract) phonetic dimension, and to refer to discrete categories of metrical organization derived from syntagmatic constraints on the occurrence of these paradigmatically contrasting categories. In short, "phonological" for me means the symbolic representation of such paradigmatic content categories as [+Hightone] (which in the pitch accent system of English is in opposition to the category [-Hightone] — see below), and of such syntagmatic categories as "stress foot" (which in English describes a metrical structure constraining where pitch accents can occur in the alignment between intonational melody and segmental string).

I think the other participants in this symposium would agree with me in calling the graph paper representation of pitch range "phonetic" rather than "phonological". However, they may not all agree with me on my reasons for this, and so it is worth belaboring the point by giving one more illustration of my prejudice about abstractness. The example involves the representation of microprosody — the so-called "intrinsic" F0 variation associated with voiced versus voiceless obstruents. My reading of the literature on phonological development convinces me that all abstraction comes with a cost, and that new speaker/hearers acquiring a language for the first time will not abstract away a symbolic representation of knowledge unless it is unavoidably useful in speech communication. In particular, they will not abstract away a set of discrete phonological categories unless there is a compelling phonological reason to do so, some reason such as the categories' participation in the "dual structure" of a small inventory of meaningless entities of contrast that are used to compose a larger inventory of meaningful entities. So although I would include microprosody within the set of phenomena that we must model when we model intonation, I would use a superpositional phonetic representation to do so, and would certainly reject any proposal to specify an intermediate phonological structure between the pitch contour and the symbolic phonological representation of the categorical contrast between [+voice] and [-voice]. That is, I think we should assume that if speakers intend to produce these "intrinsic" effects, they are controlling them in terms of the task of producing the [+voice] or [-voice] category, and when listeners factor out the various influences on the pitch pattern, they perceive these "intrinsic" effects in terms of the voicing contrast directly without parsing any intermediate phonological categories of [raised pitch] versus [lowered pitch] . (Of course it is possible that sociolinguistic pressure might induce a particular speech community to begin parsing such intermediate phonological categories to represent how different subsets of the speaker/hearers sound different, but that is another story — the story of reanalysis and incipient tonogenesis.)

ON STRESS AND ACCENT

I have belabored this point, because it is relevant for the discussion of accent and accentual prominence, a set of phenomena concerning which we are perhaps in most disagreement. My discussion of microprosody is superficially similar to the argument that Grønnum [1] uses against a symbolic representation of pitch accent in Danish. That is, if I understand her correctly, she is saying that the rise in pitch anchored at the beginning of each stress group is predictable from the stress pattern in much the same way that the raised pitch onset associated with a preceding voiceless obstruent is predictable from the representation of the [-voice] category, and therefore there is no need to include in the model of intonation any structure representing that stress-group initial rise.

To explain why I disagree, I need to make clearer the distinction I understand between the two types of phonological category — those of metrical organization and those that discretize some dimension of phonetic content. Let me illustrate this with a discussion of vowel timbre categories. In English there are phonotactic constraints on where certain types of vowels can occur, constraints which are economically represented by positing an abstract metrical category "stress foot" , defined as a grouping of syllables that is headed by an obligatory initial "stressed" syllable where the full range of vowel contrasts is possible. These kinds of metrical properties are about the most abstract kind of phonological structure imaginable. The child acquiring the language must in effect abstract them from the phonotactic constraints on the content features defining them. Now imagine a language in which the foot is defined similarly, but the obligatory initial syllable always contains [a] and the optional trailing syllables always contain [i]. By Grønnum's argument, we would need no representation of vowel timbre categories in this language because the difference

between [a] and [i] is already represented in the stress contrast. True, but I think a more plausible model of what children would do in acquiring this language is to build a phonetic representation of the vowel timbre space that allows them to abstract away from such things as speaker-related variability to a categorical contrast between [a] and [i], and then abstract away a metrical property "stress" to represent the constraint on where [a] and [i] occur. The analogous hypothesis about Danish is that children build a representation of pitch range to abstract away something like the categories [+Hightone] (henceforth "H") and a contrasting [-Hightone] ("L") so that they can then abstract away an understanding of the stress group in terms of the distribution of these tones. The child acquiring the Copenhagen dialect, for example, might decide that L only occurs on the group-initial stressed syllable and H only on the next syllable in the stress group.

This distinction between metrical properties and content features is even more critical in English, where there are several paradigmatically contrasting pitch accent types signaling different pragmatic functions. For example, the L+H* accent (where the "*" means that the rise is aligned to put the H tone on the stressed syllable) signals a choice of value along some semantic scale and a commitment to the pragmatic relevance of that value, whereas the L*+H (i.e. the same rise but aligned very similarly to the Copenhagen Danish stress-group marker) signals the same sense of scalar choice but lack of commitment [10]. This difference can have enormous consequences for the implied presuppositions of a statement. For example, in the context of the assertion *No one in his right mind works on intonation.* the response *Beckman works on intonation.* said with a L+H* accent on *Beckman* means that Beckman is a relevant counterexample, whereas the same sentence produced with a L*+H accent means that this is not really a relevant counterexample and thus implies that Beckman is dotty. Pierrehumbert & Hirschberg [11] describe contrasting pragmatic meanings associated with four other pitch accent types, including a contrast between a single-tone H* accent and a single-tone L* accent. These contrasts in pitch accent type must be included as part of the phonology of intonation. Thus, it is only the pattern of association between accents and syllables that can be relegated to the phonology of stress, to represent the constraint that an accent can occur only on the head of a stress foot, and the fact that pitch accent placement defines a categorical level of metrical prominence over and above the specification of stressed versus weak syllables in the lexicon. (There are further constraints on pitch accent placement which are related to such discourse phenomena as focus, and which are very similar to the facts that Grønnum describes about the "suppression" of the pitch rise on neighboring stress-groups to indicate emphatic contrast.)

I think most people who work on the phonology of English intonation would agree on the above characterization of pitch accents as something that must be represented in the phonology of the intonation pattern directly and independently of the relationship between pitch accent placement and stress. Also, while not everyone agrees on the exact inventory, there is a strong consensus that H and L tone levels are the right way to represent the accentual contrasts (see, [12] as well as Pierrehumbert's and Ladd's analyses). This is an important point of agreement, because some superpositional models of typologically similar intonational systems assume a different representation of pitch accent, in terms of rises or falls as holistic units (see Möbius's representation of the pitch accents of German [13]). One of the most compelling motivations for H and L as the "phonemes" of English intonation is that these occur alone as the intonational morphemes H* and L*. This analysis of single-tone accents is supported by intonation contours where several L* accents occur in succession with no intervening rise in pitch or several H* accents occur in succession with no intervening dip. (Several of the F0 contours of utterances being modeled in Möbius's paper make me think that German has comparable sequences of single-tone accents. Note, for example, the flat F0 pattern over the sequence *sieben Minuten entfernt* in Fig. 5.)

ON LOCAL PROMINENCE

The tone-level analysis of English accents is also supported by data on the way that F0 peaks associated with H* accents vary under changes in overall pitch range. In the classic experiment [4], subjects produced the sentence *Anna came with Manny* in response to contexts that induced one or another fixed pragmatic relationship between the names. The sentence was in both cases a sequence of two intonational phrases, so that both names were accented (and hence focused), but in half of the productions, *Anna* was assumed background and *Manny* the answer to the context question, and in the other half, *Manny* was the background focus and *Anna* the answer. The two renditions were produced many times at each of ten different overall vocal effort levels. When the F0 value of the second peak was plotted as a function of the first peak, the two accents showed a very tight clustering around the two regression curves for the contrasting pragmatic relationships, suggesting that the speakers were controlling the location of the accent peak within the overall pitch range. By contrast, when the extent of the F0 rise into the first peak was plotted against the extent of the second rise, there was no clear clustering. A recent study of cricothyroid muscle activity and subglottal pressure in productions of a somewhat different set of pragmatic relationships [3] offers further support for this interpretation.

While the experiment supports the representation of accent production in terms of H versus L target values at values relatively high or low within the overall pitch range, it also raises knotty questions about how to model the variability within the overall pitch range for the same accent in the two renditions. The peak was higher when the accent was the answer focus as compared to when it was the background focus (although the first peak was always at least as high as the second, whether answer or background). Liberman & Pierrehumbert themselves first modeled this variability by representing the overall pitch range in terms of a choice of scalar value for a "reference line" and then representing the two H* tones as differing in local phonetic "prominence" — another scalar value specified accent by accent and governing the distance of the accent's target tones above the reference line. Others have proposed alternative models, however. For example, a revision to Liberman & Pierrehumbert's original model suggested in [14] specifies a reference line for each intonational phrase (in effect allowing each phrase to have its own local pitch range), so that the answer/background peak relationship is modeled as a speaker strategy for choosing particular values for the reference lines of the two phrases.

The original proposal to represent the accent/background relationship in terms of choice of accentual prominence values was based on the desire to limit the degrees of freedom in the model, since accentual prominence was already being used to model one component of "declination". That is, Liberman & Pierrehumbert found that (for another large set of utterances by the same speakers) they could model the cumulative exponential decline over utterances with varying numbers of accents produced in three different overall pitch ranges, simply by specifying a proportional reduction within the overall pitch range of the prominence of each successive accent (as suggested earlier by Pierrehumbert [15]). This was an encouraging result for linear models, because Bruce [16] had just shown that the same local "downstep" worked better than Gårding's superpositional model [17] in accounting for the decline in the portion after the focal accent in a large dataset of Swedish utterances.

Pierrehumbert & Beckman's revision to Liberman & Pierrehumbert's original proposal introduced a new parameter — local pitch range — and was motivated by our difficulties in modeling downstep and its interaction with focus in a large dataset of Japanese utterances. We found that in order to model L tone targets when one accentual phrase is downstepped and pragmatically subordinated to its neighbors, we had to specify a local pitch range "topline" for each accentual phrase, a value like the prominence value specified for each pitch accent in Pierrehumbert's original model but affecting all of the tones within the phrase rather than just the tones of the accent.

This approach is reminiscent of superpositional models such as Grønnum's [1] or Gårding's [17] in that the phonetic specification of local pitch range is embedded within the specification of the speaker's overall pitch range for that degree of vocal effort. It differs from these more radically superpositional models, however, in that the embedding is phonologized in the metrical representation of stress and phrasing. The phonology of intonation itself is not superpositional, because the local graph paper is not phonologized. Rather, local pitch range is represented by a single continuously variable phonetic value (the reference line height) locally specified for the phrase.

Ladd's [18] reanalysis of the *Anna came with Manny.* data similarly involves more local pitch range specification. However, he differentiates the more local kind of variation from the specification of overall pitch range by making only the latter a continuous phonetic dimension. That is, he phonologizes local pitch range by constraining the values it can take in terms of l/h vs h/l relationships specified on branching nodes in a hierarchical phonological structure with the accents as leaf nodes. My own prejudice is that this is phonologizing relationships that belong to be represented elsewhere. For example, "answer/background" is really a relationship of pragmatic subordination between the two phrases. The F0 data that we are modeling when we examine this relationship give an impression of discretely constrained relationships between the local pitch range values, but this is an artifact of contrasting mini-discourses that felicitously allow only two choices for the degree of pragmatic subordination between the compared phrases. Differentiating my interpretation from Ladd's would require something like an experiment in which speakers are successfully induced by the context to produce more degrees of pragmatic subordination between the two accents.

WHERE FROM HERE?

There are many other directions in which this comparison of models could go. For example, in the discussion above I have referred to large datasets from three typologically unlike languages: English, Swedish, and Japanese. We could extend the comparison of models if we gathered comparably comprehensive datasets from yet other typologically different languages. Shih's [19] data for Mandarin Chinese is one such dataset, and presents problems for all of our models. Her results show a lowering of successive tone targets, but not after tone 1, and to different extents elsewhere, so that targets are lowered more after tone 3 than after tone 2 or 4. This "differential downstep" is difficult to accommodate in any superpositional model that represents the lowering by a declination built into the graph paper at any level of grouping of tones. Shih's data also show focus having the general effect of locally raising high tone targets (an effect that linear models should accommodate ideally). However, in strings of tone 1 the targets following the focus also are raised and only gradually decline back to a more neutral value. That is, tone height does not immediately return to neutral (so that focus could be modeled as affecting only the focused tone's "prominence") nor does it stay at the higher level until some metrical domain edge (so that focus could be modeled as increasing the local pitch range value).

Another area where we need much more data is the behavior of intonational events that are low in the pitch range. Here our phonetic representation serves us ill, because F0 is not well-defined for the irregular phonations often seen in extremely low-pitched parts of the utterance. In the original English model described in [4], the degrees of freedom were limited by having a constant "baseline" value representing the bottom of the speaker's overall pitch range. This was motivated by the data's apparent reconfirmation of the long-standing finding that F0 values at the end of declarative contours in English seem relatively invariant compared to peak values. Hirschberg & Pierrehumbert's simulation of varying endpoints to reflect degree of topic embedding in a monologue [20] suggests that the invariance is an artifact of measuring endpoint values at the last point before creak makes F0 a bad measure of pitch. We need to find a better representation of pitch in these regions, perhaps by first doing basic perception studies to see how perceived pitch varies for different kinds of creaky phonation.

Also, I should like us to begin looking at physiology. What we know of F0 control makes it extremely doubtful that we will not find any very simple mapping between particular model parameters and particular muscles or parts of muscles. On the other hand, our preliminary data on subglottal pressure and EMG activity levels [2, 3] suggest that the variation in overall pitch range associated with changes in vocal effort levels can be separated from some sorts of variation in local accentual prominence in being created in large part by the increased subglottal pressure at louder vocal effort levels. It would be interesting to see whether physiological data support an understanding of more local pitch range variation as a local change in vocal effort.

I am sure the other participants in this symposium can add other areas in which they would like to see more data to enrich our models and our debate. So to respond to Ladd's question, I think we should leave this debate where it stands right now, and go have a beer so we can start fresh tomorrow in the laboratory.

REFERENCES

[1] Grønnum, N. (1995). "Superposition and subordination in intonation — a non-linear approach," This symposium.
[2] Erickson, D., Honda, K., Hirai, H., Beckman, M. E., & Niimi, S. (1994). "Global pitch range and the production of low tones in English intonation," *ICSLP '94*, vol. 2, pp. 651-654.
[3] Beckman, M. E., Erickson, D., Honda, K., Hirai, H., & Niimi, S. (1995). "Physiological correlates of global and local pitch range variation in the production of high tones in English," *13th ICPhS.*
[4] Liberman, M., & Pierrehumbert, J. (1984). "Intonational invariance under changes in pitch range and length," In M. Aranoff & R. Oehrle, eds., *Language Sound Structure*, Cambridge, MA: MIT Press.
[5] Ross, E. D., Edmonson, T. A., Seibert, G. B., & Chan, J.-L. (1992). "Affective exploitation of tone in Taiwanese: an acoustical study of 'tone latitude'," *J. Phon.*, vol. 20, pp. 441-456.
[6] Jun, S.-A., & Oh, M. (1994). "A prosodic analysis of three sentence types with 'WH' words in Korean," *ICSLP '94*, vol. 1, pp. 323-326.
[7] Miura, I., & Hara, N. (in press). "Production and perception of rhetorical questions in Osaka Japanese," *J. Phon.*
[8] Hirschberg, J., & Ward, G. (1992). "The influence of pitch range, duration, amplitude and spectral features on the interpretation of the rise-fall-rise intonation contour in English," *J. Phon.*, vol. 20, pp. 241-251.
[9] Herman, R. (1995). "Final lowering in Kipare," *Ohio St. U. Working Papers in Linguistics*, no. 45, pp. 36-55.
[10] Ward, G., & Hirschberg, J. (1985). "Implicating uncertainty: the pragmatics of fall-rise intonation," *Language*, vol. 61, pp. 747-776.
[11] Pierrehumbert, J. & Hirschberg, J. (1990). "The meaning of intonation contours in the interpretation of discourse," In P. R. Cohen, J. Morgan, & M. E. Pollack (Eds.), *Intentions in Communication*, pp. 271-311. Cambridge, MA: MIT Press.
[12] Gussenhoven, C. (1984). *On the Grammar and Semantics of Sentence Accents*, Dordrecht: Foris.
[13] Möbius, B. (1995). "Components of a quantitative model of German intonation," This symposium.
[14] Pierrehumbert, J. B., & Beckman, M. E. (1988). *Japanese Tone Structure*, Cambridge, MA: MIT Press.
[15] Pierrehumbert, J. (1980). *The Phonology and Phonetics of English Intonation*, Doctoral dissertation, MIT.
[16] Bruce, G. (1982). "Developing the Swedish intonational model," Working Papers, Lund, no. 22, pp. 51-116.
[17] Gårding, E. (1979). "Sentence intonation in Swedish," *Phonetica*, vol. 36, pp. 207-215.
[18] Ladd, D. R. (1995). "'Linear' and 'overlay' descriptions: an Autosegmental-Metrical middle way," This symposium.
[19] Shih, C.-L. (1988). "Tone and intonation in Mandarin," *Working Papers, Cornell Phonetics Laboratory*, no. 3, pp. 83-109.
[20] Hirschberg, J., & Pierrehumbert, J. (1986). "The intonational structuring of discourse," *24th ACL*, pp. 136-144.

COMPONENTS OF A QUANTITATIVE MODEL OF GERMAN INTONATION

Bernd Möbius
AT&T Bell Laboratories, Murray Hill, NJ, USA

ABSTRACT

In this paper a quantitative description of German intonation is presented. It will be demonstrated that intonation contours can be efficiently analyzed, and predicted, by interpreting the components and parameters of Fujisaki's model in terms of linguistic features and categories. It will also be argued that a superpositionally organized model is particularly suitable for a quantitatve description.

STRUCTURE OF INTONATION

Tone Sequences Or Layered Components?

Two major classes of intonation models have evolved in the course of the last two decades. There are, on the one hand, hierarchically organized models which interpret *F*0 contours as a complex pattern resulting from the superposition of several components. Their counterparts are usually seen in the models which claim that *F*0 contours are generated from a sequence of phonologically distinctive tones, or categorially different pitch accents, that are locally determined and do not interact.

Two quotations illustrate the competing points of view:

"[...] the pitch movements associated with accented syllables are themselves what make up sentence intonation [...] there is no layer or component of intonation separate from accent: intonation consists of a sequence of accents, or, to put it more generally, a sequence of tonal elements." ([9], p. 40)

"[...] Standard Danish intonational phenomena are structured in a hierarchically organized system, where components of smaller temporal scope are superposed on components of larger temporal domain [...] These components are simultaneous, parametric, non-categorical and highly interacting in their actual production." ([25], p. 2)

Ladd [9, 10] argues that although the tone sequence and the superpositional models diverge in formal and notational terms, they nevertheless may be more similar from a descriptive point of view than usually admitted. Although I agree with the argument ([24], p. 1041) that the two types of intonation models not only differ in formal respect but from a conceptual point of view as well, I don't think they are ultimately incompatible. As a matter of fact, in more recent publications (e.g., [11]) Ladd proposed a metrical approach that incorporates both linear and hierarchical elements.

The main difference between the 'pure' linear and overlaying models can be seen in how the relation between local movements and global trends in the intonation contour is defined, or, in other words, in the view of the relation between word accent and sentence intonation. The underlying problem is that word- and utterance-(or phrase-)prosodic aspects all express themselves by one and the same acoustic variable: the variation of fundamental frequency as a function of time. There is no way of deciding either by acoustic measurements or by perceptual criteria whether *F*0 movements are caused by accentuation or by intonation. A separation of these effects, however, can be done on a linguistic, i.e. more abstract, level of description. Here rules can be formulated that predict accent- or intonation-related patterns independent of, as well as in interaction with, each other.

Autosegmental theory allows for the independence of various levels of suprasegmental description and their respective effects on the intonation contour by an appropriate phonological representation. According to Edwards and Beckman [2], the most promising principle of intonation models ought to be seen in the capability to determine the effects of each individual level, and of their interactions. Although probably not intended by the authors, this is precisely the most important argument in favor of a hierarchical approach and of superpositional models of intonation. Thus, the conceptual gap between the different theories of intonation does not seem to be too wide to be bridged.

After presenting supporting data, I will continue this line of argument in the concluding section.

Motivation For A Superpositional Approach

Even among researchers representing different types of intonation models there is widespread agreement on the fact that the *F*0 contour of an utterance should be regarded as the complex result of effects exerted by a multitude of factors. Some of these factors are related to articulatory or segmental effects but others clearly have to be assigned to linguistic categories.

In contradiction to the explicit assumption in [20] that intonation is determined exclusively on a local level, there is ample evidence for non-local factors. In a study of utterances containing parentheses [8], the authors show that the intonation contour is interrupted by the parenthesis, and resumed right afterwards in a way the contour would have looked like in the 'same' utterance without parenthesis. Also, in [12] the authors explain how the first accent peak in an utterance is adjusted depending on the underlying syntactic constituent structure. Furthermore, there is some evidence that the speaker pre-plans the global aspects of the intonation contour, not only with respect to utterance-initial *F*0 values but to phrasing and inter-stress intervals as well [23].

These considerations obviously favor models that directly represent both global and local properties of intonation. These models also provide a way of extracting prosodic features related to the syntactic structure of the utterance and to sentence mode. Generally speaking, the analytical separation of all the potential factors considerably helps decide under which conditions and to what extent the concrete shape of a given *F*0 contour is determined by linguistic factors (including lexical tone), non-linguistic factors, such as, e.g., intrinsic and coarticulatory *F*0 variations, and speaker-dependent factors.

Superpositionally organized models lend themselves to such a quantitative approach: Contours generated by such a model result from an additive superposition of components that are in principle orthogonal to, or independent of, each other. The components in turn can be related to certain linguistic or non-linguistic categories. Thus, the factors contributing to the variability of *F*0 contours can be investigated separately. In addition, the temporal course pertinent to each individual component can be computed independently. A production-oriented model providing components for accentuation on the one hand and sentence or phrase intonation on the other hand and generating the pertinent patterns by means of parametric commands appears to be particularly promising.

The only approach exploiting the principle of superposition in a strictly mathematical sense is the model proposed by Fujisaki and co-workers (e.g., [5, 3, 4]). This particular model has several advantages. Since it satisfies the principle of superposition, the respective effect of a given factor can be determined for a predefined temporal segment or for a given linguistically or prosodically defined unit, such as a phrase or a stress group. For every desired point in time in the course of an utterance, the resulting *F*0 value can be computed. The values of the model parameters (see following section) are constant at least within one stress group. This data reduction can be an important aspect for certain applications like speech synthesis. The smooth contour resulting from the superposition of the model's components is appropriate for the approximation of naturally produced *F*0 contours.

Generally speaking, adequate models are expected to provide both predictive and explanatory elements [1]. In terms of prediction, models have to be as precise and quantitative as possible, ideally being mathematically formulated. A model provides explanations if it is capable of analyzing a complex system in such a way that both the effects of individual components and their combined results become apparent. Fujisaki's model meets both requirements; and all effects can be described uniquely by their causes.

The model does not, however, explain by itself why a given component behaves the way it does. The particular approach and the application presented in this paper aim at providing these explanations, especially by applying a linguistic interpretation of the model's components.

Another explanatory approach can be seen in the potential physiological founda-

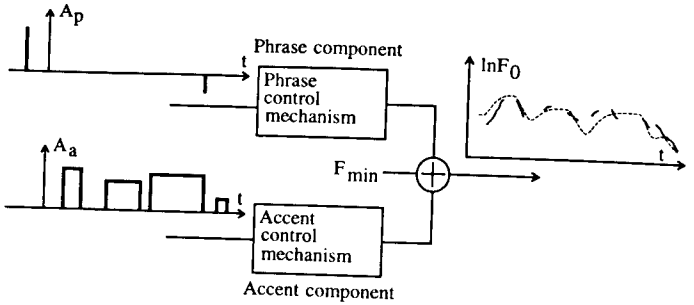


Figure 1. Block diagram of Fujisaki's quantitative model that additively superposes a basic Fo value (F_{min}), a phrase component, and an accent component on a logarithmic scale (lnFo). The control mechanisms of the components respond to impulse (phrase component) and rectangular commands (accent component), respectively (A_P = amplitude of phrase commands; A_a = amplitude of accent commands; t = time).

tion, in terms of laryngeal structures and interactions of laryngeal muscles, as discussed by Fujisaki [3]. His model is the only one I am aware of that explicitly includes a quantitative simulation of the *Fo* production and control mechanisms inherent in a human speaker; the approach is based on work by Öhman and Lindqvist [18].

The model represents each partial glottal mechanism of fundamental frequency control by a separate component. Although it does not include a component that models intrinsic or coarticulatory *Fo* variations, such a mechanism could easily be added in case it is considered essential for natural-sounding synthesis.

A QUANTITATIVE MODEL OF INTONATION

Since Fujisaki's model has been described by the original authors on many occasions, I will restrict myself to only presenting the most important properties of the model. I will focus instead on motivating the linguistic interpretation of the components as it emerged from applying the model to the analysis of German intonation.

The model additively superposes a basic *Fo* value (F_{min}), a phrase component, and an accent component on a logarithmic scale (Figure 1). The control mechanisms of the two components are realized as critically damped second-order systems responding to impulse functions in the case of the phrase component, and rectangular functions in the case of the accent component. These functions are generated by two different sets of parameters: the timing and amplitudes of the phrase commands as well as the damping factors of the phrase control mechanism on the one hand, and the amplitudes and the timing of the onsets and offsets of the accent commands as well as the damping factors of the accent control mechanism on the other hand. All these parameter values are constant for a defined time interval: the parameters of the phrase component within one prosodic phrase, the parameters of the accent component within one accent group, and the basic value F_{min} within the whole utterance.

The *Fo* contour of a given utterance can be decomposed into the components of the model by applying an analysis-by-synthesis procedure. This is achieved by successively optimizing the parameter values, eventually leading to a close approximation of the original *Fo* curve. Thus, the model provides a parametric representation of intonation contours.

Linguistic Interpretation

As I have argued in more detail elsewhere [14, 17], the quantitative description of intonation can be more efficient if modeling a given *Fo* contour and extracting the pertinent parameters is subjected to the constraints given by a linguistic and prosodic interpretation in the first place and by the criterion of optimal approximation in a mathematical sense only in the second place.

Here are the key elements of my interpretation of the model:

- The phrase component of the model represents the global slope and the slow variations of the *Fo* contour in the utterance. Obviously, the phrase component is very suitable to describe F0 declination since the phrase contour reaches its maximum rather early and descends monotonically along the major part of the utterance. Therefore, the contour that results from adding the basic value F_{min} to the phrase component serves as a baseline of the intonation contour, the magnitude of the phrase command amplitude being a direct measure for *Fo* declination in the utterance.

- Besides the obligatory utterance-initial phrase command, additional phrase commands are only provided at major syntactic boundaries, e.g., between main and subordinate clauses, thereby resetting the declination line. The procedure of inserting phrase commands wherever the criterion of optimal approximation seems to demand it [5] is rejected.

- The conspicuous final lowering of *Fo* which is regularly observed in declarative utterances and often in wh-questions is modeled by a negative phrase command. Likewise, we provide positive utterance-final phrase commands for other sentence modes, such as yes/no and echo questions. Thus, the phrase component of the model can be related to the linguistic category *sentence mode*, via the shape of the phrase contour and the underlying commands and parameter values. There are both global (the overall slope) and local (final rise or lowering) cues that contribute to differentiating between sentence modes.

- Local *Fo* movements that are associated with accented syllables are represented by the accent component and superposed onto the global contour. Closely following Thorsen's definition of *stress groups* [22] I apply an accent group concept, an accent group being defined as a prosodic unit that consists of an accented syllable optionally followed by any number of unaccented syllables. Accent groups are independent of word boundaries but sensitive to major syntactic boundaries, as will be shown below.

The concept of accent groups fits in the hierarchical structure of the model. While the linguistic category *sentence mode* is reflected in the phrase component, the linguistic feature *word accent* is manifested in the locations and shapes of accent commands. Consequently, the *Fo* course of a given accent group should be modeled by the contour generated by exactly one accent command. Thus, the parameter configurations of the accent component can be interpreted as correlates of the linguistic feature word accent.

ANALYSIS OF GERMAN INTONATION

Estimation Of Parameter Values

In principle, the parameter values that approximate the *Fo* contour of a given utterance can be determined automatically or by hand. Nevertheless, only an automatic procedure guarantees that the optimal values are extracted in an objective and reproducible way. Preliminary experiments showed that there are considerable intra- and interindividual divergencies when an interactive, i.e., partly manual method is used. Therefore, the parameter values of the model are determined by means of a computer program [19] that automatically approximates measured *Fo* contours by successively optimizing the parameters within the framework of the linguistic interpretation of the model. Input information is the *Fo* data for the given utterance and the locations of accent group boundaries.

Based on the principle of superposition, determination of the phrase command parameters and the basic value F_{min}, which is the first step in the algorithm, can be separated from the subsequent determination of the accent command parameters. The contour resulting from F_{min} and the phrase parameters is approximated to the measured *Fo* curve. Once the parameters of the phrase component are optimized, the resulting differential signal is interpreted by the accent component of the model.

The accent component is made up of partial contours that are in turn generated by accent commands. Each accent group is modeled by the contour resulting from exactly one accent command. Algorithmically speaking, the individual accent groups are processed from left to right in a non-iterative way.

Figure 2 illustrates the close approximation of a measured *Fo* contour.

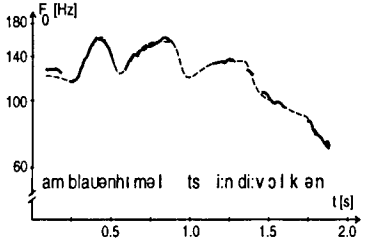


Figure 2. Close approximation (dashed line) of the Fo contour of the declarative utterance "Am blauen Himmel ziehen die Wolken" (male voice).

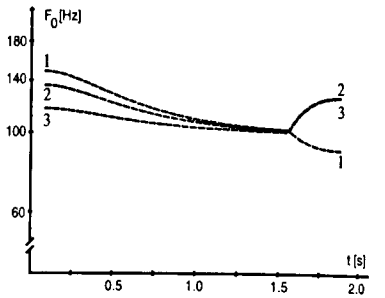


Figure 3. Typical phrase contours for the interrogative modes wh-question (1), yes/no question (2), and echo question(3), for Fmin=100 Hz.

Results

The speech materials cover declarative sentences with one or two major syntactic clauses, the latter realized as two prosodic phrases, and three types of interrogatives, namely echo, yes/no, and wh-questions. Speech data for six male and three female speakers were collected under 'laboratory' conditions.

The potential sources of variation of the parameter values were explored by means of statistical procedures, taking into account both linguistic and speaker-dependent factors. The results have been presented at full length elsewhere [14], so only the major trends and findings will be presented here.

Damping factors. The damping factors of the phrase and accent components are treated as constants. My experiments confirm the claim that the approximation of *Fo* contours is not impaired by this assumption [3]. For the phrase component, a standard value of 3.1/s is both appropriate for the purpose of approximation and reasonable as far as the physiological foundation of the model is concerned. A constant value of 16/s corresponding to the arithmetic mean for all speakers and all accent groups is suitable for the damping factor of the accent component.

Basic value *Fmin*. For all speakers, the dispersion of the basic value *Fmin* is relatively small, yielding 50% of the observed values within the range of about 3.0 Hz around the arithmetic mean for the respective speaker. This finding suggests that it is reasonable to keep *Fmin* constant for a given speaker. Typical values are 75-80 Hz for male speakers and 145-150 Hz for female speakers.

Phrase command timing and amplitude. Since the phrase component serves as a baseline to the intonation contour with the peak of each phrase contour coinciding with the beginning of the utterance, or the prosodic phrase, the exact timing of a phrase command directly depends upon the value of the damping factor (3.1/s). Therefore, the first phrase command is set at 323 ms before the onset of the utterance. This agrees with findings from studies on *Fo* production and control which reveal prephonatory activities of the laryngeal muscles [7].

Phrase command amplitudes are largely speaker, or rather speaker type, dependent. Sentence mode is the most important linguistic factor; it is globally signaled by the contour of the phrase component. While phrase contours of wh-questions are very similar to those of declaratives, yes/no-questions and the syntactically unmarked echo questions show a much less steep declination (see also [22] for Danish). Typical phrase contours for these three interrogative modes are shown in Figure 3. No consistent dependency of phrase command amplitude upon utterance duration or speech tempo was observed.

Accent command amplitude. The values of the accent command amplitudes split the speakers into two groups. The location of the accent group in the utterance turned out to be the most important linguistic factor. Utterance-final accent commands show significantly smaller amplitudes than accent commands in any other position in the utterance. Other important factors are the part-of-speech

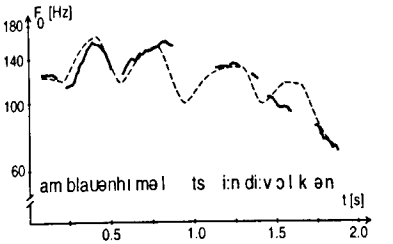


Figure 4. Rule-generated Fo contour (dashed line) compared to the original Fo contour of the declarative utterance "Am blauen Himmel ziehen die Wolken." (male voice); cf. Figure 2.

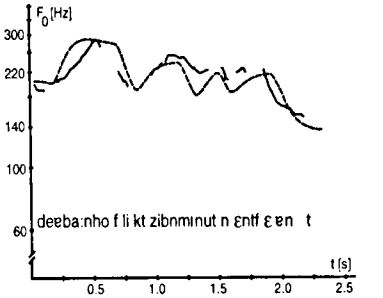


Figure 5. Rule-generated Fo contour (dashed line) compared to the original Fo contour of the declarative utterance "Der Bahnhof liegt sieben Minuten entfernt." (female voice).

class for the word carrying the accent, with nouns requiring higher amplitudes than other classes, and the presence of a phrase boundary. Amplitudes of accent commands preceding a phrase boundary tend to be about 25% higher than in other positions.

Accent command duration. Duration of an accent command can be reliably predicted from the duration of the respective accent group. There is a high correlation ($r = 0.84$) between these two variables, i.e., more than 70% of the variance observed in durations of accent commands can be explained by accent group duration. An effect of phrase-final lengthening is observed for several speakers.

Accent command position. The most important factor controlling the relative temporal position of an accent command within a given accent group is the location of the accent group in the utterance. While in non-final positions the temporal distance between the beginning of the accent group and the command onset is about 10% of the accent group duration, it tends toward zero in utterance-final accent groups.

Fo Synthesis By Rule

Parameters are adjusted by rules based on the analysis described above. The rules capture speaker dependent as well as linguistic features, such as sentence mode, sentence accent, phrase boundary signals, or word accent, and generate an artificial intonation contour for a given target utterance. The input information needed is the location of accented syllables in the utterance, the durations of accent groups, and, although less important, part-of-speech information for the words that carry accents.

Since the rules are based on the results of statistical analyses, the parameter values they provide are averages, producing contours that were not actually observed for any real speaker. On the other hand, they were shown to capture speaker-dependent features; they produce intonation patterns that to a fair degree correspond to what the modeled speaker could have produced. Thus, one should expect a mixture of frequently very good predictions with occasionally rather poor ones, the latter being due to either insufficient data or inadequate predictive power for a particular context.

Illustrations of *Fo* contours generated by rule are given in Figures 4, 5, and 6.

The adequacy of the rules was tested in a series of perceptual experiments whose results are presented elsewhere [15, 16]. The rules have been implemented in the German concatenative speech synthesis system HADIFIX developed at IKP Bonn [21].

CONCLUSION

In conclusion, I resume the controversial discussion of tone-sequential and superpositional intonation models, taking the prosodic marking of phrase boundaries as a starting point. The results presented here indicate that major syntactic boundaries invoke a resetting of declination, or the *Fo* baseline, which is realized in my quantitative model by inserting a phrase command. Additionally, signaling of the

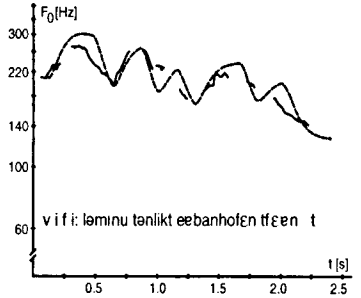


Figure 6. Rule-generated F0 contour (dashed line) compared to the original F0 contour of the wh-question "Wieviele Minuten liegt der Bahnhof entfernt?" (female voice).

boundary is enhanced by parameters of the accent component that are sensitive to major syntactic boundaries. The hierarchical structure of the model enhances elaborating the respective effects of the factors involved, even though both strategies of signaling phrase boundaries make use of the same phonetic variable.

It is important to note, though, that the notion of *hierarchy* is not necessarily an appropriate criterion for differentiating tone sequence and superpositional models, especially since its meaning is ambiguous. Both types of concepts contain hierarchical elements in the sense that utterances consist of prosodic phrases which in turn consist of accent groups or pitch accents; and even the most influential tone sequence model [20] provides a non-local element, i.e. declination. There is another meaning of *hierarchy*: making choices in various components of the prosodic system of a given language, higher levels having priority over, and setting constraints for, lower levels (cf. [25]). In the superpositional model presented here, however, there is no preponderance of one component over another.

Furthermore, since all the models discussed here (explicitly or implicitly) assume a mechanism of pre-planning in speech production, the difference between them should rather be seen in terms of how they represent this mechanism. Tone sequence models provide a higher *F0* onset in longer utterances but the relations between the individual pitch accents are not affected. According to [6], utterance length determines the slope of declination, short utterances having a steeper baseline, but not the utterance-initial *F0* value.

The formulae given by [13] in their version of the linear tone sequence approach are based on the analysis and approximation of intonation contours. Meaning is only assigned to the relations between pitch accents which are in turn defined by the feature of downstepping. However, it seems to be more appropriate to also assign meaning to the arguments in the formula, i.e., to the variables and constants. Genuinely superpositional models meet this requirement: The output behavior of the model as a response to the sum of several input signals can be predicted from the responses to each of the individual input signals.

Arguing in favor of a hierarchical organization of prosodic systems does not imply a rejection of phonological approaches. On the contrary, the integration of a superpositionally organized intonation model with an underlying phonological representation of the prosodic system of a given language is ultimately desirable. The phonological foundation of the quantitative model for German presented here remains a desideratum.

ACKNOWLEDGMENTS

The experiments presented in this paper were done at IKP, Univ. of Bonn, supported by grants from the German Research Council (DFG) and the German Federal Ministry of Research and Technology (BMFT). The author wishes to thank Grazyna Demenko, Wolfgang Hess, Julia Hirschberg, Matthias Pätzold, Thomas Portele, and Jan van Santen, for support and valuable discussions.

REFERENCES

[1] Cooper, F.S. (1983): "Some reflections on speech research", In P.F. MacNeilage (ed.), *The production of speech*, New York: Springer, pp. 275-290.

[2] Edwards, J., Beckman, M.E. (1988): "Articulatory timing and the prosodic interpretation of syllable duration", *Phonetica*, vol. 45, pp. 156-174.

[3] Fujisaki, H. (1983): "Dynamic characteristics of voice fundamental frequency in speech and singing", In P.F. MacNeilage (ed.), *The production of speech*, Berlin: Springer, pp. 39-55.

[4] Fujisaki, H. (1988): "A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour", In O. Fujimura (ed.), *Vocal physiology: voice production, mechanisms and functions*, New York: Raven, pp. 347-355.

[5] Fujisaki, H., Hirose, K., Ohta, K. (1979): "Acoustic features of the fundamental frequency contours of declarative sentences in Japanese", *Annual Bulletin of the Research Institute for Logopedics and Phoniatrics (Tokyo)*, vol. 13, pp. 163-172.

[6] Grønnum, N. (1990): "Prosodic parameters in a variety of Danish standard languages, with a view towards Swedish and German", *Phonetica*, vol. 47, pp. 182-214.

[7] Jafari, M., Wong, K.-H., Behbehani, K., Kondraske, G.V. (1989): "Performance characterization of human pitch control system: an acoustic approach", *Journal of the Acoustical Society of America*, vol. 85, pp. 1322-1328.

[8] Kutik, E.J., Cooper, W.E., Boyce, S. (1983): "Declination of fundamental frequency in speaker's production of parenthetical and main clauses", *Journal of the Acoustical Society of America*, vol. 73, pp. 1731-1738.

[9] Ladd, D.R. (1983): "Peak features and overall slope", In A. Cutler, D.R. Ladd (eds.), *Prosody: models and measurements*, Berlin: Springer, pp. 39-52.

[10] Ladd, D.R. (1983): "Phonological features of intonational peaks", *Language*, vol. 59, pp. 721-759.

[11] Ladd, D.R. (1993): "In defense of a metrical theory of intonational downstep", In H. van der Hulst, K. Snider (eds.), *The phonology of tone. The representation of tonal register*, Berlin: Mouton de Gruyter, pp. 109-132.

[12] Ladd, D.R., Johnson, C. (1987): "'Metrical' factors in the scaling of sentence-initial accent peaks", *Phonetica*, vol. 44, pp. 238-245.

[13] Liberman, M.Y., Pierrehumbert, J. (1984): "Intonational invariance under changes in pitch range and length", In M. Aronoff, R. Ochrle (eds.), *Language sound structure*, Cambridge: MIT Press, pp. 157-233.

[14] Möbius, B. (1993): *Ein quantitatives Modell der deutschen Intonation—Analyse und Synthese von Grundfrequenzverläufen*, Tübingen: Niemeyer.

[15] Möbius, B. (1993): "Perceptual evaluation of rule-generated intonation contours for German interrogatives", *Working Papers (Dept. of Linguistics and Phonetics, Univ. Lund) (=Proceedings of the ESCA Workshop on Prosody, Lund, 27-29 Sept. 1993)*, vol. 41, pp. 216-219.

[16] Möbius, B., Pätzold, M. (1992): "F0 synthesis based on a quantitative model of German intonation", *Proceedings of the International Conference on Spoken Language Processing (Banff, Alberta)*, vol. 1, pp. 361-364.

[17] Möbius, B., Pätzold, M., Hess, W. (1993): "Analysis and synthesis of German *F0* contours by means of Fujisaki's model", *Speech Communication*, vol. 13, pp. 53-61.

[18] Öhman, S.E.G., Lindqvist, J. (1966): "Analysis-by-synthesis of prosodic pitch contours", *Royal Inst. of Technology (Stockholm), STL-QPSR*, vol. 4 (1965), pp. 1-6.

[19] Pätzold, M. (1991): Nachbildung von Intonationskonturen mit dem Modell von Fujisaki - Implementierung des Algorithmus und erste Experimente mit ein- und zweiphrasigen Aussagesätzen (Ms., Univ. Bonn).

[20] Pierrehumbert, J. (1980): The phonology and phonetics of English intonation (Diss., MIT, Cambridge).

[21] Portele, T., Steffan, B., Preuss, R., Sendlmeier, W.F., Hess, W. (1992): "HADIFIX - a speech synthesis system for German", *Proceedings of the International Conference on Spoken Language Processing (Banff, Alberta)*, vol. 2, pp. 1227-1230.

[22] Thorsen, N. (1979): "Lexical stress, emphasis for contrast, and sentence intonation in advanced standard Copenhagen Danish", *Annual Report of the Institute of Phonetics (Univ. Copenhagen), ARIPUC*, vol. 13, pp. 59-85.

[23] Thorsen, N.G. (1985): "Intonation and text in Standard Danish", *Journal of the Acoustical Society of America*, vol. 77, pp. 1205-1216.

[24] Thorsen, N.G. (1986): "Sentence intonation in textual context - supplementary data", *Journal of the Acoustical Society of America*, vol. 80, pp. 1041-1047.

[25] Thorsen, N.G. (1988): "Standard Danish intonation", *Annual Report of the Institute of Phonetics (U Copenhagen), ARIPUC*, vol. 22, pp. 1-23.

"LINEAR" AND "OVERLAY" DESCRIPTIONS: AN AUTOSEGMENTAL-METRICAL MIDDLE WAY

D. Robert Ladd
Department of Linguistics, Edinburgh University

ABSTRACT

(1) All current descriptions of intonation involve strings of local events - pitch accents and phrase-final pitch movements. (2) Pitch range effects reflecting hierarchical structure affect individual accents, not whole domains. These two facts argue for neither an overlay model nor a radically linear one, but one with a "metrical" structure that specifies the relative height of prosodic constituents.

INTRODUCTION

Overlay vs. Linear?...

Let me begin by focusing in on what I see as the central disagreement at issue in this symposium. It is not, as Beckman's paper [2] reminds us, whether "superpositional" or "overlay" structure exists in F0. Anyone dealing with intonation must inevitably deal with its superpositional aspects. Anatomy (or culture, or both) makes some speakers' voices high and others low, some lively and others flat. Individual speakers may raise their voice or lower it, and may talk animatedly or monotonously. In many languages speakers may aim at a variety of distinctively different pitch levels or patterns, and in all languages (as far as anyone can determine; cf. [20]) there are segmental influences on the fine detail of F0. All of these effects can be fairly readily distinguished, and they interact in a way that is appropriately described as superposition.

I think we all agree on that much; we disagree on two more specific points. First, there is disagreement about which phenomena belong to which superposed layer. For example, are paragraph cues like raising and lowering the voice, or are they like choosing between distinctively different pitch patterns? This issue is unfortunately largely beyond the scope of this paper, though I will return to it tangentially at the end. Second, even if we agree which phenomena belong to the narrowly linguistic system we call intonation, we disagree about whether a superpositional model is appropriate for that system alone. We can all at least agree that intonation includes cues to phrasing and prominence and sentence-type, but some of us (e.g. [3,10,14,17] want to describe these in terms of strings of phonological "events", while others (e.g. [7,8,16,18,19]) want to model them as part of the general superpositional nature of F0. This is the issue I wish to discuss here.

...or Hierarchical plus Linear?

Specifically, I wish to defend a basically linear view of intonational phonology, but one in which hierarchical organisation - tree structure, roughly speaking - plays a significant role. I will focus on two phenomena: (1) mismatches between the phonetic extent of a pitch phenomenon and its functional "domain", and (2) hierarchical phrase-level effects or "paragraph cues". The first of these poses problems for the overlay approach, while the second is particularly difficult to accommodate in the radically linear approach associated with Pierrehumbert and Beckman's work [2,3,4,17]. Unfortunately, neither of these is a straightforward empirical problem, because in some sense F0 is really pretty easy to model: many approaches yield synthetic intonation that sounds fairly acceptable, and by any measure many models are capable of approximating natural data. However, the theoretical arguments can now be based on a much larger body of solid empirical data than was true twenty or even ten years ago.

Again, let me begin with a point of agreement. I believe that at some level of abstraction we are all describing intonation in terms of strings of events, and that therefore in some sense we are all operating with a linear phonology. For example, Grønnum [8,18,19], Möbius [16], and Fujisaki [7] all assume a string of accents, associated with specific syllables in the string of words. In the phonology, these could be represented autosegmentally, and an overlay model of intonational phonetics can be regarded as the "phonetic realisation" of that phonological string.

In saying this, I am not trying to argue that "those overlay people are basically doing what we linear types do". If anything, I acknowledge the possibility that "we linear types are basically doing what those overlay guys do". That is, I am quite prepared to accept that a Fujisaki-style "accent command" is a possible phonetic model of an autosegmental L+H* accent. In fact, the Anderson et al. implementation of the L+H* accent [1] looks remarkably like a Fujisaki-style accent command. My point is simply that all of us are looking at accents in very similar ways, and that one of the fundamental empirical tasks in which we are all engaged is to model *the phonetic details of accents in succession*. Where we disagree is over how to do it.

THE RELEVANCE OF DOMAINS

The basic assumption of the overlay approach is that every prosodic domain has characteristic pitch features, and that these pitch features *extend over the whole domain that they characterise*. Thus accent is a property of words, and dictates the shape of the pitch contour of words. "Declination" is a property of phrases and utterances, and perhaps also of paragraphs as well, and dictates the overall trend of pitch throughout the domain. And lexical tone, in languages that have it, is supposed to be a property of individual syllables.

This assumption is not unreasonable, but it is also not a single assumption. It is possible to imagine pitch features that characterise prosodic domains but do not extend over the whole domain they characterise. For example, we might imagine a sharp local rise marking the end of every intonational phrase, or the beginning of every paragraph, or whatever. That is, even if we admit the existence of a hierarchy of phonological domains - syllables, words, phrases, utterances - it does not follow that each one has its own slope or shape. Conversely, we can imagine pitch features that extend phonetically over something that is not a "domain" at all, but only an essentially accidental stretch of speech from one local event to another. Both these kinds of phenomena exist, and both are problematic for the overlay view.

The Hat Pattern

Taking the second case first, consider the "hat pattern" in the IPO description of Dutch intonation [9]. This consists of a rising pitch movement on one accented word, followed by a high level stretch, followed by a falling pitch movement on the next accented word. (This is, or is similar to, the intonation pattern shown in Möbius's Fig. 4.) In autosegmental or linear terms, the hat pattern is a high (H*) accent followed by a H*+L or H+L* accent, and presents no problems for the theoretical approach.

In an overlay model using Fujisaki's or Möbius's "accent commands", the hat pattern can easily be modelled - *phonetically* - by placing the beginning of the accent command at the first accented word and the end of the accent command at the second accented word. But (as Möbius would be the first to agree) this description makes no sense functionally or phonologically. For Möbius, each accent is the heart of a domain called an "accent group", and each accent group is supposed to have its own accent command. The hat pattern uncontroversially contains two accented words, but has the phonetic shape of a single accent group. This is a paradox for the overlay model.

I emphasise that the issue is not simply the ability to model F0 contours as phonetic objects. That, as I said above, is relatively easy to do (compared to, say, modelling segment duration). The real issue is to relate the parameters of one's phonetic model of F0 to distinct aspects of intonational function - i.e. linguistic categories such as phrase, accent, degree of emphasis, focus, and so on. If "accent group", as a linguistic construct, is a central part of one's phonetic model of accent, then it is a problem if we have to ignore the accent group in order to make the phonetic model "work". (I should add that Möbius is very careful to constrain his use of phonetic paramaters in a way that makes sense linguistically. However,

many proponents of overlay models - including Fujisaki - are rather less so.)

Edge Tones

There is also clear empirical evidence of pitch phenomena that relate functionally to a large domain but have localised phonetic manifestations. The best examples are *edge tones* of various kinds. I use "edge tone" as a cover term for what are variously called boundary tones, phrase tones, phrase accents, non-prominence-lending pitch movements, and so on. These have gone from being a source of theoretical bemusement 20 or 30 years ago to being a generally accepted part of the theoretical landscape today.

For example, in their early work on what became the IPO analysis of Dutch intonation, Cohen and 't Hart [6] explicitly pointed out the difference between the fall at the end of a statement and the rise at the end of a question as follows: "This so-called question rise need not occur in dominant words or even on prominent syllables, as opposed to final falls. In other words, this rise should not be taken to replace a final fall, but must be seen as an added feature" (p. 189). In the terms used here, the final fall is a pitch accent while the question rise is an edge tone. In current IPO terms, the final fall is accent-lending and must occur at the stressed syllable of a word that is prominent in the utterance; the question rise occurs at the very end of the utterance irrespective of the location of stress, is non-accent-lending, and must therefore be preceded somewhere in the utterance by an accent-lending pitch movement. Whichever way we express the difference, it is not a difference that would surprise anyone today - but in 1967 it required special comment.

Note that Möbius's model in effect includes edge tones, in the form of phrase commands at the *end* of a phrase. For the most part, phrase commands occur at the beginning of phrases, and serve to model the course of declination. In addition, however, Möbius uses them at phrase ends: a normal phrase command to model a phrase-final rise in pitch (a high boundary tone or IPO type 2 rise), and a *negative* phrase command to model a phrase-final drop in pitch. As I noted above, Möbius insists on some sort of linguistic or functional constraints on the location of the two types of command: accent commands are for phonetic effects related to prominent words, and phrase commands for phonetic effects related to phrases. His use of a phrase command to model phrase-final pitch movements is therefore commendable in principle; the only problem is that, as Liberman and Pierrehumbert have shown [14], the phrase command is really not a very accurate phonetic model of what happens at the ends of phrases. In effect, for phonetic accuracy, Möbius might better use half an accent command to model these boundary phenomena, but this route is closed to him on theoretical grounds.

Significance of Edge Tones

One of the first works to demonstrate the need for edge tones, Gösta Bruce's PhD thesis [5], also makes clear their significance for the overlay-vs.-linear debate. Bruce's specific concern was to develop an account of how lexical accent distinctions in Stockholm Swedish are manifested phonetically in different sentence contexts, but his solution to this problem lays the foundation for a more general theory of how word-level and sentence-level features interact.

In Swedish, the main stressed syllable of each word, in addition to being stressed, bears one of two accents, often called simply Accent 1 and Accent 2. The phonetic difference between the two accents is very striking in some environments and exceedingly subtle in others, but it typically involves a difference in the pitch contours of words and is therefore often described as a difference of "pitch accent". In Stockholm Swedish, the phonetic difference between the two accent types in citation form is superficially a difference between single peaked (Accent 1) and double peaked (Accent 2) pitch contours, as can be seen in Fig. 1.

Bruce's work made clear that the citation form contours involve an interaction between word level accent features and phrase- or sentence-level intonation features. In some sense this was never in doubt, but Bruce's breakthrough was to state explicitly what the accentual and intonational features are. Specifically, he established that the genuinely distinctive feature for the two accent types is the *alignment of an*

Figure 1. F0 contours of citation forms of "Accent 1" and "Accent 2" in Stockholm Swedish. Adapted from Bruce [5]. These contours are for two-syllable words stressed on the first syllable. The thinner and thicker line sections show the duration of the consonants and vowels respectively.

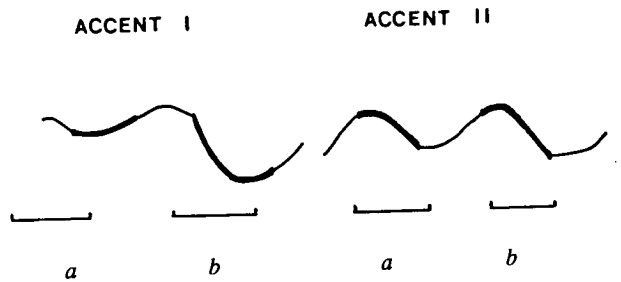


Figure 2. Bruce's analysis of the contours from Fig. 1, showing (a) the different alignment of the accentual fall, and (b) the high-low sequence that signals the end of the phrase and of the utterance.

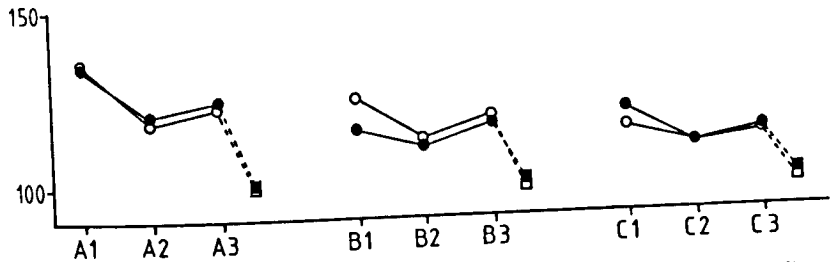


Figure 3. Sample data from Ladd's study of hierarchical effects in nested declination [10]. The circles show the mean F0, for one speaker, of the three accentual peaks in three consecutive clauses A, B, C; the squares show the clause-final low F0 endpoint. Filled circles and squares show data from the A and B but C *structure; open ones show the* A but B and C *structure. For more detail see text.*

underlying pitch peak with the stressed vowel. In Accent 1 the peak precedes the onset of the stressed vowel by a considerable extent, so that if there are no preceding unstressed syllables the stressed vowel simply begins mid or low, while in Accent 2 the peak and the subsequent drop in pitch more or less coincide in time with the stressed vowel and are therefore always present in the phonetic F0 contour.

The single peaked and double peaked word contours in citation forms result from the interaction of these invariant word accent features with features of sentence intonation. Crucially, these "features of sentence intonation" are not overall trends, but edge tones - a late peak or "phrase accent", and a final fall to the bottom of the range. When these edge tones occur after Accent 2, in which the accent has already produced a clear peak and fall on the accented vowel, the result is a "second" peak. But when they follow Accent 1, in which the accentual high may not be realised as such, the result is a phonetic rise across the accented vowel to the single peak in the utterance. This analysis is shown in Fig. 2.

In addition to providing an elegant and convincing solution for a long-standing problem of Scandinavian phonology, Bruce's analysis clearly shows the nature of the interaction between pitch features whose function relates to domains of different sizes: the word accents and the sentence level intonation features interact not by overlaying small-domain features on large-domain ones, but as a *sequence of phonetic events*. The sentence-level features affecting the word accent contours in citation form are not global shapes or trends but localisable phonetic events. This does not, of course, establish that phonological structure based on superposition is impossible, but it strengthens the argument for a rigorously linear phonological model, because it shows that such a model provides an accurate account of a case that prima facie might be expected to fall within the scope of the overlay approach. That is, if large-domain phonological specifications are unnecessary in these cases, then parsimonious theorising suggests we ought to try to do without them altogether.

HIERARCHICAL PITCH RANGE

The second kind of phenomenon I want to discuss is the manipulation of pitch range to convey the overall organisation of discourses - what we can informally lump together under the heading *paragraph cues*. I have proposed elsewhere [10,11,12,13] that these are properly thought of in terms of *abstract relations of relative height* between prosodic constituents of various sizes. That is, paragraph cues involve phonological relations in a metrical tree or similar structure, the phonological relations being used to express not relative prominence (as in standard metrical phonology, e.g. [15]) but relative pitch range. Thus the distinction between downstepping and non-downstepping accents can be represented as the difference between the following two "metrical" relations:

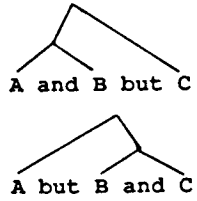
(Here T* is any pitch accent, the h-l relation means the second pitch accent is downstepped, and the l-h relation means it is not.) I have suggested that this approach could be extended to relations between higher level prosodic domains as well, and that the details of the nested relative-height relations could be used to account for detailed differences of relative height. My previous proposals are admittedly pretty sketchy about exactly how hierarchical relations translate into quantitative parameters, a failing I am not going to remedy here.

This metrical/hierarchical approach, it seems to me, falls neatly between the radical linear approach and the overlay approach. It agrees with the overlay approach in potentially being able to accommodate the indefinite nesting of prosodic domains of increasing size (phrase, sentence, paragraph, etc.). But it agrees with the linear approach in conceiving of the whole problem as being one of specifying *local* phonetic properties of individual accents - in this case, specifying the pitch range of accents based on their position in a hierarchical prosodic structure and the details of the pitch range relations specified in that structure - rather than specifying global phonetic properties of whole domains.

I believe that this metrical approach to relative pitch range is superior to either the strictly linear or overlay approaches. Let me briefly try to defend that claim.

The Overlay View

First let me compare the metrical view with the overlay approach. In [10] I investigated the details of "nested downtrends" - declination-within-declination effects of the sort that form an important argument for the overlay view (e.g. [18,19]). Specifically, I compared three-clause sentences of the form *A and B but C* or *A but B and C*, where A, B, and C are the clauses. There was very clear evidence of nested declination in all cases. However, the results also showed that the duration of the pauses between the clauses, and the height of the initial accent peaks in clauses B and C, is affected by the hierarchical organisation: in both structures the *but*-boundary is "stronger" (in the sense of being preceded by a longer pause and followed by a higher accent peak) than the *and*-boundary at the comparable location in the other structure (see Fig. 3). That is, the results seem to reflect a difference of hierarchical structure best represented as follows:



According to the overlay view, we should expect to be able to account for the observed differences in terms of a declination component for the sentence as a whole, one for each clause, and, crucially, one for the intermediate constituent *A and B* or *B and C*. But given the phonetic details of my results - specifically, the fact that the pitch range differences were concentrated on the initial accents of the B and C clauses - I see no ready way of modelling this with simple overlaid components, and no adherent of the overlay approach has come forward since the publication of my study to show how it can be done. I believe that close investigation of similar cases would reveal equally systematic manipulation of the pitch range of *individual* accents, and that the only way to account for these in an overlay model would be to posit implausibly complex and specific long-domain components, with bumps and dips in just the right places.

The Radical Linear View

The case for the radical linear approach to paragraph cues is, if anything, even worse. Proponents of this view have put themselves in the position of having to ignore a great deal of lawful or systematic phonetic variation, by maintaining that it is all gradient and paralinguistic. Specifically, Beckman and Pierrehumbert argue that pitch range effects like those just discussed are the result of individual phrase-by-phrase settings of a gradiently variable parameter that sets the initial pitch range for the phrase as a whole. The variability of this parameter is not phonological, but is said to be paralinguistic and to reflect the newness of the topic.

It is clear from many studies that overall pitch range is a powerful signal of speaker interest or emotional involvement. Extending this paralinguistic function of pitch range into the realm of grammar, Beckman and Pierrehumbert [3] claim that "pitch range is raised when initiating a new topic, and lowered when concluding one". This implies that "for discourse segments consisting of only one topic, a downtrend is accordingly predicted" (p.300). In explaining the data from the *A and B but C* study, they would presumably have to say that the downtrend is moderated at the *but*-boundary because the topic of the clause that follows is new, or at least not quite as old as the topic of the clause that follows an *and*-boundary.

Actual refutation or falsification of such a proposal is difficult. Nearly everyone would agree that paralinguistic factors are involved in intonation somehow, and at our present stage of understanding it is difficult to draw the line between those factors and phenomena that are genuinely linguistic. However, in my view the paralinguistic account of data like the *A and B but C* data is distinctly implausible. There are three reasons for this.

First, the match between structurally

defined boundary strength and pitch range is quite precise, yet the paralinguistic explanation rules out explicit reference to structure. If we claim that the pitch range data depend only on the speaker's estimation of the newness or interest of a particular clause, we must surely also give some explanation for the close coincidence between the structure and the paralinguistic signals of newness.

Second, it is insufficient to come up with a paralinguistic explanation only for pitch range, since the data showed that there were consistent effects on the duration of inter-phrase pauses as well. In my view, a far more plausible explanation is that details of both duration and pitch range are governed by linguistic differences in the hierarchical organisation of prosodic domains.

Finally, as has been demonstrated in Pierrehumbert's own work [17], it is actually quite easy to distinguish quantitatively between uses of pitch range that are unarguably paralinguistic, such as raising the voice for added overall emphasis, and what I am calling relations of relative height. Experiments in which speakers produce specific contour types with different overall pitch ranges show clearly that the constant pitch level proportions of one peak to another are maintained, regardless of the experimental manipulation of overall range. In my view this makes it implausible, *pace* Beckman and Pierrehumbert [4], to lump all these pitch range effects together as "paralinguistic".

Moreover, as I have discussed elsewhere [12], the problems with the Beckman-Pierrehumbert approach go beyond the difficulties of accounting for detailed structural effects in nested declination. An even more serious conceptual difficulty, in my view, is that Beckman and Pierrehumbert model downtrends in two quite different ways. At the lowest level of prosodic organisation - within what they call the "intermediate phrase" - downtrends are the result of phonologically triggered "catathesis" or *downstep*. At all higher levels, downtrends are the result of the paralinguistic signalling of newness, and merely "mimic" phonologically determined downstep. Why one should mimic the other is never made clear; it would seem more appropriate, given two similar phenomena, to try to ascribe them to similar causes.

This is not an isolated problem. There are other apparently distinctive pitch range relations that can hold between prosodic domains of quite different sizes. One is the "answer-background" relation posited by Liberman and Pierrehumbert [14]. This is seen in the relative pitch range on the accent peaks in the sentence *Anna came with Manny* (answering a question such as "What about Manny? Who came with him?"). This pitch range relation can hold between long and complex phrases, as in a possible rendition of the sentence *I'd actually like to stop talking and go out in search of a beer, if only I could get my point across to my fellow panellists*. But for Beckman and Pierrehumbert this latter utterance would involve at least 3 or 4 separate paralinguistic choices of pitch range, one for each intermediate phrase in succession. This view leaves us with no way to express the fact that the pitch range relation between the whole first half of the sentence and the whole second half is intuitively the same at that between *Anna* and *Manny*.

Another distinctive pitch range relation - perhaps it is the same as the preceding one, although it feels different because the lower phrase ends with a low final boundary instead of a high - is used in alternative questions. The alternatives can be as small as a single accent, or as long as an entire complex utterance. An example of two accents related this way is: *Do you want coffee or tea?* An example of two utterances related this way is: *Do you think we ought to leave this debate where it stands right now? Or would it be better to carry on until one side acknowledges that the other is right?* Cases like these are easy to accommodate in a theory in which pitch range relations can (indeed, must) hold between constituents at all levels of the prosodic hierarchy for which pitch features are specified. But in the radical linear view, most of the phenomena just described have to be treated as unsystematic and paralinguistic, and similarities ascribed to "mimicry". This raises the issue of which phenomena belong centrally to intonation; unfortunately, as I said at the beginning of the paper, there is no space to discuss this here.

CONCLUSION

There are no real conclusions yet, and there is space only for an observation. In my view, the most important point to keep in mind as we work toward a resolution of the linear/overlay debate is that the issue is not merely one of phonetic modelling, but of phonetic modelling constrained by assumptions about linguistic structure and function. This means that there are two approaches to evaluating competing models: one is to assess their accuracy as phonetic models, and the other is to examine their linguistic assumptions. The latter approach is the one I have taken here.

REFERENCES

[1] Anderson, M.; Pierrehumbert, J. B.; Liberman, M. (1984). "Synthesis by Rule of English Intonation Patterns." In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2.8.2-2.8.4. New York: IEEE.

[2] Beckman, M. E. (1995). "Local shapes and global trends". This volume.

[3] Beckman, M. E.; Pierrehumbert, J. B. (1986) "Intonational Structure in English and Japanese." *Phonology Yearbook* 3, 255-310.

[4] Beckman, M. E.; Pierrehumbert, J. B. (1992) "Comments on chapters 14 and 15." In G. J. Docherty and D. R. Ladd (eds.), *Gesture, Segment, Prosody: Papers in Laboratory Phonology II*. Cambridge: Cambridge University Press, pp. 387-397.

[5] Bruce, G. (1977). *Swedish word accents in sentence perspective*. Gleerup, Lund.

[6] Cohen, A.; 't Hart, J. (1967). "On the anatomy of intonation". *Lingua* 19, 177-192.

[7] Fujisaki, H.; Sudo, H. (1971). "A generative model for the prosody of connected speech in Japanese." *Annual Rept. Engineering Resch. Inst., Univ. Tokyo* 30, 75-80.

[8] Grønnum, Nina (1995). "Superposition and subordination in intonation - a non-linear approach". This volume.

[9] 't Hart, J.; Collier, R.; Cohen, A. (1990). *A perceptual study of intonation: An experimental-phonetic approach*. Cambridge: Cambridge University Press.

[10] Ladd, D. R. (1988). "Declination 'reset' and the hierarchical organization of utterances." *JASA* 84, 530-544.

[11] Ladd, D. R. (1990). "Metrical representation of pitch register." In J. Kingston and M. Beckman (eds.), *Papers in Laboratory Phonology I*. Cambridge: Cambridge University Press, pp. 35-57.

[12] Ladd, D. R. (1993). "In defense of a metrical theory of intonational downstep." In H.v.d.Hulst and K.Snider (eds.), *The Representation of Tonal Register*, Dordrecht: Foris Publications, pp. 109-132.

[13] Ladd, D. R. (1993). "Constraints on the gradient variability of pitch range (or) Pitch Level 4 Lives!". In P. Keating (ed.), *Papers in Laboratory Phonology III*. Cambridge: Cambridge University Press, pp. 43-63.

[14] Liberman, M.; Pierrehumbert, J. B. (1984). "Intonational invariance under changes in pitch range and length." In M. Aronoff and R. Oerhle (eds.) *Language sound structure*. MIT Press, Cambridge, pp. 157-233.

[15] Liberman, M.; Prince, A. (1977). "On stress and linguistic rhythm." *Linguistic Inquiry* 8, 249-336.

[16] Möbius, B. (1995). "Components of a quantitative model of German intonation". This volume.

[17] Pierrehumbert, J. B. (1980). *The Phonology and Phonetics of English Intonation*. MIT PhD Thesis, published 1988 by Indiana University Linguistics Club.

[18] Thorsen (Grønnum), N. (1980). "A study of the perception of sentence intonation - evidence from Danish." *JASA* 67, 1014-1030.

[19] Thorsen (Grønnum), N. (1985). "Intonation and text in Standard Danish." *JASA* 77, 1205-1216.

[20] Whalen, D. H.; Levitt, A. G. (1995). "The universality of intrinsic F0 of vowels". To appear in *Journal of Phonetics*.

SUPERPOSITION AND SUBORDINATION IN INTONATION A NON-LINEAR APPROACH

Nina Grønnum
Institute of General and Applied Linguistics
University of Copenhagen

ABSTRACT

After a brief presentation of the preliminaries and my presuppositions, a model of speakers' production of standard Danish intonation is presented. Its basic property is the layered, superpositional organization of its components - where the manifestation of components at lower levels of the linguistic hierarchy is subordinate to components at higher levels.

INTRODUCTION

A symposium at the quadrennial international congress of phonetic sciences is an appropriate occasion to remind oneself of and clearly state the underlying assumptions, the tacit goals, the ultimate ambition, and the implicit restrictions of our descriptions of intonational structure. They almost certainly are not identical across authors, and so mine are presented below. They are to be understood as programmatic in nature, rather than axiomatic, and - for reasons of space - they are also rather telegrammatic.

Since other views of intonational structure are presented and summarized by the other participants at this symposium I shall concentrate on my own, and only note in passing that similar views about intonation as a layered structure are to be found in [3, 5, 6, 8, 11, 12, 15]. Furthermore, the documentation for the descriptive adequacy of the model does not find room here and will have to be sought in previous publications, all of which are referenced and most of which are summarized in [7].

PRELIMINARIES

Intonation

'Intonation' encompasses all the linguistically relevant, suprasegmental, non-lexical aspects of the fundamental frequency (F_0) variation - or its perceptual correlate, the pitch variation - through the course of spoken utterances. 'Suprasegmental' removes the intrinsic F_0 characteristics of segments and their coarticulation. The segmentally induced variation is generally supposed to be beyond the speaker's conscious control, although it is equally generally assumed to be perceptually relevant for the identification of segments, and probably should be included in the low level rules for synthetic speech, if it is to sound natural. 'Non-lexical' excludes syllable tones and word accents.

The crucial term here is 'linguistically relevant'. But the line cannot always be unambiguously drawn between linguistic and para-linguistic in intonation. Prototypical functions in either domain are easily established: linguistically relevant is the cuing of (1) various types of prominence (reduced stress, normal stress, default sentence accent, focal accent, emphasis for contrast), (2) prosodic boundaries at various levels and (3) speech act function (imperative, declarative, interrogative). Typical para-linguistic meaning conveyed by certain aspects of F_0/pitch variation (inter alia) will be everything that characterizes an individual speaker, like sex, age, and present state of health and mind. But - to mention just one example - how to classify speaking style? Is that a linguistic or a para-linguistic parameter? Or is that not a decision which should be made universally? I have excluded speaking style from amongst the intonational parameters of my model, but more out of necessity - for want of relevant data, than for any more principled reason. -- Beckman [this symposium] takes a different attitude to para-linguistic aspects.

Models

Intonation can be modelled from various perspectives, for various purposes: we can aim at speakers' production or their perception; or models may be adapted to the demands of synthetic speech or automatic speech recognition,

respectively, procedures which do not necessarily parallel human processes. Furthermore, speaker and listener behaviour may be modelled at various levels of abstraction. All of us, in this symposium, have been mainly concerned with the modelling of production data, and mainly acoustic data at that, and the following is limited to models of human production of intonation.

To be a model, a description of intonation must do more than merely depict or replicate singular events, individual items and individual speakers - it must at least be a generalized account, generalized over typical speakers of a given speech community speaking in a given speech style and also over different exemplars of a given utterance type. I think it uncontroversial to also demand that it account not only for available data, but correctly predict new utterances as well. - Over and above that, the level of ambition may be highly variable, in terms of multiplicity of input to the model and also in claims about its universality. - Models, accordingly, are representations or formulae which mediate, back and forth, between a stage upstream in the human production of linguistic utterances where all the morpho-syntactic and lexical information has been supplied, and the next one where significant and distinctive intonational information is inserted, and from which the phonetic F_0 implementation can be derived by rule at a lower level and presumably be translated into neural commands to activate the physiological production system, alternatively into commands to drive a speech synthesizer. This level is often referred to as the 'phonological level'. It is an unfortunate term, however, with its connotations of 'minimal units to bring about a difference in (lexical) meaning' and - particularly - 'the double articulation' of language. Firstly, I do not think it feasible or expedient to phonologize differences in F_0 or pitch contours which are merely the acoustic or perceptual correlates of a contrast in another linguistic dimension, namely stress. Secondly, outside the realm of stress, intonation does not differentiate but carries meaning, though not of a lexical sort, cf. above. But it does not do so autonomously, only in an intricate interplay with syntax, semantics and pragmatics. Finally, I think it highly probable that intonation, except explicit local boundary phenomena, is perceived in relational terms as holistic gestalts. All this makes an analogy to paradigmatic segmental contrast rather forced. If 'intonational phonology' is merely synonymous with 'abstract representation of intonation', then the latter is perhaps a more fortunate term.

Universals

The general search for universals in linguistics has had an impact in intonation research as well. I am perhaps more sceptical than most about statements to the effect that, e.g., "all languages have boundary tones; all languages have sentence accents; questions end in final rises", etc. Languages sound so vastly different, intonation-wise, and I do not see why these huge impressionistic differences should not be reflections of fundamental principled differences in the abstract representation of their intonation, in the underlying components which speakers manipulate and the way they are organized. But to the extent that a model can be adapted without undue complication to typologically different intonation systems, it is of course the more powerful one.

Adequacy

There are several routes to models of intonation, from divine inspiration, through qualified introspection if the language is a familiar one, to laborious and time-consuming analyses of acoustic data, or a mixture of these. Likewise, there are several measures of a production model's adequacy. First of all, does it produce an acceptable output? (Note that it may do so without laying any claims to psychological reality.) This can be tested in speech synthesis. Then, how likely is it as a representation of speakers' behaviour, granted what we know about the physiology of F_0 production and other aspects of speech? How cognitively real is it? How well does it

lend itself to typological research, i.e. how easily are parameters added or deleted or modified without debilitating the model's functioning?

A MODEL OF STANDARD DANISH INTONATION

The model presented here is derived from acoustic analyses of a considerable amount of speech material from a fair number of speakers, and some perceptual experiments. It is a generalized description of the central linguistically relevant aspects of F_0 contours in short texts in informal but distinct monologue. It is also a hypothesis about speakers' internal representation, about the nature of the components and how they interact to turn a string of semantically coherent sentences into a series of prosodically coherent utterances. It lays no claim to universality, though by uncomplicated incorporation of extra parameters and by proper quantification and adjustment, it will account for a rather rich variety of Danish regional languages. It produces acceptable sounding synthetic speech.

The ultimate ambition, of course, is to be able to account for any style of speech and any syntactic structure in Danish. -- The model's strongest present limitation is its restriction to informal but distinct monologue, i.e. one-way communication, and read speech at that, based as it is on analyses of speech produced under laboratory conditions. However, two fundamental features, the recurrence of the F_0 pattern associated with stress and the quasi-rectilinear slope of utterance intonation contours have been demonstrated also in informal spontaneous speech in interviews ([4]), and I think it justified to assume that the model can serve at least as point of departure for the description of Danish speakers in any speech style. -- Syntactic boundaries below the sentence level are not incorporated in the model either. Again, this is not from any a priori theoretical preclusion of their relevance, but for lack of relevant data. From an early study it appeared that syntactic boundaries within simple, though long, sentences have no direct reflection in the intonation, but obviously there are prosodic boundaries affiliated with a number of syntactic ones in complex sentence structures, and the model should be expanded accordingly when the necessary experiments have been performed.

The textual contour

Short texts are characterized by a gradual global descent, the textual contour, *T* in the figure. Its onset and offset are defined by the first and last stressed syllable in the first and last utterance, respectively. It typically spans half an octave. Superposed on *T* are the individually sloping utterance contours, *U1-3*. Utterance onsets (defined by the first stressed syllable) lower gradually through the text and typically span 3 semitones. Utterance offsets (defined by the last stressed syllable) likewise lower gradually but typically span only 2 semitones, to the effect that utterance slopes are slightly steeper initially than finally in the text - in utterances of equal length and function. In texts of more than four utterances the medial part of the textual contour levels out. It is unreasonable to expect that speakers be able to manipulate lowering of successive utterance onset to a degree finer than 1 semitone, particularly in view of their relatively large mutual temporal distance.

Coordinate main clauses are less slanted relative to the global textual contour than are a sequence of terminal declaratives (not depicted in the figure).

The utterance contour

Utterance contours are defined solely by the string of stressed syllables, because: (1) Local rise-falls depend for their existence upon the presence of unstressed syllables in the stress group. In a succession of stressed syllables there is no upwards deflection of the F_0 course between them. (2) The stressed syllables are frequency scaled in relation to each other without regard to the presence or not of any rise-falls in the surroundings. For a given intonation type (cf. below), the range spanned by the contour is constant. In utterances of up to five stress groups they are equidistantly spaced in frequency, in intervals which are inversely proportional to their number.

But their timing will depend on stress group length (syllable structure and number of unstressed syllables), cp. *a1* and *a2* in *U3*. The local deflections (the 'highs') in the F_0 course have no independent role in the shaping of grosser trends in F_0 contours.

Further characteristics of utterance slopes are: If the utterance is long, its contour will not lower continuously but be broken into prosodic phrase contours, with a slight resetting between them, cf. *U2*, granted that the break does not cut up the utterance in an unacceptable manner, cf. below. The organization of the phrase contours relative to the superordinate utterance contour is analogous to the organization of utterances in the text: phrases descend along the utterance contour while each phrase is associated with its own slope, so the first phrase onset (defined by the first stressed syllable) is higher than the last phrase onset, and the first phrase offset is higher than the last phrase offset, cf. *P1* and *P2* in *U2*. Intermediate phrases have intermediate onsets and offsets, but above four prosodic phrases the utterance contour must level out medially, so as not to frustrate speakers' control over step down magnitude between phrase onsets.

As mentioned above, the syntax-prosody interface in complex sentence structures is largely unexplored in Danish, though see [14]. But the following is valid for simple sentences in isolation or in combination. (1) The syntactic structure of short sentences is not reflected in their intonation contour. The order of constituents does not matter, nor their internal structure. (2) Longer utterances are produced as a descending sequence of sloping prosodic phrases, but the conditions governing the location of the breaks are complex. (a) A prosodic phrase must contain at least two stress groups. (b) Prosodic phrases tend to be of equal length. (c) But this tendency is easily overruled: the prosodic boundary cannot occur within a syntactic constituent, nor between syntactic constituents which are semantically coherent. Thus, "Der går mange store Røde Kors busser til Grosny i Tjetjenien i aften." (Many big Red Cross buses depart for Grozny in Chechenya this evening.) will most likely be produced in one sweep, in spite of its length, and in spite of the formal boundary before the place complement, because buses are intimately associated with their destination (or their point of departure). (3) Coordinate main clauses are more likely to be separated by a resetting of the intonation contour than subordinate and main clause constructions (irrespective of their ordering). (4) Individual utterance contours are steeper, also the text final one, and demonstrate a greater amount of resetting between them, in a succession of terminal declaratives than in a corresponding string of coordinate main clauses. This difference, induced by different lexico-syntactic boundary conditions, at least hints at a solution to Ladd's *ABC*-problem [this symposium].

Utterance contours vary between most steeply sloping (in declarative sentences used conventionally) and horizontal (in sentences which are not marked lexically or syntactically for their interrogative function). Other questions and non-final clauses fill in the intermediate space, with a clear tendency for a trade-off between lexical/syntactic markers of their function and the slope of the intonation contour, cf. *a*, *b* and *c* in *U1*. The steepest contours typically span 4 semitones initially in the text and 3 semitones finally. The choice of contour slope for a given utterance is determined by syntactic and - not least - pragmatic factors, in accordance with, I propose, principles of markedness and typicality as follows: By definition, unmarked intonation is associated with syntactically unmarked sentences used conventionally. That makes the steepest contour - which accompanies conventional declaratives - unmarked and any less falling contour marked. Typical intonation is the contour which accompanies any given sentence type when it is used conventionally. Thus, a conventional Yes/No question will have a slope somewhere between unmarked and the horizontal, i.e. it is marked but typical. Any deviation from the typical intonation will have im-

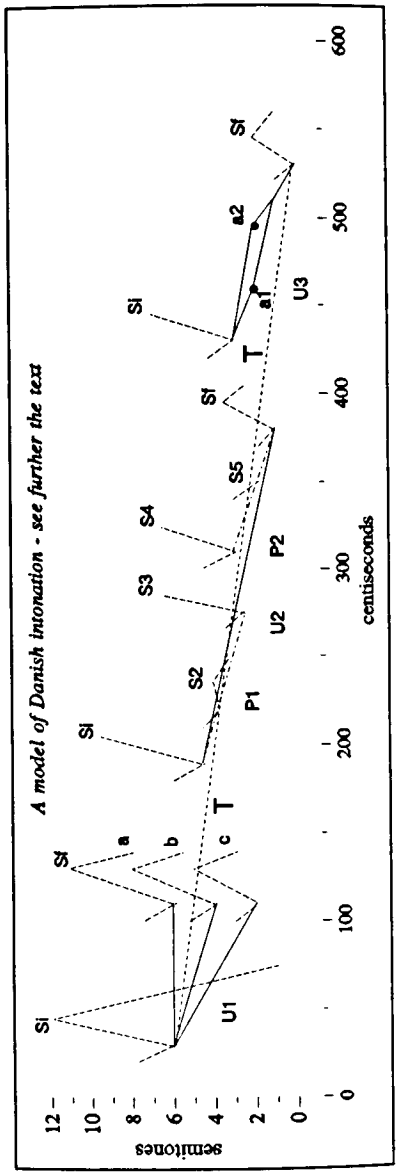
plications for illocutionary force and go counter to conventional usage. Thus, the most strongly marked intonation contour, the horizontal, will turn any sentence into an interrogative speech act.

The stress group pattern

F_0 is the most explicit among the acoustic cues to stress and prominence. The onset of any stressed vowel coincides with the onset of a recurrent melodic pattern which extends over all succeeding unstressed syllables within the same sentence, irrespective of their morphological or syntactic affiliation, until the onset of the next stressed vowel. This stress group pattern is subject to truncation or extension: A maximally developed pattern describes a brief initial fall succeeded by a steep rise to the first post-tonic and a steep fall through succeeding post-tonic syllables, cf. the initial stress group in *U1*. The shorter the stress group, the less extensive the F_0 pattern, so with a short stressed vowel and absence of post-tonic syllables, all that remains is the slight and brief initial fall, cf. *S5* in *U2*.

The implementation of F_0 patterns is extensively sensitive to their prosodic environment. (a) Rises are higher on marked contours, ceteris paribus, cp. *a*, *b* and *c* in *U1*. (b) Rises are successively lower from initial to final utterances, ceteris paribus, cf. the initial and final stress groups, respectively, in *U1-3*. (c) Rises lower progressively through an utterance, but any further differentiation according to prosodic phrasing - in the shape of higher post- than pre-boundary rises - is absent, cf. *S3* and *S4* in *U2*. (Either because it is too taxing for the production system or because a prosodic boundary per se is not intended.)

It is hard to say how much of this variation is directly speaker controlled, introduced by phonetic rule, and how much can be ascribed to general speech production principles which reduce articulatory explicitness through time or in unmarked vs. marked contexts. But whichever the output control mechanism, the variation is predictable, and though stress group patterns are an integral part of a model of speaker performance, they



are not part of the abstract underlying representation properly speaking. It is equally hard to determine the degree to which stress group pattern variation is perceptually relevant or redundant. In principle it can be entirely automatic and

yet perceptually relevant, whereas the reverse - rule governed but perceptually redundant - is perhaps less likely. I opt for perceptual relevance, because: The patterns described above pertain to utterances where the stressed syllables are equally prominent perceptually. When prominence varies, within the realm of normal stress, so does the magnitude of the rise: the higher the rise, ceteris paribus, the more salient the stressed syllable. Of course, this can only work on a background of *expected* neutral rise magnitudes in the various contexts.

Focus, on the other hand, is cued by suppression of the succeeding rise-fall, cf. *S2* in *U2*. Suppression of both preceding and succeeding rise-falls will create an emphasis for contrast (not displayed). -- In spontaneous speech focusing can also take another shape: the whole stress group is lifted out of (above or below) the contour, but the stress group patterns are not modified ([4]).

Note that syntactically or prosodically determined sentence accents, in the shape of a particularly prominent F_0 excursion finally (whatever the constituent), are absent in Standard Danish.

Intonation cues are global, not local

Standard Danish does not exhibit specific local tonal cues to either speech act function or boundaries (whatever the unit), in terms of final highs or lows. Intonational markedness and completion are inherent in the global course of utterance contours, possibly in conjunction with the derived variation in the magnitude of stress group pattern excursions. This property is shared by the majority of regional Danish variants.

Subordination and superposition; look-ahead and non-locality

Why this layered system of simultaneous, interacting, non-categorial intonational components of varying structural and temporal scope, where larger scope components carry and set the scale for smaller scope ones? And where the implementation of F_0 events is performed on the basis of upcoming as well as preceding events and is sensitive to syntactic and semantic structuring? Why not an abstract representation in terms of a linear sequence of categorially different, non-interacting pitch accents whose manifestation is exclusively locally determined, implemented on a left-to-right basis? And where there is no intonation component separate from the accents, and where global trends thus are the result of iterative application of local downstep rules ([1, 10, 13]) or a downstep morpheme ([2]) or locally determined range reduction ([9])? It would definitely be computationally simpler.

First of all 'pitch accent', with its connotation of phonologically distinct differences, is not an appropriate term for stress group patterns in Danish, because (a) there would be only one category and it would always align in the same fashion with the segmental material and (b) its phonetic manifestation is entirely predictable. In other words, a speaker cannot - at least not in the speech style analyzed so far - make a choice between various types of pitch accent. And when the magnitude of stress group pattern rises is manipulated to cue varying degrees of prominence, we are dealing with a scalar, not a binary, phenomenon. Speakers are not making a choice in those circumstances either of a particular pitch accent from a set of phonologically distinct ones, but are simply subordinating the manifestation of stress to the demands for signalling more or less prominence. -- I have seen Bornholm speakers invert their slowly falling-rising F_0 patterns across a whole utterance with a resulting change in perceived speech style or register, and I am certain a similar mechanism, a 'long component' or 'setting', is operative in Standard Danish. But that does not prove the existence of another pitch accent. It only shows that the manifestation of stress interacts with parameters at other (para-linguistic?) levels of description. Finally, if pitch accent were phonological stress would not be, a proposal with no serious merit. So, once more: stress group patterns are not part of the abstract representation of intonation. Therefore, with proper scaling the model will cover the majority of regional vari-

ants of Standard Danish as well, namely all the 'global' types, cf. above. The principal difference among them is in the shape and the magnitude of the stress group pattern and its alignment with segments. Whatever the patterns, they are superposed on a contour defined by the stressed syllables, without interfering with its course, but the intonation contour (its location in the text, its slope, its length), on the contrary, is decisive for the manifestation of stress group patterns. Furthermore, the manifestation of prominence degrees above normal stress demonstrates yet more syntagmatic interaction: stress group patterns are shrunk before emphasis for contrast, but not before a mere focus, i.e. the realization depends on the nature of the succeeding stress group. If all of this is not a paragon of superposition and subordination, I do not know what is.

Secondly, utterance and text intonation contours cannot be computed on a purely left-to-right basis either. As mentioned above, slope varies in inverse proportion to length, or - in other words - stressed syllables in an utterance are more closely spaced in frequency in longer utterances, or - in yet other words - the frequency location of a stressed syllable is sensitive to the number of succeeding ones. So, e.g., the second stressed syllable is higher in an utterance of four than one of three stressed syllables, ceteris paribus. A similar look-ahead and pre-planning is operative in the combination of utterances into prosodically coherent texts, cf. above. In this manner, utterance and text intonation are characterized by compression and expansion in time (contrary to stress group patterns which are truncated or extended), a process which is inconceivable without pre-planning. It is also attested in the difference between co-ordinate main clauses vs. a sequence of terminal declaratives, the former being less slanted relative to the global textual contour. -- Look-ahead and interaction turn up in the temporal structure as well: Speaking rate is somewhat accelerated before a (non-initial) focused item.

Thirdly, intonation contours, except the maximally marked one, are *always* associated with a more or less negative slope. Thus, e.g., there are no declaratives without a global downtrend.

These facts should make an account of sentence intonation in Danish in terms of (varying degrees of) local downstep or range reduction, triggered by certain pitch accent configurations, a formal possibility but empty of the significance it carries in, e.g., tone languages. -- Note that the superposition principle does not preclude features of a local kind, like Ladd's 'edge tones' [this symposium]. Thus, there are Danish regional varieties that definitely have a very local final high or low (as the case may be) which does not interact with what precedes it. The point is that edge tones are not universal, and when their corresponding function is carried by global trends, as in a number of Danish varieties, the linear representation is at a loss. Whereas their existence and incorporation is rather straightforward in a superposition model, precisely because they occur at domain boundaries. Likewise, a hierarchical organization does not, of course, preclude lexical tonal differences, cf. [6].

To sum up: the superpositional and sequential models do not differ in the acknowledgement of look-ahead, but they differ in its representation. For example, in the linear sequence model longer utterance onsets are higher than shorter ones but pitch accent relations are unaffected. In the superpositional approach utterance onsets need not vary, but the stressed syllables are less descending, the slope is less steep in longer than in shorter utterances. In [10 (p. 231)] the authors stated that a pitch accent can only look back to a previous pitch accent, a phrase to a previous phrase, etc., and apparent instances of anticipation should be explained by, e.g., feature spreading or temporal overlap in the realization of the segments in question [which really only amounts to passing the buck, NG]. The hierarchical concept entails a more direct interaction between subconstituents and superordinate structures.

Psychological reality?

A linear representation is computationally somewhat simpler and thus easier to implement in rule synthesis, than layered structures involving look-ahead, but that is not necessarily indicative of how speakers and listeners process intonation. And look-ahead and pre-planning are definitely operative and evidenced in other aspects of speech (syntax, slips of the tongue, 'phrasal stress reduction'), so why not in intonation?

Secondly, we are both speakers and hearers, and although we do not a fortiori produce and perceive in the same terms, it is at least not unlikely. I hypothesize that utterance intonation in Danish is holistically conceived, and that the concept is cognitively simpler, to the speaker and listener, than the summation of its atomic elements (the local ups and downs). Anecdotal evidence hints that this hypothesis would be worth testing: When linguistically naive Danes are asked to characterize the melody of various Danish utterances, they typically provide overall shapes. They have to be pushed hard - with exaggerated patterns - to hear that there are local pitch movements associated with stressed syllables. The local humps are not conceived as part of the melody and listeners seem to disregard the contribution to pitch that comes from stress.

REFERENCES

[1] Beckman, M. and J. Pierrehumbert (1986), "Intonational structure in English and Japanese," *Phonology Yearbook*, vol. 3, pp. 255-309.

[2] van den Berg, R., C. Gussenhoven and T. Rietveld (1991), "Downstep in Dutch: Implications for a model," in *Gesture, Segment, Prosody* (eds. G. Docherty and D.R. Ladd), Cambridge University Press, pp. 335-359.

[3] Bruce, G. (1977), *Swedish Word Accents in Sentence Perspective*, Lund: Gleerup.

[4] Dyhr, N.J. (1992), "An acoustical investigation of the fundamental frequency in Danish spontaneous speech," in *Nordic Prosody VI* (eds. B. Granström and L. Nord), Stockholm: Almqvist & Wiksell, pp. 23-32.

[5] Fujisaki, H., K. Hirose and K. Ohta (1979), "Acoustic features of the fundamental frequency contours of declarative sentences in Japanese," *Ann. Bull. Res. Inst. Logopedics and Phoniatrics*, vol. 13, pp. 163-173.

[6] Gårding, E. (1994), "On parameters and principles in intonation analysis," *Working Papers, Dept. of Linguistics, Lund University*, vol. 40, pp. 25-47.

[7] Grønnum, N. (1992), *The Groundworks of Danish Intonation*, Copenhagen: Museum Tusculanum Press.

[8] 't Hart, J. and R. Collier (1979), "On the interaction of accentuation and intonation in Dutch," *Proc. Ninth Int. Cong. Phonetic Sciences*, vol. II, pp. 395-402.

[9] Ladd, D.R. (1993), "In defense of a metrical theory of intonational downstep," in *The Phonology of Tone. The Representation of Tonal Register* (eds. H. van der Hulst and K. Snider), Berlin: Mouton de Gruyter, pp. 109-132.

[10] Liberman, M.Y. and J. Pierrehumbert (1984), "Intonational invariance under changes in pitch range and length," in *Language Sound Structure* (eds. M. Aronoff and R.T. Oehrle), Cambridge, Mass.: M.I.T. Press, pp. 157-233.

[11] Möbius, B. (1993), *Ein quantitatives Modell der deutschen Intonation*, Tübingen: Max Niemeyer Verlag.

[12] Öhman, S.E.G. (1968), "A model of word and sentence intonation," *STL-QPSR, Royal Institute of Technology, Stockholm*, pp. 6-11.

[13] Pierrehumbert, J.B. (1980), *The Phonology of English Intonation*, M.I.T. Doctoral Dissertation.

[14] Reinholt Petersen, N. and P. Molbæk Hansen (1994), "Fundamental frequency resettings, pauses, and syntactic boundaries in read-aloud Danish prose," *Acta Linguistica Hafniensia*, vol. 27, 2, pp. 383-401.

[15] Vaissière, J. (1983), "Language-independent prosodic features," in *Prosody: Models and Measurements* (eds.: A. Cutler and D.R. Ladd), Berlin: Springer-Verlag, pp. 53-66.

MECHANISMS OF DEVELOPMENTAL CHANGE IN SPEECH AND LANGUAGE

Patricia K. Kuhl
University of Washington, Seattle, WA

ABSTRACT

One of the most interesting aspects of language development is the transition that occurs in phonetic perception during the first year of life. At birth, infants are capable of distinguishing all phonetic contrasts in the world's languages. By adulthood, these abilities are severely restricted. This "language-general" to "language-specific" pattern of change is mirrored in speech production. This paper focuses on the role of early language experience on infants' perceptual and perceptual-motor systems in bringing this transition about. The data show that by the time infants begin to master the higher levels of language, such as sound-meaning correspondences, contrastive phonology, and grammatical rules, their perceptual and perceptual-motor systems are already tuned to a specific language. These results are described in a developmental theory at the phonetic level that holds promise for higher levels of language.

INTRODUCTION

Research on developmental speech perception and speech production has revealed an interesting pattern of change. Speech exhibits a language-general pattern that becomes language-specific by the end of the first year of life [1,2,3]. What accounts for the transition?

Early in life infants discern differences between all the phonetic units used in the world's languages, and demonstrate exquisite sensitivity to acoustic change in the region of the boundaries between phonetic categories [4]. By 12 months of age, infants fail to discriminate foreign contrasts they once discriminated [3]. As adults, our abilities are greatly reduced; we often find it difficult to perceive differences between sounds not used to distinguish words in our native language [3]. Adult native speakers of Japanese have difficulty discriminating American English /r/ and /l/ [5], though Japanese infants make the distinction [6].

Speech *production* follows a similar pattern. Regardless of culture, all infants progress through a set of universal stages during the first year [7]. By the end of the first year, however, the utterances of infants reared in different countries begin to diverge, reflecting the ambient language [8,9]. In adulthood, the speech motor patterns that contribute to one's "accent" are very difficult to alter [9].

Work in my laboratory focuses on the role of early linguistic experience in bringing about this language-general to language-specific change in speech perception and production. The thesis is that linguistic experience results in an interesting kind of learning. Given linguistic input, the perceptual and perceptual-motor systems underlying speech show self-organization accompanied by a loss in flexibility. The vehicle for change is argued to be *representations* stored in memory that capture the regularities of a specific language. These representations alter the perceptual and perceptual-motor skills of infants. The findings demonstrate that in the absence of formal language understanding or use, infants' perceptual and perceptual-motor systems are strongly biased towards the characteristics of the ambient language. A model is described at the phonetic level that shows how this structure accounts for the transition and aids the acquisition of phonology.

Language Experience Affects Speech Perception Early in Life

Research in my laboratory has uncovered an effect that helps explain how language experience affects speech perception and production. The effect shows that linguistic experience alters the perceived distances between speech stimuli. In effect, our results suggest that linguistic experience "warps" the perceptual space underlying speech. The result is that perceptual categories are formed, ones that begin to mirror the phonological categories of the ambient language. The experimental data that support these claims derive from a phenomenon I have called the *perceptual magnet effect*. It shows that phonetic "prototypes" (the best or most representative instances of phonetic categories) play a unique role in speech perception. They function like "perceptual magnets" for other sounds in the category [10] (Figure 1). When listeners hear a phonetic prototype and attempt to discriminate it from sounds that surround it in acoustic space (1A), the prototype displays an attractor effect on the surrounding sounds. It perceptually pulls other members of the category toward it, making it difficult to hear differences between the prototype and surrounding stimuli (1B). Poor instances from the category (nonprototypes) do not function in this way. A variety of experimental tasks produce this result [11,12]. Other studies confirm listeners' skills in identifying phonetic prototypes and show that they are language specific [13,14].

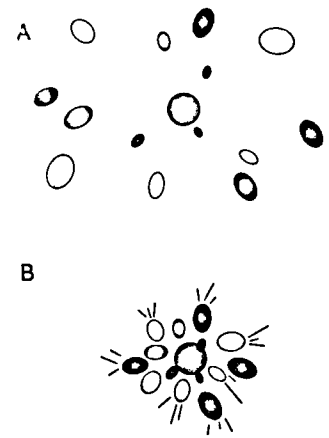


Figure 1. The perceptual magnet effect: When a variety of sounds in a category surround the category prototype (A), they are perceptually drawn toward the prototype (B). The prototype appears to function like a magnet for other stimuli in the category.

Developmental studies show that the perceptual magnet effect is exhibited by 6-month-old infants for the sounds of their native language [10]. Moreover, cross-language experiments show that the magnet effect is the product of linguistic experience [15]. The cross-language experiment was conducted with infants in America and Sweden. The infants from both countries were tested with two vowel prototypes, an American English vowel prototype, /i/ (as in "peep"), and a Swedish vowel prototype, /y/ (as in "fye"). Adults from both cultures perceived the foreign vowel as a nonprototype. The results demonstrated that the perceptual magnet effect was affected by exposure to a particular language as early as 6 months after birth. By the age of six months, American infants demonstrated the magnet effect only for the American English /i/; they treated the Swedish /y/ like a nonprototype. Swedish infants of the same age showed the opposite pattern, demonstrating the magnet effect for the Swedish /y/ and treating the American English /i/ as a nonprototype.

Recent work by Polka and Werker [16] both support and extend these findings. They tested Canadian English infants in a discrimination task involving two German phonetic contrasts. Their results confirm the presence of the magnet effect in 6-month-old infants and show a decline in the discrimination of foreign-language vowel contrasts between 4 months and 6 months, earlier for vowels than for consonants but following the same pattern.

Studies conducted on adult listeners by Iverson and Kuhl have begun to suggest the mechanism underlying the magnet effect. These studies employ multidimensional scaling (MDS) techniques to examine how the magnet effect distorts perception [17,18] (Figure 2). The studies show that the best instances of phonetic categories yield increased perceptual clustering while the category's worst instances yield reduced perceptual clustering. For example, Iverson and Kuhl [17] computer synthesized a set of syllables beginning with /r/ and /l/. The syllables were created by varying critical acoustic components of the signals, the second (F2) and third (F3) formant frequencies. The syllables were spaced at equal intervals in a 2-dimensional grid (2A). Listeners identified each syllable as beginning with either /r/ or /l/, rated its category goodness, and estimated the perceived similarity for all possible pairs

of stimuli using a scale from "1" (very dissimilar) to "7" (very similar). Similarity ratings were scaled using MDS techniques.

The results revealed that *perceived* distances differed from *real physical* distances. The physical (acoustic) differences between pairs of stimuli were equal (2A); however, perceived distance was not equal, it was "warped" (2B). The perceptual space around the best /r/ and the best /l/ was greatly reduced, while the space near the boundary between the two categories was expanded (see [19] for similar findings on cognitive categories outside the domain of speech).

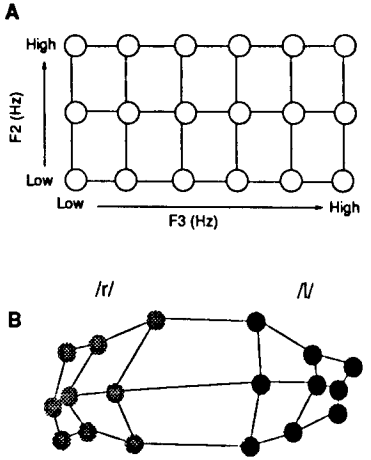


Figure 2. Consonant tokens of /r/ (gray dots) and /l/ (black dots) were generated to be equally distant from one another in acoustic space (A). However, when listeners perceive them, distance is distorted (B). Perceptual space is reduced near the best instances of /r/ and /l/ and expanded at the boundary between the two. The resulting "perceptual map" (B) differs for speakers of different languages.

The results suggest that linguistic experience results in the formation of *perceptual maps* specifying the perceived distances between stimuli. These maps increase internal category cohesion while maximizing the distinction between categories. The critical point for theory is the hypothesis that the map is defined differently for speakers of different languages [18]. Native speakers of Japanese tested with the same American /r/ and /l/ stimuli show a different perceptual map, one without perceptual clusters around the American /r/ and /l/ prototypes.

NATIVE LANGUAGE MAGNET (NLM) THEORY

I have proposed a 3-step theory of speech development, called the Native Language Magnet (NLM) theory [1,2]. NLM describes infants' initial state as well as changes brought about by experience with language. It explains how infants' developing native-language speech representations alter both speech perception and production. The schematic illustration provided here presents the hypothetical case for vowels, but the same principles and theory applies to consonant perception.

Phase 1 describes the initial state of infants' speech perception abilities (Fig. 3). At birth, infants partition the sound stream into gross categories separated by natural auditory-perceptual boundaries. The lines in Figure 3 show these hypothesized perceptual boundaries. The lines separate the vowels of all languages. According to NLM, these perceptual boundaries are innately specified in auditory processing and do not depend on specific language experience. The boundaries initially structure perception in a phonetically-relevant way, which is extremely helpful for infants. However, they are not due to a "language module" or other language-specific device but to more basic auditory perceptual processing mechanisms. This is argued to be the case because experiments on animals show that they exhibit boundary effects in the same places in acoustic space [20].

The data on human infants supporting this initial stage in the model stems from two kinds of studies: first those showing that early in life infants discriminate natural native- and foreign-language vowel and consonant contrasts, and second, those on "categorical perception" showing that infants exhibit increased sensitivity to change in the region of phonetic boundaries for both consonants and vowels [21]. Both show that at birth infants are capable of distinguishing among the consonants and vowels of the world's languages.

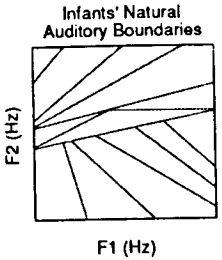


Figure 3. At birth, infants perceptually partition the acoustic space underlying phonetic distinctions in a language-universal way. These "boundaries" allow them to discriminate all phonetically relevant differences in language.

Phase 2 in the model illustrates hypothetical vowel spaces that exist at 6 months of age for infants reared in three very different language environments, Swedish, English, and Japanese (Fig. 4). According to the model, infants at this age show more than the innate boundaries exhibited in Phase 1. Our data indicate that by 6 months, infants have heard hundreds of thousands of instances of particular vowels. According to NLM, infants represent this information in memory in some form. This is illustrated in Figure 4. These diagrams show infants' stored representations, which reflect the distributional characteristics of the vowels infants have heard. Infants being raised in Sweden, America, and Japan hear different vowels. Thus, their stored representations differ. In each case, linguistic experience has produced stored representations that reflect the vowel system of the ambient language. According to NLM, in Phase 2 language-specific magnet effects are exhibited by infants.

Phase 3 shows how magnet effects recursively alter the initial state of speech perception. Magnet effects cause certain perceptual distinctions to be minimized (those near the magnet attractors) while others are maximized (those near the boundaries between two magnets). The consequence is that some of the boundaries that initially divided the space perceptually "disappear" as the space is reconfigured to incorporate a language's particular magnet placement. This is schematically illustrated in Figure 5 in which certain boundaries that existed in Phase 1 have been erased. It is important to note, however, that even though these boundaries have been erased, the model does not hold that sensory perception has changed. Instead, it is argued that higher order memory and representational systems have altered infants' abilities. In other words, magnet effects *functionally* erase certain boundaries — those relevant to foreign but not native languages. By Phase 3, a perceptual space once characterized by simple boundaries has been replaced by a warped space dominated by magnets.

The important point for theory is that infants at 6 months of age have no awareness of phonemes or the fact that sound units are used contrastively in language to name things. Yet the infant's perceptual system has organized itself to reflect language-specific phonetic categories. At the next stage in linguistic development, when infants acquire word meanings by relating sounds to objects

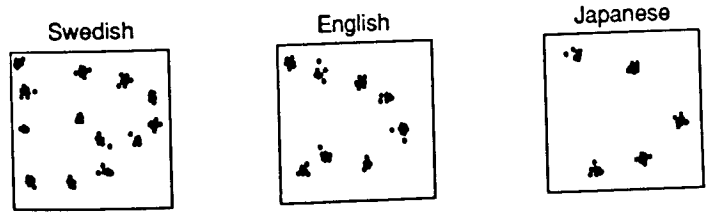


Figure 4. By 6 months of age, infants reared in different linguistic environments show an effect of language experience. Infants store incoming vowel information in memory in some form. The resulting representations are language-specific, and reflect the distributional properties of vowels in the three different languages.

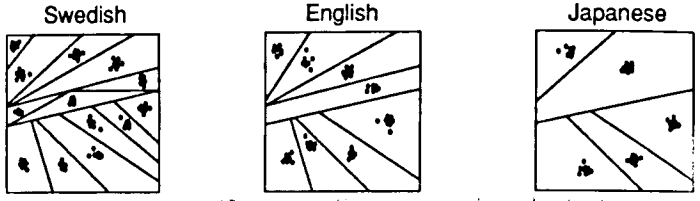


Figure 5. After language-specific magnet effects appear (shown by the dots), some of the natural boundaries that existed at birth 'disappear.' Infants now fail to discriminate foreign-language contrasts they once discriminated.

and events in the world, the language-specific mapping that has already occurred in their perceptual systems will greatly assist this process.

NLM theory offers a potential explanation for the reorganization in speech perception observed by Werker [3]. A developing magnet pulls sounds that were once discriminable towards it, making them less discriminable. Magnet effects should therefore developmentally precede changes in infants' perception of foreign-language contrasts; preliminary data indicate that they do [16]. The magnet effect also helps account for the results of studies on the perception of sounds from a foreign language by adults [3,5,9]. For example, NLM theory explains Japanese listeners' difficulty with American /r/ and /l/ by predicting that the magnet effect for their category prototype (which is neither American /r/ nor /l/) will attract both /r/ and /l/, making the two sounds difficult for native-speaking Japanese people to discriminate. Best [5] has made predictions about the relative discriminability of foreign-language contrasts by examining the relationship of specific foreign sounds to native-language categories; these predictions are consistent with NLM.

Effects of Linguistic Experience on Speech Production

As adults, we produce speech motor patterns that are difficult to alter. When do infants acquire these life-long speech-motor patterns? When do they forge the initial link between the perception and production of speech? Data can be adduced by the earliest age at which ambient language affects spontaneous speech production. It is known that by 1 year of age language-specific patterns of speech production appear in infants' spontaneous utterances [8,9]; by 30 months, detailed patterns that differentiate sounds in two different languages are observed [22].

Recent studies, however, suggest that the initial perceptual-motor link is in place much earlier. In another paper in this volume, Kuhl and Meltzoff describe research showing that infants can imitate the gross spectral forms of vowels at 12-, 16-, and 20-weeks of age. This indicates the presence of an auditory-articulatory link very early in life. In the Kuhl and Meltzoff studies [23], infants watched a video of a woman articulating either /i/, /a/, or /u/ for 5 minutes on each of three consecutive days. Infants' utterances were analyzed both perceptually (phonetic transcription) and instrumentally (spectrographic analysis). Two findings emerged. There was developmental change in infants' vowel productions between 12- and 20-weeks of age. The areas of vowel space occupied by infants' /a/, /i/, and /u/ vowels become progressively more tightly clustered at each age. Second, the data suggest that infants attempted to imitate the vowels they heard. The total amount of exposure was only 15 minutes; yet this appeared to be sufficient to alter infants' productions. If 15 minutes of laboratory exposure to a vowel is sufficient to influence infants' vocalizations, then listening to ambient language for weeks would be expected to provide a powerful influence on infants' production of speech. Kuhl and Meltzoff interpret the data as suggesting that infants' stored representations of speech alter not only infant perception, but speech production as well. Infants' representations serve as targets that guide motor production. Stored representations are thus viewed as the common cause for both the tighter clustering observed in infant vowel production and the tighter clustering observed in infant vowel perception. Additional data on infants' and adults' auditory-visual perception of speech, also support this conclusion (see [23] for further discussion)

This pattern of learning and self-organization, in which perceptual patterns stored in memory serve as guides for production, is strikingly similar to that seen in other domains involving auditory-perceptual learning, such as birdsong [24], visual-motor learning, such as gestural imitation of articulatory movements in the absence of sound [25], and imitation from memory [26]. In each of these cases, perceptual experience establishes a representation that guides sensory-motor learning. In the case of infants and speech, perception affects production in the earliest stages of language learning, reinforcing the idea that the perceptual-motor link is in place early in life [2,23,27,28,29].

Learning Prosodic Regularities

Infants' abilities to learn do not begin the day they are born. Learning commences prenatally with the more global, prosodic aspects of language. By the time infants are born, exposure to sound *in utero* has resulted in a preference for native-language over foreign-language utterances [30,31]. The mother's voice [32] and simple stories she read during the last trimester [33] are also recognized by infants at birth. Studies on the acoustics of speech and the intrauterine environment suggest that intense (above 80dB), low-frequency sounds (particularly below 300 Hz, but as high as 1000 Hz with some attenuation) penetrate the womb [34]. This means that the prosodic patterns of speech, including voice pitch and the stress and intonation characteristics of a particular language and speaker, are transmitted to the fetus, while the sound patterns that allow phonetic units and words to be identified are greatly attenuated. (This can be compared to listening to speech through the wall of a room — a human voice can be identified, but words cannot be made out.)

Postnatally, infants' learning of the prosodic aspects of speech provides additional information about language-specific sound patterns. Jusczyk and his colleagues have focused on infant learning of the sound patterns typical of native-language words, phrases, and sentences [35]. This work shows that between 6 and 9 months of age, infants develop listening preferences for sound patterns typical of the native language. In one study [36], both American and Dutch infants listened significantly longer to native- as opposed to foreign-language words. At 6 months of age, infants showed no listening preferences. Other work shows listening preferences at 9 months, but not at 6 months, for words that follow the predominant stress pattern of the language [37]. These studies indicate that prior to the time that infants learn the meanings of individual words or phrases, they recognize general perceptual characteristics that describe such units in their native language.

CONCLUSIONS

In the first year of life infants learn much about the perceptual characteristics of their native language. Perceptual learning subsequently alters the perception and production of speech. According to the *Native Language Magnet* theory, perceptual learning early in life results in the formation of stored representations that capture native-language regularities. These stored representations act like *perceptual magnets* for similar patterns of sound. Magnet effects distort physical distance, creating perceptual maps in which distance has been altered. Perceptual maps shrink distances near a category's most typical instances and stretch distances between categories. Perceptual maps differ in adults who speak different languages. The magnet effects and the perceptual maps they produce also affect speech production. This helps explain why, as adults, we do not hear or produce foreign-language sounds very well. During the language-learning period, our perceptual maps are tuned to our native language. The model ascribes an important role for language input. Future work will be aimed at testing and refining the model.

ACKNOWLEDGMENT

Research reported here was supported by grants from NIH (HD-22514, HD-18286, and DC 00520). Correspondence

should be sent to Patricia K. Kuhl, Department of Speech and Hearing Sciences, University of Washington, Seattle, WA 98195

REFERENCES

[1] Kuhl, P. K. (1993), "Innate predispositions and the effects of experience in speech perception: The native language magnet theory", in *Developmental neurocognition: Speech and face processing in the first year of life*, edited by de Boysson-Bardies, B., de Schonen, S., Jusczyk, P., McNeilage, P., Morton, J., pp. 259-274. Dordrecht, Netherlands: Kluwer.

[2] Kuhl, P. K., Meltzoff, A. N. (1995), "Evolution, nativism, and learning in the development of language and speech", in *The biological basis of language*, edited by Gopnik, M. New York: Oxford University Press.

[3] Werker, J. F., Polka, L. (1993), "The ontogeny and developmental significance of language-specific phonetic perception", in *Developmental neurocognition: Speech and face processing in the first year of life*, edited by de Boysson-Bardies, B., de Schonen, S., Jusczyk, P., McNeilage, P., Morton, J., pp. 275-288. Dordrecht, Netherlands: Kluwer.

[4] Eimas, P. D., Miller, J. L., Jusczyk, P. W. (1987), "On infant speech perception and the acquisition of language", in *Categorical perception: The groundwork of cognition*, edited by Harnad, S., pp. 161-195. New York: Cambridge University Press.

[5] Best, C. T. (1993), "Emergence of language-specific constraints in perception of non-native speech: A window on early phonological development", in *Developmental neurocognition: Speech and face processing in the first year of life*, edited by de Boysson-Bardies, B., de Schonen, S., Jusczyk, P., McNeilage, P., Morton, J., pp. 289-304. Dordrecht, Netherlands: Kluwer.

[6] Tsushima, T., Takizawa, O., Sasaki, M., Shiraki, S., Nishi, K., Kohno, M. et al. (1994), "Discrimination of English /r-l/ and /w-y/ by Japanese infants at 6-12 months: Language-specific developmental changes in speech perception abilities". *Proceedings of the International Conference on Spoken Language Processing*. Tokyo: Acoustical Society of Japan.

[7] Ferguson, C. A., Menn, L., Stoel-Gammon, C. (1992), *Phonological development: Models, research, implications*, Timonium, MD: York.

[8] de Boysson-Bardies, B. (1993), "Ontogeny of language-specific syllabic productions", in *Developmental neurocognition: Speech and face processing in the first year of life*, edited by de Boysson-Bardies, B., de Schonen, S., Jusczyk, P., McNeilage, P., Morton, J., pp. 353-363. Dordrecht, Netherlands: Kluwer.

[9] Flege, J. E. (1993), "Production and perception of a novel, second-language phonetic contrast", *Journal of the Acoustical Society of America*, vol. 93, pp. 1589-1608.

[10] Kuhl, P. K. (1991), "Human adults and human infants show a 'perceptual magnet effect' for the prototypes of speech categories, monkeys do not", *Perception & Psychophysics*, vol. 50, pp. 93-107.

[11] Iverson, P., Kuhl, P. (1995), "Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling", *Journal of the Acoustical Society of America*, vol. 97, pp. 553-562.

[12] Sussman, J. E., Lauckner-Morano, V. J. (1995), "Further tests of the 'perceptual magnet effect' in the perception of [i]: Identification and change/no-change discrimination", *Journal of the Acoustical Society of America*, vol. 97, pp. 539-552.

[13] Kuhl, P. K. (1992), "Psychoacoustics and speech perception: Internal standards, perceptual anchors, and prototypes", in *Developmental psychoacoustics*, edited by Werner, L. A., Rubel, E. W., pp. 293-332. Washington, DC: American Psychological Association.

[14] Miller, J. L. (1994), "On the internal structure of phonetic categories: A progress report", *Cognition*, vol. 50, pp. 271-285.

[15] Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., Lindblom, B. (1992), "Linguistic experience alters phonetic perception in infants by 6 months of age", *Science*, vol. 255, pp. 606-608.

[16] Polka, L., Werker, J. F. (1994), "Developmental changes in perception of nonnative vowel contrasts", *Journal of Experimental Psychology: Human Perception and Performance*, vol. 20, pp. 421-435.

[17] Iverson, P., Kuhl, P. K. (1994), "Tests of the perceptual magnet effect for American English /r/ and /l/", *Journal of the Acoustical Society of America*, vol. 95, p. 2976.

[18] Kuhl, P., Iverson, P. (1995), "Linguistic experience and the 'perceptual magnet effect'", in *Speech perception and linguistic experience: Theoretical and methodological issues in cross-language speech results*, edited by Strange, W. Timonium, MD: York Press.

[19] Nosofsky, R. M. (1988), "Exemplar-based accounts of relations between classification, recognition, and typicality", *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 14, pp. 700-708.

[20] Kuhl, P. K. (1991), "Perception, cognition, and the ontogenetic and phylogenetic emergence of human speech", in *Plasticity of Development*, edited by Brauth, S. E., Hall, W. S., Dooling, R. J., pp. 73-106. Cambridge, MA: MIT Press.

[21] Kuhl, P. K. (1987), "Perception of speech and sound in early infancy", in *Handbook of infant perception: Vol 2. From perception to cognition*, edited by Salapatek, P., Cohen, L., pp. 275-382. New York: Academic Press.

[22] Stoel-Gammon, C., Williams, K., Buder, E. (1994), "Cross-language difference in phonological acquisition: Swedish and American /t/", *Phonetica*, vol. 51, pp. 146-158.

[23] Kuhl, P. K., Meltzoff, A. N. (this volume), "Vocal learning in infants: Development of perceptual-motor links for speech".

[24] Konishi, M. (1989), "Birdsong for neurobiologists", *Neuron*, vol. 3, pp. 541-549.

[25] Meltzoff, A. N., Moore, M. K. (1994), "Imitation, memory, and the representation of persons", *Infant Behavior and Development*, vol. 17, pp. 83-99.

[26] Meltzoff, A. N. (1988), "Infant imitation and memory: Nine-month-olds in immediate and deferred tests", *Child Development*, vol. 59, pp. 217-225.

[27] Suomi, K. (1993), "An outline of a developmental model of adult phonological organization and behaviour", *Journal of Phonetics*, vol. 21, pp. 29-60.

[28] Vihman, M. M. (1993), "Variable paths to early word production", *Journal of Phonetics*, vol. 21, pp. 61-82.

[29] Studdert-Kennedy, M. (1993), "Discovering phonetic function", *Journal of Phonetics*, vol. 21, pp. 147-155.

[30] Cutler, A., Mehler, J. (1993), "The periodicity bias", *Journal of Phonetics*, vol. 21, pp. 103-108.

[31] Moon, C., Cooper, R. P., Fifer, W. P. (1993), "Two-day-olds prefer their native language", *Infant Behavior and Development*, vol. 16, pp. 495-500.

[32] DeCasper, A. J., Fifer, W. P. (1980), "Of human bonding: Newborns prefer their mothers voices", *Science*, vol. 208, pp. 1174-1176.

[33] DeCasper, A. J., Spence, M. J. (1986), "Prenatal maternal speech influences newborns' perception of speech sounds", *Infant Behavior & Development*, vol. 9, pp. 133-150.

[34] Lecanuet, J. P., Granier-Deferre, C. (1993), "Speech stimuli in the fetal environment", in *Developmental neurocognition: Speech and face processing in the first year of life*, edited by de Boysson-Bardies, B., de Schonen, S., Jusczyk, P., McNeilage, P., Morton, J., pp. 237-248. Dordrecht, Netherlands: Kluwer.

[35] Jusczyk, P. W. (1993), "From general to language-specific capacities: the WRAPSA model of how speech perception develops", *Journal of Phonetics*, vol. 21, pp. 3-28.

[36] Jusczyk, P. W., Friederici, A. D., Wessels, J. M. I., Svenkerud, V. Y., Jusczyk, A. M. (1993), "Infants' sensitivity to the sound patterns of native language words", *Journal of Memory and Language*, vol. 32, pp. 402-420.

[37] Jusczyk, P. W., Cutler, A., Redanz, N. J. (1993), "Infants' preference for the predominant stress patterns of English words", *Child Development*, vol. 64, pp. 675-687.

THE PERCEPTUAL-MAGNET EFFECT: AN EMERGENT CONSEQUENCE OF EXEMPLAR-BASED PHONETIC MEMORY

Francisco Lacerda
Institute of Linguistics, Stockholm University, S-106 91 Stockholm, Sweden

ABSTRACT

This paper uses a mathematical model of infant speech perception to examine the assumptions and consequences of Kuhl's Native Language Magnet theory (NLM). A basic assumption of the NLM theory is that perceptual space is partitioned into phonetically relevant categories that are represented by category prototypes — the category's "best exemplar". The category prototypes function as "perceptual magnets" that attract exemplars falling within their zone of influence. As a consequence, discrimination as a function of the auditory distance between a prototype and other exemplars is low in the neighborhood of the prototype and has an increasing positive derivative as the exemplar moves away from the prototype (discrimination is proportional to the square or the cube of the auditory distance between the prototype and the exemplar).

While Kuhl's description of the perceptual magnet effect in terms of prototypes is elegant, I argue that the magnet effect emerges from a simple similarity metric operating on collections of exemplars stored in memory, without the need to refer to special exemplars.

INTRODUCTION

With this paper I would like to open a debate on conceptual aspects of Kuhl's Native Language Magnet theory (NLM) that has recently become very influential in the description of categorization and discrimination phenomena.

According to Kuhl [8], the phonetic perceptual space is organized in terms of particular exemplars, prototypes, that function as references for different classes of speech sounds. Prototypes are regarded as particular good vowel-category representatives — "focal" exemplars [15;1] — against which other instances of vowel sounds are compared in the course of the perception process. Kuhl's important addition to a traditional prototype-based classification process is the assumption that each prototype has its specific neighborhood of influence. The prototype is pictured as a "perceptual-magnet" that exerts its attraction force on neighboring auditory representations [7]. Stimuli producing auditory representations in that neighborhood are attracted to the prototype. In contrast, non-prototypical sounds are not supposed to exhibit the magnetic effect. In other words, discrimination, as a function of the psychoacoustic distance, is low and increases slowly within a prototype's neighborhood. Thus, variants producing auditory representations in that region tend to be perceived as more similar to the prototype than what might be expected on the basis of the auditory distance per se. As a consequence, the perceptual space appears warped in the neighborhood of a prototype whereas in the neighborhood of a non-prototype discrimination increases proportionally to the psychoacoustic distance.

Looking at categorization processes in terms of focal prototypes clearly captures the functional aspects of classification phenomena. Kuhl's introduction of the perceptual-magnet notion extends the traditional view of prototype-based classification processes by accounting for non-linear effects. Nevertheless a prototype-based approach raises issues that are rooted in the very notion of prototype. One issue, for instance,

concerns its application to language acquisition processes. To account for effects of learning, prototypes must be re-tuned as a result of the learning procedure. In this case the magnet must be re-located during the language acquisition process at the same time that it functions as a prototypical reference. Another issue concerns the general motivation of prototypes in describing perceptual phenomena. Is the concept of a prototype really needed to account for categorization processes or can a prototype-like behavior emerge from collections of exemplars?

In this paper I will focus on the latter issue and present a sketch of how the magnet-effect can be derived from exemplars stored in memory and a simple distance metric.

PROTOTYPES

One problem I see with the prototype approach is that the very nature of the language acquisition process requires prototypes that can be rearranged in the perceptual space. If prototypes are assumed to be determined by the acoustic properties of the speech sounds — like the case of "point vowels" — it is necessary to assume a loss mechanism to discards non-functional prototypes. Following Kuhl's magnetic analogy (surely in more literal sense than she claims) the prototype must have larger mass than the variants in order to be stable enough in the perceptual space. Otherwise the system cannot be described in terms of a relatively stable magnet towards which "lighter" magnets are attracted. As the disparity of masses involved diminishes, the system becomes a two-mass system in which both elements clearly interact with each other. Thus, in order to achieve stability of the perceptual-magnets it seems necessary to postulate forces that anchor the prototypes at the appropriate locations or a process by which a prototype-like structure emerges in the perceptual space. Possible acoustic-articulatory arguments for a priori locations of vowel prototypes may be found in Stevens' [16] quantal theory of speech — but why should perceptual prototypes necessarily match acoustic-articulatory constraints? An alternative perceptual account for specific vowel-prototype locations might be based on Lindblom's [14] modeling of vowel systems. However, this account would involve circularity since Lindblom starts out with a given number of vowels (matching a priori Kuhl's prototypes) and tries to determine their positions under the constraints of both maximal perceptual distance among the vowels in the system and articulatory feasibility. But prototypes cannot, at any rate, be determined solely by constraints in the perceptual and articulatory system.

To account for the language acquisition process, prototypes will have to be moveable entities. Actually, as revealed by Kuhl and Meltzoff's [9] recent research, prototypes are highly plastic entities since 3, 4 and 5 months-old infants rearranged their prototype locations after only 3x5 minutes audio-visual exposure to model presentations. At first glance, this extreme plasticity by 5 months of age is hard to reconcile with the establishment of stable language-dependent prototypes by 6 months of age [12]. Given the normal signal variability, establishing stable prototypes within one month's period should, in principle, be a difficult task for the infant. Yet, taking into consideration that from the infant's point of view there may be only a limited number of functionally relevant audio-visual combinations, maybe the task is afterall less demanding than it first appears. At any rate, to account for this reorganization during the language acquisition process, prototypes must be seen as plastic entities, suitable to modification by adequate exposure to language but this diminishes the

referential role of prototypes. If prototypes are the distal effect of external stimulation, there is no obvious conceptual reason to use prototypes instead of the very exemplars on which they are based.

EXEMPLAR-BASED MODEL

To present my argument that prototypes are implicit in exemplar-based categorization processes, I will introduce a very simple perception model in which the perceptual magnet effect emerges from an exemplar-based categorization process.

Model assumptions

Let us simplify the exemplar-based perceptual model by addressing classification and discrimination of one-dimensional elements. This one-dimensional case represents a situation in which there is a determinant main dimension that allows discrimination of the stimuli. An example would be discriminating vowels in terms of degree of opening. Although vowels can be represented by multi-dimensional points in a formant space, degree of opening can be satisfactorily discriminated by considering F_1 alone. The general case in which co-variation in several dimensions must be considered can, in principle, be treated as a combination of appropriate one-dimensional cases.

My basic assumption is that representation of exemplars are stored in memory and that an external labeling function is also available during the learning phase. The plausibility of a memory representation of specific exemplars is supported by Jusczyk's recent results indicating that infants store specific information about voices and words [3;4;5]. In addition, access to a labeling function is typical of the learning situation. During the language acquisition process, the infant is exposed to allophonic variation in its ambient language and to correlated category information that is available from other modalities. Thus, stimulus variability along a perceptual dimension for a given category is assumed to produce distributed memory representation of exemplars. This distribution represents the frequencies with which particular values of that dimension were observed for the category. In the following example I will assume that stimuli generate normal distributions with mean μ and standard deviation σ, as indicated by the function Class(x, μ, σ):

$$\text{Class}(x,\mu,\sigma) \quad \frac{1}{\sqrt{2\cdot\pi}\cdot\sigma}\cdot e^{\frac{(x-\mu)^2}{2\cdot\sigma^2}}$$

The function describes the relative frequency with which a **x**-values were stored in memory.

The model categorizes new stimuli using a similarity metric that is based on the labels of the memorized exemplars that are found within the immediate neighborhood of the new stimulus. The new stimulus is given the label of the category contributing with the largest number of exemplars in the stimulus' neighborhood, provided that category's dominance is above a pre-established minimum decision threshold[1]. If similarity is below the threshold, no decision is made and the new stimulus is classified as "unknown". In this paper I will assume that the decision threshold is 0. This metric behaves like a cohort model for lexical access. The number of neighbors belonging to category **A** (μ=0) found in the neighborhood ε of the stimulus is given by

$$\text{NeighbA}(x0,\varepsilon) \quad \text{TotalA}\int_{x0-\varepsilon}^{x0+\varepsilon} \text{Class}(x,0,\sigma)\,dx$$

[1] If there are many competing categories, the relative similarity to any of them may be so low that no decision should be made by the model.

In this expression **x0** represents the actual value of the stimulus along the relevant dimension and **TotalA** the total number of exemplars in the category.

If the alternative category is **B**, with μ=3, then the similarity of stimulus **x0** to category **A** and to category **B** is defined as

$$sA(x0) := \frac{\text{NeighbA}(x0,\varepsilon)}{\text{NeighbA}(x0,\varepsilon) + \text{NeighbB}(x0,\varepsilon)}$$

The similarity of elements **x0** to each of the categories **A** and **B** is depicted in figure 1 for the case in which categories **A** and **B** are described in table 1.

Table 1. Specifications of two categories, A and B, with normally distributed exemplars (mean μ and standard deviation σ). The categories have different numbers of exemplars.

Category	μ	σ	Exemplars in the category
A	0	1	1000
B	3	1	100

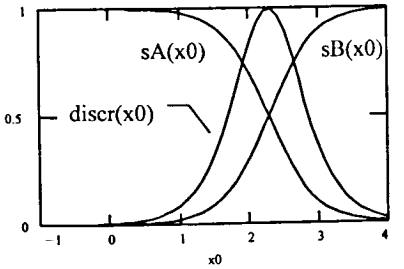


Figure 1. Similarity functions to classes A and B and discrimination function. Note that the category boundary is shifted towards category B because A contains a larger number of exemplars.

If similarity to a category is 1 all the neighbors come from that category. The figure also displays a discrimination measure, **discr(x0)**, that is based on the local variation in the number of exemplars coming from different categories. For the two categories case, the measure is defined as

$$\text{discr}(x0) := \frac{\left|\frac{d}{dx0}sA(x0)\right| + \left|\frac{d}{dx0}sB(x0)\right|}{2\cdot\text{Const}}$$

where the constant, **Const**, is arbitrary and represents here the maximum slope of the similarity functions and is used to make the discrimination function to fit the interval [0;1].

An interesting property of this exemplar model is that it can simulate Kuhl's perceptual magnet effect without reporting to a specific prototype. In fact, assuming that a prototype, in Kuhl's terms, is the center of the category distribution, the discrimination curve, **discr(x0)**, suggests that discriminability will be lower in the neighborhood of the prototype than for stimuli falling on the outer skirts of the prototype's category. The category limits are dependent on both on the spreading and on the number of exemplars defining the category.

According to the present model, the perceptual-magnet effect occurs as a consequence of the distance metric that is applied to the perceptual space in which representations of stimuli are stored. Under the plausible assumption that the perceptual representations of exemplars belonging to two different categories are only partially overlapping along a relevant perceptual dimension[2], the above described similarity measure will generate the perceptual-magnet effect when assigning novel stimuli to those categories. Thus, the warping of the perceptual space invoked to describe

[2] This condition is always met. Stimuli belonging to different categories must be distinguishable in some way. There may not be necessarily a single dimension that distinguishes them but at some level of complexity, including contextual dependencies, there will always be a difference between stimuli that signal different categories.

the magnet effect emerges as a corollary of the application of the present similarity metric the entire perceptual domain.

In conclusion, it seems that the observation of a perceptual magnet effect is not necessarily linked to the existence of language-specific prototypes for the different classes of speech sounds. The focal prototypes used in Kuhl's NLM theory are elegant functional higher level entities but they demand a specific non-linear metric. In my opinion, the exemplar-based account that I sketch here is a preferable approach to account for the perceptual magnet effect, since it is based on "simple" memory representations and uses a more "straight forward", cohort-based, perceptual distance.

Accounting for native language tuning

One of the problems faced by the prototype approach is the need to redefine the prototype to account for the infant's tuning towards its ambient language [12]. Prototypes must be relocatable in the perceptual space to enable the infant and the young child to learn the ambient language and also to enable re-tuning in the event of change of ambient language during the early stages of the language acquisition process. Since the prototypes' relocation process is contingent on language exposure, it should be possible to account for it on the basis of the exemplars that underlie the process, without the need to invoke the prototype notion. Within the current exemplar-based model, re-tuning is a consequence of memory decay affecting "old" exemplars. Thus, the influence of the ambient language can be modeled by including a memory decay term in the exemplar distribution, a term that fades out the representations of non-activated exemplars. Computations including this term are in principle analogous to the timeless model and will be discussed elsewhere [13].

Accounting for the species-specific perceptual magnet effect

In Kuhl's original article introducing the perceptual-magnet effect [7] it was demonstrated that the effect could not be observed for non-human species (monkeys). In my opinion, the fact that "human adults and infants show perceptual-magnet effect while monkeys do not" can be accounted for within an exemplar-based framework. One of the implications of the exemplar model is that the perceptual-magnet effect arises as a consequence of the exemplar labeling and in conjunction with the memorization process. The extent of the perceptual-magnet effect, as predicted by the exemplar model is dependent on the relation between the number of stimuli in the category and the number of stimuli in the competing categories (everything else being equal). Thus, the exemplar model predicts that the perceptual-magnet effect should be observed for monkeys if the stimuli used are made meaningful for the animals. Otherwise, a discrimination test of the type described by Kuhl [7] cannot be expected to reveal any perceptual-magnet effect for monkeys because the effect is a consequence of an underlying labeling process which, in this case, may have been irrelevant for the monkeys. Hence, within the framework of the present exemplar-based model, the behavioral differences observed between the monkeys in Kuhl's [7] experiment and the monkeys in Kuhl and Padden's [10;11] earlier experiments or the quails in Kluender, Dihel and Killeen's [6] may be due to different amounts or different types of training.

TESTING THE MODEL ON EXPERIMENTAL DATA

This section is a simple numerical exercise to illustrate how the current

model may account for some of Kuhl's experimental data. I used the data in Kuhl's (1991, fig. 3) [7] category goodness ratings for the American /i/ vowel, along the vector extending from the prototypical /i/ to the non-prototype /i/, to define the parameters of the exemplar model sketched above.

According to the assumptions of the exemplar model, goodness ratings are a rough estimate of the number of category exemplars at the stimulus location and can therefore be used to estimate the frequency distribution of the exemplars in the perceptual space. I used the variation in the category goodness, provided by Kuhl's subjects, to derive a discrimination function between the elements falling along that vector and the reference element (prototype). The results of this computation are shown in figure 2. The scale of the modeled discrimination function is arbitrary and was adjusted to force the maximum of the modeled discrimination function to be close to 1.

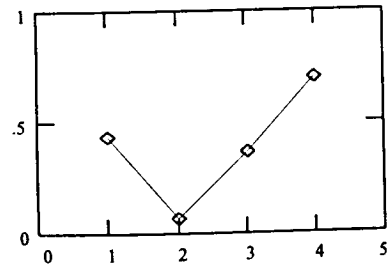


Figure 2. Discrimination sensitivity (arbitrary units) as a function of the distance (30 mel rings) to the prototype

To derive a generalization function comparable to Kuhl's experimental generalization scores, the discrimination sensitivity function in figure 2 must be integrated across rings, according to

$$AccD_{ring} := \sum_{i=1}^{ring} discr_i$$

and transformed into a generalization function, (1-AccD). The computed generalization curve is shown in figure 3.

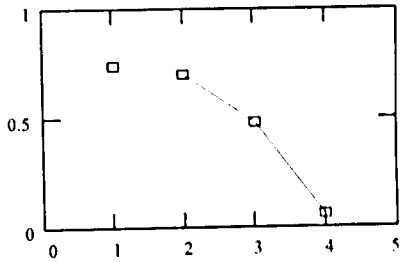


Figure 3. Computed generalization (arbitrary scale) as a function of the distance to the prototype.

The shapes of the modeled and Kuhl's experimental generalization curves resemble each other. The discrepancies between the model prediction and the experimental data may be due to the fact that goodness ratings distribute along a rank scale, but do not necessarily fulfill the interval scale requirement that underlies my model computations. To overcome this difficulty, I generated new curves based on data recently published by Sussman and Lauckner-Morano [17]. These authors reassessed Kuhl's [7] work and provide mean percent /i/-responses based on subjects' responses in a dichotomic /i/ *vs.* not-/i/ task. These percentages are more likely to meet the requirements of an interval scale. Under this assumption, the model-based discrimination function takes the form of that is displayed in figure 4.

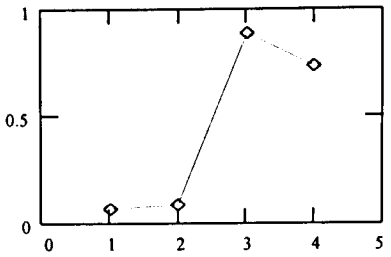


Figure 4a. Discrimination sensitivity vs. ring location predicted from Sussman and Lauckner-Morano's (1995) data.

The corresponding generalization function predicted by the model is shown in figure 4b.

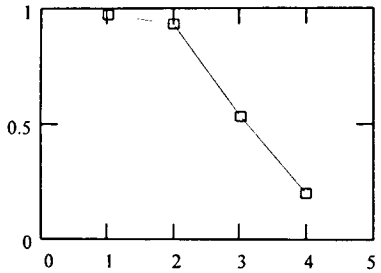


Figure 4b. Generalization associated with the discrimination sensitivity shown in fig. 4a.

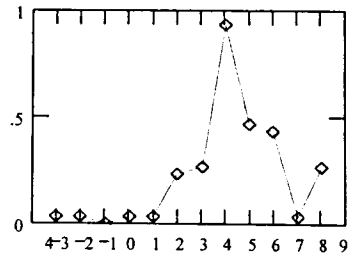


Figure 5a. Discrimination sensitivity (arbitrary units) vs. ring location. X-origin: Kuhl's (1991) prototype

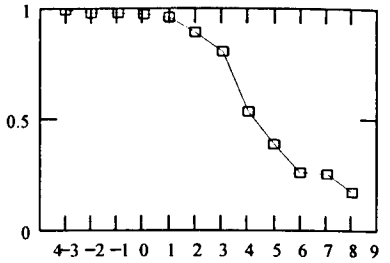


Figure 5b. Generalization computed from the discrimination function displayed in fig. 5a.

Figures 5a and 5b show the discrimination sensitivity and the generalization functions computed from Iverson and Kuhl's [2] data. The origin of the X-axis is the location of Kuhl's [7] prototype. The scale of the Y-axis is arbitrary. As illustrated by figures 4 and 5, when the model predictions are based on a more plausible interval scale, the general agreement between the predicted discrimination and the experimental is clearly improved.

CONCLUSIONS

The exemplar-based perception model that was sketched here is likely to provide a more parsimonious account of the perceptual magnet effect than Kuhl's original prototype-based account. What I tried to present here was the outline of an exemplar-based perception model that has some interesting theoretical properties but that is not, at this stage, calibrated in meaningful numerical simulations.

One important feature of the exemplar model is that it allows a rather straightforward reorganization of the listener's perceptual space as a consequence of the amount of experience with exemplars defining the relevant linguistic categories. While it should be kept in mind that a model description is obviously not an attempt to mimic neurophysiological reality, the type of computations required by this exemplar-based model are likely to be less alien to neurophysiology than the operations associated with the prototype model. Another important consequence is that an exemplar-based metric implicitly accounts for the warping of the perceptual space in the neighborhood of the prototypes.

The potential use of the present model as a unified description of other perceptual phenomena is currently under investigation [12].

In summary, while prototypes are adequate entities to describe the functional features of the perceptual magnet effect they are not necessary to explain it.

ACKNOWLEDGMENTS

I would like to thank Prof. Björn Lindblom for all the inspiring discussions during the preparation of this paper. This work is supported by The Bank of Sweden Tercentenary Foundation, grants 90-150 and 94-435.

REFERENCES

[1]Estes, W. (1994), *Classification and Cognition*, Oxford University Press, New York.

[2]Iverson, P. and Kuhl, P. (1995), "Mapping the perceptual magnet effect for speech using signal detection theory and multi-dimensional scaling", *JASA 97*, 553-562.

[3]Jusczyk, P. (1992), "Developing phonological categories from the speech signal", in Ferguson, C., Menn, L., and Stoel-Gammon, C. (Eds.), *Phonological development: Models, research, implications*, York Press, Maryland, 17-64.

[4]Jusczyk, P. (1995), "In the beginning, was the word...", to appear in Lacerda, F., von Hofsten, C. and Heiman, M. (Eds.), volume in preparation.

[5]Jusczyk, P., Hohne, E., Jusczyk, A.M. and Redanz, N. (1993), "Do infant's remember voices?", *JASA 93*, 2373.

[6]Kluender, K., Diehl, R. and Killeen, P. (1987), "Japanese quail can learn phonetic categories", *Science 237*, 1195-1197.

[7]Kuhl, P. (1991), "Human adults and human infants show a 'perceptual magnet effect' for the prototypes of speech categories, monkeys do not", *Perception and Psychophysics 50*, 93-107.

[8]Kuhl, P. (1994), "Learning and representation in speech and language", *Current opinion in Neurobiology 4*, 812-822.

[9]Kuhl, P. and Meltzoff, A. (1995), "Evolution, nativism and learning in the development of language and speech", in Gopnik, M. (Ed.), *The biological basis of language*, Oxford University Press, New York (in press: referred in Kuhl (1994).

[10]Kuhl, P. and Padden, D. (1982), "Enhanced discriminability at the phonetic boundaries for the voicing feature in macaques", Perception and Psychophysics 35, 542-550.

[11]Kuhl, P. and Padden, D. (1983), "Enhanced discriminability at the phonetic boundaries for the place feature in macaques", JASA 73, 1003-1010.

[12]Kuhl, P. Williams, K., Lacerda, F., Stevens, K. and Lindblom, B. (1992), "Linguistic experience alters phonetic perception in infants by 6 months of age", *Science 255*, 606-608.

[13]Lacerda, F. (1995), in preparation.

[14]Lindblom, B. (1986), "Phonetic universals in vowel systems", in Ohala, J. and Jaeger, J. (Eds.), *Experimental Phonology*, Academic Press, Orlando, 13-44.

[15]Rosch, E. (1978), "Principles of categorization", in Rosch, E. and Lloyd, B. (Eds.), *Cognition and categorization*, Erlbaum, Hillsdale, NJ, 27-48.

[16]Stevens, K. (1972), "The quantal nature of speech: Evidence from articulatory-acoustic data", in Denes, P. and David Jr, E. (Eds.), *Human communication: A unified view*, McGraw-Hill, New York, 51-66.

[17]Sussman, J. and Lauckner-Morano, V. (1995), "Further tests of the 'perceptual magnet effect' in the perception of [i]: Identification and change/no-change discrimination", *JASA 97*, 539-552.

DEVELOPMENTAL PATTERNS IN INFANT SPEECH PERCEPTION

Linda Polka
School of Communication Sciences and Disorders
McGill University, Montréal, Québec, Canada

ABSTRACT

This report highlights recent research investigating developmental changes in vowel perception during the first year of life. The findings provide further insights into language-specific influences in infant speech perception and also reveal language-independent perceptual biases that infants bring to the task of vowel perception. Possible interpretations of these findings are discussed and some new research questions are posed.

DEVELOPMENT OF CONSONANT PERCEPTION

Over the past 20 years, researchers have learned a great deal about the development of phonetic perception through cross-language studies of consonant perception. There are now a number of well established findings in this literature regarding the effects of age and language experience. For example, we know that, with few exceptions, young infants (aged 6 months or less) typically show the ability to discriminate both native and non-native consonant contrasts [1,2]. In addition, adults often show difficulty discriminating some non-native consonant contrasts, including contrasts that young infants have successfully discriminated, thus revealing a profound effect of language experience on phonetic perception [3, 4, 5]. It has also been clearly demonstrated that a decline in discrimination of non-native consonant contrasts can be observed as early as 8-10 months of age and is well established by 10-12 months [3, 4, 5, 6]. Studies showing that adults still possess the ability to discriminate non-native consonant contrasts when specific task or stimulus conditions are employed or training is provided, further indicate that declines in discrimination are best interpreted as a reorganization, rather than a loss of perceptual function [7]. Together these findings suggest that in the course of learning a specific language there is a perceptual attunement to the consonant categories of the native language which begins in the first year of life. This perceptual attunement serves to maintain or facilitate the discrimination of native consonant contrasts, but results in a reduced ability to discriminate some, though not all, non-native contrasts.

Two exceptions to the general developmental pattern just described are also informative. First, English infants failed to show a discrimination decline for contrasting non-native phones that English adults could readily discriminate but did not perceive as speech [8]. This suggests that perceptual attunement is evident only for phones that can somehow be assimilated to the native language. Second, infants have shown a decline in discrimination for a non-native contrast that adults readily discriminate [9]. In this case, adults did not perceive either phone to be similar to a specific native phonetic category, but they detected differences between the non-native phones that correspond to a phonemic feature contrast in their native language. This shows that infant attunement to the native language is much less sophisticated that than of adults and reflects a sensitivity to phonetic regularities rather than an ability to process phones according to a system of phonemic contrasts.

CROSS-LANGUAGE STUDIES OF VOWEL PERCEPTION

Recently, research in our lab has investigated developmental changes in cross-language vowel perception during infancy. A general question guiding this research is whether similar patterns of perceptual development are observed for vowels and consonants. This question is relevant because every spoken language is structured using vowels and consonants as segmental units. However, vowels and consonants also differ in their linguistic and communicative functions and in their acoustic properties. To the extent that perceptual attunement to native phonetic categories develops in synchrony across diverse segmental units, similar patterns of development for vowels and consonants are to be expected. In this case, we would expect to observe a decline in discrimination for some non-native vowel contrasts between 6-8 and 10-12 months of age.

To the extent that functional or acoustic factors guide or modulate the development of phonetic perception, it would be expected that vowels and consonants are associated with distinct developmental patterns. With respect to linguistic function, vowels play a more central role than do consonants as carriers of prosodic or suprasegmental information. Moreover, we now know that infants show language-specific responsiveness to some prosodic features of their native language quite early in life, before they evidence language specific attunement in consonant discrimination [10,11]. These abilities imply that, from a every early age, infants deploy considerable attention to vocalic portions of the speech stream. Therefore, linguistic influences on vowel discrimination may become evident earlier in development that they do for consonants. Recent findings reported by Kuhl et al [12] support this hypothesis.

Differences in the acoustic structures of vowels and consonants might also lead to different patterns of perceptual development. Vowels are quite prominent acoustic patterns compared to consonants in that they are typically longer and louder than consonants. Although vowels are typically perceived categorically in more naturalistic conditions (e.g. in syllable context), they are often associated with relatively high levels of within-category discrimination when studied in the categorical perception paradigm [13]. As well, cross-language studies of vowel contrasts in the categorical perception paradigm have shown language-specific effects in identification but not discrimination [14]. These findings suggest that, for acoustic reasons, language-specific attunement might not be evident at the level of vowel contrast discrimination. In this case, we would expect infants not to show a decline in discrimination of non-native contrasts across the first year of life.

Our studies addressing these issues began in Canada with a set of experiments which examined English listeners' discrimination of two German (non-English) vowel contrasts, /y/ vs. /u/ and /U/ vs. /Y/. Multiple natural exemplars produced in a /dVt/ context by a male native German speaker (from Southern Germany) were used as stimuli. The first experiment examined discrimination of these German vowels contrasts by monolingual English-speaking adults and native speakers of German [15]. English adults' discrimination of /dut/ vs. /dyt/ was close to perfect and equal to that of native German adults, revealing no effect of language experience. Discrimination of /dUt/ vs. /dYt/ was better than chance but was also significantly poorer than the German-speaking adults, revealing a small effect of language experience. English adults were also asked to match the German vowels to English vowel categories and rate the quality of the match. These data revealed that English adults perceived German /u/ vs. /y/ and /U/ vs. /Y/ as a good vs. a poor example of similar high back vowels in English (i.e. /u/ and /U/). This corresponds to the category-goodness difference assimilation pattern as described by Best [5]

Next, age-related changes in English-learning infants' ability to discriminate these German vowel contrasts was evaluated in two experiments [16]. The first experiment compared English-learning infants of 6-8 and 10-12 months on their ability to discriminate the two German vowel contrasts in the conditioned headturn procedure. The younger infants were better able to discriminate the non-native contrasts than were the older infants, consistent with previous studies with consonants. However, performance at 6-8 months also fell below levels that have been reported for non-native consonant contrasts which suggested that some decline in discrimination performance was already underway by 6-8 months. This hypothesis was tested in a second experiment in which English infants at 4 and 6 months of age were tested on the two German contrasts using a habituation

looking procedure. These data showed that 4 month olds discriminated both German vowel contrasts whereas 6 month olds failed to show evidence of discrimination for either non-native vowel contrast. Both age groups discriminated an English vowel contrast. Thus, the overall pattern of change in infant vowel discrimination across these two experiments was consistent with previous consonant work, indicating a shift from a language-general toward a language-specific pattern during the first year of life. However, our results show this shift to be underway earlier in development for vowels than for consonants.

In a second set of experiments, Ocke Bohn and I attempted to replicate and extend these findings [17]. This research, which was conducted in Montreal, Canada and in Kiel, Germany, was designed to assess the generality of the developmental pattern observed in the first study and to gather more direct evidence for language-specific influences on infant vowel perception. English-learning and German-learning infants at 6-8 and 10-12 months of age were tested on discrimination of an English (non-German) contrast, /dɛt/ vs. /dæt/, and a German (non-English) contrast, /dyt/ vs. /dut/. The English vowels were produced by a native Montreal anglophone. The German vowels were produced by a native German speaker from Northern Germany, a different German dialect from the first study. Identical instrumentation for conducting the headturn instrument was set up in Kiel to test German infants. Data were then collected using identical procedures with English-learning babies in Montreal. Monolingual adults were tested in both cities.

Discrimination of both contrasts was equally good for both German and English adults. Identification and rating data showed that English adults also perceived the German /u/ vs. /y/ as a good vs. poor example of English /u/, however the perceived difference in goodness of fit to English /u/ was larger for this contrast, due to lower ratings of German /y/ in this dialect than in the Southern dialect. For German adults, English /ɛ/ was an acceptable, though not a good, example of German /ɛ/ and English /æ/ was perceived as a poor match to either German /a/ or /ɛ/ or as failing to match any German vowel. This corresponds to the categorizable vs. uncategorizable assimilation pattern as described by Best [5].

Contrary to our expectations, the Kiel babies and Montreal babies did not perform differently on either the German or the English vowel contrast. The 6-8 and 10-12 month olds also did not perform differently in either the Kiel sample or the Montreal sample. Thus, the age-related differences found in the first study were not replicated with a new contrast nor with the same contrast produced in a different dialect. Both age and language groups had greater difficulty discriminating the English contrast than the German contrast. This study showed that infant discrimination accuracy varies for different vowel contrasts, independent of language experience, and does not always change between 6 and 12 months of age.

DEVELOPMENTAL PATTERNS IN INFANT VOWEL PERCEPTION

Overall our findings to date reveal similarities as well as differences in the development of vowel and consonant discrimination. The evidence for an influence of language experience by 6 months shown in our first study is consistent with Kuhl et al's findings of language-specific effects on infant perception of within-vowel category differences [12]. It is interesting that these language-specific effects in vowel perception are evident around the same age that language-specific processing of various aspects of prosodic structure aref foudn, e.g. [11]. This synchrony may be interpreted as evidence of an attentional focus on vocalic information in early infancy. However, further declines between 6-8 and 10-12 months of age show that perceptual attunement for vowels also continues through the later half of the first year, just as has been observed for consonants. That we find both converging and diverging results for vowels and consonants at different ages raises the question of whether a single processing mechanism can account for both the early and later changes in infant vowel discrimination. This issue is discussed further in [18].

As mentioned earlier, previous consonant studies have not always shown there to be a decline in discrimination of a non-native contrast between 6-8 and 10-12 months. We now know that the same observation applies to cross-language vowel discrimination. However, it is interesting to note that, to date, the contrasts which have failed to show a decline, and therefore a language effect, have differed for vowels and consonants with respect to the similarities that adults perceive between the non-native phones and their native language phonetic categories. As outlined in the model proposed by Best [9], such differences can be informative as to the kinds of phonetic regularities that infants begin to detect in the native language with increasing age and language experience.

In the case of consonants, Best has reported on two non-native contrasts, to date, in which no discrimination decline was observed in English-learning infants. One contrast, a Zulu click contrast, was not assimilable to the native language by English adults [8]. The other contrast was the Ethiopian ejective stop contrast, /p'ɛ/ - /t'ɛ/. English adults perceived /p'ɛ/ to be highly similar to English /p/ and /t'ɛ/ to be highly similar to English /t/. All other consonants contrasts that have been tested with infants have shown developmental decline, even when adults could easily discriminate them. These contrasts include a variety of assimilation patterns in adults including 1) a single category mapping in which both non-native phones are perceived as being quite similar to the same native phonetic category, 2) a two category mapping in which each non-native phone is perceived as being similar to a different native phonetic category, and 3) a category goodness difference mapping in which the non-native phones are perceived as good vs. a poor match to the same native phonetic category.

In the case of vowels, we have failed to show a decline for a contrast that was perceived as a category goodness difference (i.e. /u/-/y/ in the Northern dialect). However, this same contrast showed a decline for tokens from a Southern dialect which was associated with a smaller difference in category goodness. In comparison, with consonants, the category goodness assimilation has been consistently associated (so far) with a perceptual decline [9]. The other vowel contrast failing to show any language effects in our work, English /dɛt/ vs. /dæt/ was assimilated by German adults as a categorizable vs. an uncategorizable vowel. To our knowledge, no consonant contrast showing this assimilation pattern has been tested with infants.

Overall, the existing data show that with age and experience infants show some attunement to native vowels in that they ignore some vowel differences that are not meaningful in their native language, provided that the differences correspond to a single native vowel and are sufficiently small, whereas they continue to discriminate other differences that don't convey word meaning in their native language. On the other hand, it seems that infants ignore a wider range of consonant differences that are not functional in the native. They only appear to continue to discriminate consonant differences if they are remarkably similar to (and perhaps indistinguishable from) specific native language phones or when presented phones that are not assimilable to the native language.

Certainly further research is needed before any strong conclusions can be drawn regarding differences in how infant perception of vowels and consonants becomes tuned to the native language. However, patterns in the existing data suggest that we should continue to entertain the hypothesis that language-specific processing is expressed differently in the development of vowel and consonant perception. On the basis of our findings and the language-specific effects demonstrated by Kuhl et al. it could be predicted language experience brings about subtle changes in the structure of vowel categories such that language effects may only be observed for discrimination of non-native vowel contrasts in which both vowels are quite similar to a single native vowel category.

PERCEPTUAL ASYMMETRIES IN INFANT VOWEL PERCEPTION

In our first set of infant experiments we noted very striking directional

asymmetries. That is, infants performed differently depending on which direction they were presented a vowel change, which we had varied simply as a matter of experimental control. We became quite interested in these directional asymmetries because they suggested an interesting connection between our work and findings reported by Kuhl et al [12] showing there to be language-specific influences on the internal structure of vowel categories by 6 months of age.

In this work, Kuhl et al started with central /i/ vowel that a group of English adults had rated as being a very good example of English /i/. Additional /i/ vowels were then created by increasing and decreasing F1 and F2 values (in equal mel steps) from this central, prototypic /i / such that the peripherals vowels formed four equally-spaced rings surrounding the central vowel in an F1 by F2 space. Ratings of the peripheral vowels (as an example of English /i/) decreased as distance from the central vowel increased.

In the same way, vowel stimuli were created with a good or prototypic example of Swedish /y/ as the central vowel and four rings of /y/ variants surrounding it. Here also, ratings of the vowels (as examples of Swedish /y/) by Swedish adults decreased as distance from the central vowel increased.

Kuhl et al found that English infants showed poorer performance in discriminating the central English /i/ vowel from the peripheral /i/ vowels surrounding it, compared to their performance in discriminating the central Swedish /y/ from each of the peripheral vowels surrounding it. Swedish infants showed the reverse pattern, i.e. better performance when discriminating the English prototype from it's peripheral variants than when discriminating the Swedish prototype from it's peripheral variants. These data were interpreted as evidence for a language-specific perceptual magnet effect. Essentially, the claim is that, with language experience, a native vowel begins to act like a perceptual magnet which appears to pull more peripheral vowels toward it, thus effectively shrinking the perceptual space surrounding the vowel prototype. The magnet effect enhances perceptual generalization to the prototype and in doing so makes discrimination of differences near the prototype more difficult.

The directional asymmetries that we observed in our first infant study (testing English infants on German /u/-/y/ and /U/ vs. /Y/), were consistent with the notion of a language-specific perceptual magnet effect. As shown in Figure 1 below, among the 6-8 month olds, discrimination was significantly poorer for infants tested with /u/ or /U/ as the reference vowel (i.e. on a vowel change from /u/ to /y/ and from /U/ to /Y/) compared to infants presented the change in the reverse direction. Thus, within each contrast, the back vowel appears to act like a perceptual magnet. Our experiments with English adults showed that German /u/ and /U/ are more typical of English vowels than are /y/ and /Y/ [15]. Thus, within each contrast the vowel which acts like a magnet was the more typical (or English-like) vowel. As such, the directional asymmetries that we observed in infant discrimination of non-native vowel contrast could be taken as further evidence that vowel perception is organized around language-specific prototypes (i.e. "best" or most typical instances) by 6 months of age.

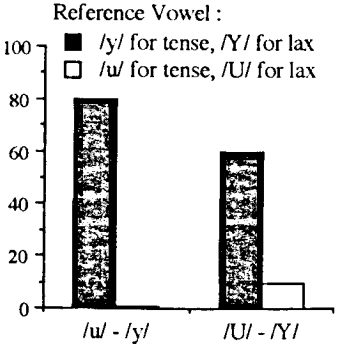


Figure 1. Proportion of English 6-month-olds reaching criterion on German /u/-/y/ and /U/-/Y/ plotted separately for infants tested with different reference vowels.

We designed our second infant study (with German and English infants) to test this hypothesis. If the directional asymmetries indicate a language-specific perceptual magnet effect, we expected to replicate the same directional effect (as shown in first study) for English infants tested on the German contrast and to find no direction effect for German infants tested on the German contrast. In addition, we expected German infants tested on the English contrast to show a direction effect in which /ɛ/ acts like a perceptual magnet. That is, we expected discrimination to be poorer in infants tested with /dɛt/ (the more German-like vowel) compared to those infants tested with /æ/ (the less German-like vowel) as the reference vowel. Likewise, we expected to find no direction effect in the English infants tested on the English vowel contrast.

For the German contrast, /dut/ - /dyt/, the directional asymmetry was replicated and was quite robust. However, as indicated in Figure 2, this asymmetry, showing /u/ to act like a perceptual magnet, was evident in both the English infants and the German infants.

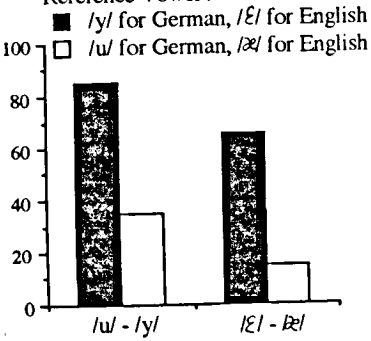


Figure 2. Proportion of infants (across both age and language groups) reaching criterion on English /æ/ - /ɛ/ and German /u/-/y/ plotted separately for infants tested with different reference vowels.

For the English contrast, /dɛt/ vs. /dæt/, we also found a strong directional asymmetry, as shown in Figure 2. However, the asymmetry was in the opposite direction from our prediction, showing poorer discrimination when /æ/ served as the reference compared to when /ɛ/ was the reference vowel. This direction effect, showing /æ/ to act like a perceptual magnet, was also evident in both German and English infants. The direction effect did not interact with age or language experience for either vowel contrasts.

Clearly, the pattern of directional asymmetries in our second study are inconsistent with the notion of a language-specific perceptual magnet effect. These asymmetries also cannot be explained as an effect of markedness because, in the English contrast, the vowel which acts like a perceptual magnet, /æ/, occurs much less frequently across languages compared to /ɛ/. Therefore, these asymmetries point to a language-independent bias that infants bring to the task of vowel perception. The only consistency that we have noted in these asymmetries is that, within each contrast, the vowel which appears to act like perceptual magnet is produced with a more extreme articulatory posture (re vowel height and front-back dimensions). Thus, there appears to be a greater perceptual stability associated with vowels produced with more extreme articulatory postures. Ocke Bohn and I are continuing to test this hypothesis in studies of German vowel contrasts by German infants. So far, we have found an asymmetry associated with German /e/ vs. /I/, which is consistent with our hypothesis. We have also noted asymmetries in other vowel studies with infants and adults which are consistent with this interpretation (see [17]).

Overall, several clear conclusions can be drawn regarding perceptual asymmetries. First, the direction of a perceptual asymmetry is not language-specific, but is an default, language-independent perceptual bias. Second, these directional effects are quite robust in infants, whereas in adults we have found little or no evidence for these asymmetries using similar testing procedures. Given these age differences, it is reasonable to predict that the magnitude of a directional effect may be altered by language experience. Our current data fail to provide a good test of this prediction.

At present there are more questions than answers surrounding the significance of these perceptual asymmetries. One possibility that we are currently considering is that these asymmetries reflect the operation of mechanisms involved in normalization for talker differences. The corner vowels

in a traditional vowel space, which correspond to extreme articulatory postures, define constraints placed on vowel productions by the size and shape of the vocal tract. These corner vowels, especially /i/ and /u/, have been shown to be particularly stable in that large deviations in articulation are associated with small changes in formant frequency [19]. Basic research in vowel perception has also suggested that the corner vowels, which are less likely to overlap acoustically with other vowels, might provide particularly clear cues to vocal tract size [20]. While adults are likely to employ a wide range of information in calibrating for different talkers, infants might rely on a more restricted set of cues in vowel normalization. Thus, the enhanced perceptual stability of more extreme vowels might reflect their reliance on particular vowel cues in calibrating for differences in vocal tract size. There are also some interesting changes in infants vowel production in the first year of life which suggest that extreme vowels play a special role in the infant's mapping of the vowel space. See [21] for a review and discuss of this work.

In future research, we will address three questions to further clarify the meaning of perceptual asymmetries. First, are there latent asymmetries in adult cross-language vowel discrimination that will become evident under test conditions which preclude ceiling performance levels? There was, in fact, a small direction effect, in the same direction as shown by the infants, in English adults' discrimination of the German /dUt/ vs. /dYt/ contrast. Our expectation is that we will be able to show directional asymmetries in adults for the other non-native vowel contrasts, either in reaction measures or in a dual task paradigm which lowers overall discrimination accuracy. This would suggest that these asymmetries reflect an inherent phonetic bias that becomes weaker, but is not lost, with age.

If asymmetries are evident in adults as well as infants, it will then be interesting to ask whether such asymmetries reflect a species-specific perceptual bias. To the extent that these asymmetries reflect auditory processing constraints, we would expect animals that possess a similar auditory system to show the same patterns. On the other hand, if an appropriate animal failed to show perceptual asymmetries, it would suggest that the directional effects are showing a phonetic bias that, perhaps, reflects a sensitivity to vocal tract constraints.

Finally, it will be also useful to explore the conditions which generate these directional effects by using other stimulus sets. For example, it will be informative to determine whether directional effects are found only in discrimination of vowels that specify a single talker. If this were the case, it would increase support for the hypothesis the biases evident in directional asymmetries reflect mechanisms used in mapping a specific talker's vowel space. Alternatively, it is possible that directional asymmetries might be also be observed in discriminating a vowel contrast in productions from multiple talkers. This outcome would imply that these perceptual tendencies may potentially contribute to the development of talker-independent phonetic categorization skills.

SUMMARY

Overall, studies to date point to similarities as well as difference in the development of vowel and consonant perception. However, a great deal more research is needed to arrive at a comprehensive understanding of the development of infant phonetic processing abilities. On the basis of our present findings and related studies, it appears that effects of language experience on a vowel perception are subtle and occur against a background of strong language-independent perceptual biases. Future research should strive to clarify the contribution of inherent perceptual biases and language-specific influences in the development of phonetic perception. With this knowledge, we can begin to explore the significance of these developmental changes in the child's acquisition of word meaning.

REFERENCES

[1] Trehub, S.E. (1976), "The discrimination of foreign speech contrasts by infants and adults", *Child Development*, vol. 47, pp. 466-472.

[2] Streeter, L.A. (1976), "Language perception of 2-month-old infants shows effects of both innate mechanisms and experience", *Nature*, vol. 259, pp. 39-41.

[3] Werker, J.F., Gilbert, J.H.V., Hunphrey, K., & Tees, R.C. (1981), "Developmental aspects of cross-language speech perception", *Child Development*, vol. 52, pp. 349-353.

[4] Werker, J.F., & Tees, R.C. (1983), "Developmental change across childhood in the perception of non-native speech sounds", *Canadian Journal of Psychology*, vol. 37, pp. 278-286.

[5] Best, C.T. (1994), "The emergence of native-language phonological influences in infants: A perceptual assimilation model", In J. Goodman & H.C. Nusbaum (Eds.) *The Development of Speech Perception: The Transition From Speech to Spoken Words*. Cambridge MA: MIT Press.

[6] Werker, J.F., & Tees, R.C. (1984), "Cross-language speech perception: Evidence for perceptual reorganization during the first year of life", *Infant Behavior and Development*, vol. 7, pp. 49-63.

[7] Werker, J. F. & Pegg, J. E. (1992), "Infant speech perception and phonological acquisition." In C. Ferguson, L. Menn, and C. Stoel-Gammon (Eds.) *Phonological Development: Models, Research, and Implications*. Parkton Maryland: York .

[8] Best, C.T., McRoberts, G.W., & Sithole, N.N. (1988). "Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants", *Journal of Experimental Psychology: Human Perception and Performance*, vol. 14, pp. 345-60.

[9] Best, C. T. (1994), "Learning to perceive the sound pattern of English," In C. Rovee-Collier & L. Lipsitt (Eds.), *Advances in Infancy Research* Norwood, NJ: Ablex .

[10] Mehler, J., Jusczyk, P.W., Lambertz, G., Halstead, N., Bertoncini, J., & Amiel-Tison, C. (1988), "A precursor of language acquisition in young infants", *Cognition*, vol. 29, pp. 142-178.

[11] Jusczyk, P. W. , Friederici, A.D. , Wessels, J.M.I., Svenkerud, V.Y., & Jusczyk, A.M. (1993), "Infants' senstivity to the sound pattern of native language words", *Journal of Memory and Language*, vol. 32, pp. 402-420.

[12] Kuhl, P.J., Williams, K.A., Lacerda, F., Stevens, K.N., & Lindblom, B. (1992), "Linguistic experience alters phonetic perception in infants by 6 months of age", *Science*, vol. 255, pp. 606-608.

[13] Repp, B.H. (1984), "Categorical perception: Issues, methods, and findings", In N.L. Lass (Ed.) *Speech and Language: Advances in Basic Research and Practice*. New York: Academic Press.

[14] Beddor, P. S. & Strange, W. (1982), "Cross-language study of the oral-nasal distinction", *Journal of the Acoustical Society of America*, vol. 71, pp. 1551-1561.

[15] Polka, L. (1995), "Linguistic influences in the adult perception of non-native vowel contrasts, *Journal of the Acoustical Society of America*, vol 97, pp. 1286-1296.

[16] Polka, L. & Werker, J.F. (1994), "Developmental changes in perception of non-native vowel contrasts", *Journal of Experimental Psychology: Human Perception and Performance*, vol. 20, pp. 421-435.

[17] Polka, L & Bohn, O. (submitted), " A cross-language study of vowel perception in English-learning and German-learning infants", *Journal of the Acoustical Society of America*

[18] Werker, J.F., Lloyd, V. Pegg, J. & Polka, L. (1995), "Putting the baby in the bootstraps: Toward a more complete understanding of the role of input in speech processing" In J. Morgan & K. Demuth (Eds.) *Signal to Syntax*. Hillsdale, NJ: Lawrence Erlbaum

[19] Stevens, K. (1989), "On the quantal nature of speech", *Journal of Phonetics*, vol 17, pp. 3-45.

[20] Lieberman, P. (1984) *The biology and Evolution of Language* Cambridge: Harvard University Press.

[21] Polka, L. "The role of initial perceptual biases and language-specifc learning in infant speech perception development", To appear in F. Lacerda (Ed.) *Transitions in Perception, Cognition, and Action in Early Infancy.*

NEUROBIOLOGY OF LANGUAGE AND SPEECH

Victoria A. Fromkin

University of California, Los Angeles

ABSTRACT

The interest in the neurobiology of language and speech goes back at least 3000 years. Its recent resurgence reflects the concern for explanation as well as description. Much current research utilizes the new technologies for studying brain/behavior relationships such as Magnetic Resonance Imaging (MRI) Positron Emission Tomography (PET), and Event Related (Brain) Potentials (ERP). This symposium discusses different aspects of this question.

BACKGROUND

The interest in the neurobiology of language and speech and in the brain/behavior interface goes back at least 3000 years.[1] Its recent resurgence reflects the concern among linguists and phoneticians in explanation as well as description. We are no longer satisfied with knowing that the obicularis oris is activated to a greater extent in the production of an initial than a final /p/ [2] although such questions remain important in our understanding of the phonetic realization of phonological units; in addition, we seek answers to questions such as those raised by Chomsky in 1988 [3]: 'What are the physical mechanisms that serve as the material basis for (the) system of (linguistic knowledge) and for the use of this knowledge?' (p 3) 'In the study of language we proceed abstractly, at the level of mind, and we also hope to be able to gain understanding of how the entities constructed at this abstract level and their properties and the principles that govern them can be accounted for in terms of properties of the brain.' (p 8)

Although we do not have any final answers to the question of the neural structures underlying linguistic units such as sentences, phrases, words, or phonological or phonetic segments, the new technologies for studying brain/behavior relationships such as Magnetic Resonance Imaging (MRI) Positron Emission Tomography (PET), and Event Related (Brain) Potentials (ERP) are beginning to provide some answers. In addition, the utilization of these new technologies to the study of language and speech disorders following focal damage to the brain, contributes to our understanding of the neurobiology of normal language and speech.

That this is hardly a new issue is shown by the fact that in the 135th Psalm, one find an implicit recognition of the left brain / language interface (although contralateral brain function was of course not understood). The verse states: "If I will forget thee, Jerusalem, let my right hand die -- let my tongue cleave to the roof of my mouth."

In the New Testament, St. Luke reports that Zacharias could not speak but could write, predating the modern observations of the independence of linguistic components by two millennia.

As pointed out in the Whitaker contribution to this session, observations of language loss with intact general

intelligence are found in the medical records written on papyrus in 1700 B.C.E. by Egyptian surgeons, long before the philosophers of ancient Greece speculated about the brain/mind relationships. Although neither Plato nor Aristotle recognized the brain's crucial function in cognition or memory as shown by Aristotle's suggestion that the brain is a cold sponge whose function is to cool the blood, in the same period , the Graeco-Roman physicians' Hippocratic Treatises (written from 400 BCE to 135 CE) reveal their understanding of the role of the brain in noting that language and speech disorders result from cerebral trauma or brain disease and that loss of speech often occurred simultaneously with paralysis of the right side of the body. They also showed an understanding of the separation of linguistic competence and performance in their observation that language loss may occur without the loss of speech and vice versa. [4]

Other writers and scholars of the ancient classical world and the mediaeval period provide us with a wealth of information on aphasia -- the loss of distinct linguistic abilities -- with a preservation of nonlinguistic cognitive functions, as well as differential impairment and preservation of different linguistic abilities. Over 2000 years ago Valerius Maximus and Pliny described the Athenian scholar who in the words of Pliny "...with the stroke of a stone, fell presently to forget his letters only, and could read no more; otherwise his memory served him well enough." [1]

Numerous clinical descriptions of patients with language deficits and preserved non- linguistic cognitive systems were published from the 15th to the 18th century. [5]

Differential breakdown of language and components of language were reported in detail throughout the 16th to the 19th century. Such descriptive reports strongly support current day views of the modularity of mind, the independence of language from other cognitive systems and general intelligence, and also, the modularity of the components of the mental grammar itself.

Whitaker's discussion in this symposium on the importance of, the pitfalls, and gaps in the history of neurolinguistics from 1600 to 1900 provides new insights regarding this important issue. I will therefore only mention the two 19th C names probably most familiar to linguists and phoneticians -- Broca [6] and Wernicke [7]. Broca's seminal paper of 1861 reported on the specific role of the left hemisphere in relation to language localized brain areas..

In 1874, Wernicke [7] provided further evidence when he pointed out that damage in the posterior portion of the left temporal lobe results in a different form of language breakdown than that occurring after damage to the frontal cortex.

The fact that focal injuries to different parts of the brain not only lead to selective cognitive disorders, but to damage of distinct components of language or of specific linguistic processing mechanisms provide a major reason for the linguistic interest in aphasia.

Blumstein's contribution to this symposium draws on the earlier findings of Broca and Wernicke and on all the research which has followed. She discusses specific phonological and phonetic deficits and the linguistic and

non-linguistic basis for these in reference to anterior and posterior lesions.

Jakobson's Legacy

The years which followed Broca's and Wernicke's discoveries stimulated neurologists throughout the world such as Broadbent [8] and Bastian [9] in Britain, Pick [10] and Salomon [11] in Germany, Moutier [12] in France, and Hughling Jackson [13,14] in the US, among others, to apply linguistic analyses to aphasia data. But Roman Jakobson [15,16, 17, 18,] was the first linguist to apply linguistic theory to aphasia research.

Following up on the insights of Baudouin de Courtenay in 1895 and Ferdinand de Saussure in 1879 [19] who had expressed the belief that a study of language pathology could contribute to linguistics, Jakobson also stressed the other side of the coin, the contribution of linguistics to the study of aphasia, stating that "any description and classification of aphasic syndromes must begin with the question of what aspects of language are impaired". [17] He despaired over the fact that" the linguist's contribution to the investigation of aphasia is still ignored" and also believed that "Linguists are also responsible for the delay in undertaking a joint inquiry into aphasia."

Jakobson would have been pleased to have seen the developments that have taken place in the last number of years, which led to the holding of this session on the neurobiology of language.

His notion of the hierarchical organization of linguistic entities proposed in his early works on phonology, found its expression in the theory of markedness discussed in relation to phonological paraphasias by Kean in this symposium. Kean also provides strong evidence for the correctness of Jakobson's view of how linguistic theory can contribute to our understanding of aphasia. [20, 21] Her paper, as well as Blumstein's, provides rich evidence for the insights provided by linguistic theory.

Except for Jakobson, few linguists followed up the early interest in linguistics by neurologists who drew on linguistic concepts in their investigations of aphasia. The first linguist to follow Jakobson's lead was Blumstein [23] who applied his theories of distinctive features and markedness to an experimental investigation of aphasic phonemic errors and who further emphasized Jakobson's view that an analysis of aphasic errors can contribute to phonological theory, itself. In her paper at this congress she, as well as Kean, provides additional evidence in support of her original finding that in aphasic speech errors (like normal errors) the direction of substitution is from marked features (nasal /n/) to unmarked (non-nasal /d/).

Blumstein and Kean show that evidence from aphasia presents a partial answer to whether the mental grammar, that is, the representation of linguistic knowledge in the mind and brain, is itself decomposed into components like those projected by linguists on the basis of language evidence alone.

Furthermore, as Blumstein's paper points out, speech deficits in aphasia may be due to either linguistic or non-linguistic causes, taking different forms in the two cases.

Early views of aphasia tended to treat the different syndromes as either expressive or comprehension disorders. Whitaker, in his paper, points out that Broca was concerned only with speech production since comprehension was considered to be out of the province of

'real science', i.e. the province of the philosophers and others concerned with the mind. (How reminiscent of the behaviorist period in American linguistics.) The early view that agrammatism was a disorder of speech production with intact speech comprehension was upset in the 1970's when controlled experimental studies showed that when comprehension depends on the syntactic structure of sentences, syntactic comprehension deficits -- asyntactic comprehension -- also arise in these patients. [24,25,26,27,28]

This suggests that a syntactic representation or processing deficit was involved, again supporting the notion of distinct and possibly independent components.

SPEECH, SIGN, AND LANGUAGE

Aphasia was originally seen as a problem in speech -- production in relation to Broca's aphasia, and comprehension/perception in relation to Wernicke's aphasia. However, both Blumstein's and Kean's papers make clear that many speech problems may be more properly viewed as language, not speech disorders.

Perhaps the most telling and dramatic findings on the brain / language / speech relationship is revealed by the research on sign language conducted by Bellugi and her colleagues [29]. The linguistic study of sign language over the last 25 years has already revealed that these languages of the deaf have all the crucial properties common to all spoken languages, including highly abstract underlying grammatical and formal principles.

Since the same abstract linguistic principles underlie all human languages -- spoken or signed -- regardless of the motor and perceptual mechanisms which are used in their expression, it is not surprising that deaf patients show aphasia for sign language similar to the language breakdown in hearing aphasics following damage to the left hemisphere.

The left cerebral hemisphere is not dominant for speech but for language, the cognitive system underlying both speech and sign. Hearing and speech are not necessary for the development of left hemispheric specialization for language.

Furthermore, while deaf patients with focal lesions show marked sign language deficits, they can correctly process non-language visual-spatial relationships. The left cerebral hemisphere is thus not dominant for speech, as had been suggested, but for language, the cognitive system underlying both speech and sign. Hearing and speech are not necessary for the development of left hemispheric specialization for language.

This has been a crucial point in determining that the left hemisphere specialization in language acquisition is not due to its capacity for fine auditory analysis, but for language analysis per se.

CT, PET, MRI, AND ERP STUDIES

Aphasia studies have been crucial in the investigation of the brain/language-speech relationship. The advent and development of new imaging technologies such as computerized tomography (CT), Magnetic Resonance Imaging (MRI), Positron Emission Tomography, and Event Related Brain Potential (ERP) studies now make possible greater access to the macroscopic neuroanatomy and neuropathology of living humans. [30] The first use of these techniques in studies of brain and language paralleled the aphasia studies approach, i.e. the 'lesion method'. As

stated by the Damasios, "The essence of the lesion method is the establishment of a correlation between a circumscribed region of damaged brain and changes in some aspect of an experimentally controlled behavioral performance. Given a preexisting theory about the operation of the normal brain and how it would mediate the performance of (a task such as speech production or comprehension) the lesion (the area of brain damage) can be seen as a probe to test the validity of the theories." (p 8) [11]

Through the use of these techniques we have a much better understanding of the localization of function and of the neuroanatomy underlying language and speech.

PET allows us to look at the normal (as well as the disordered) brain in vivo as shown by blood flow. Similarly, functional MRI's and ERPs allow one to see what is going on in the brain during various task performances or response to different kinds of stimuli. ERP experiments study scalp electrical activity as recorded from electrodes placed on the scalp according to a universally agreed on set of positions following different stimuli presented to the subjects.

A number of PET experiments have examined phonological processing. [31] Despite the sanguine view of all such studies, we should keep in mind Poepple's [26] caveat regarding the many problems which remain in the interpretation of the data. He points out that a comparison of PET studies shows that because of great variation across subjects and tasks, we can not "attribute phonetic/phonological processes to a specific region of the brain". This does not mean there may not be such a region, (but see Blumstein's conclusion that "anterior as well as posterior brain structures are implicated in the auditory processing of speech.") It does mean that while we welcome the new technology, we cannot abandon other traditional approaches and, as is obvious, the use of all new technological tools should be motivated by linguistic theory.

We are reminded of Sapir's warning in 1925 [32] "Mechanical and other detached methods of studying the phonetic elements of speech are, of course, of considerable value, but they have sometimes the undesirable effect of obscuring the essential facts of speech-sound psychology." At the same time, as shown in Blumstein's paper, instrumental acoustic analysis is vital in trying to uncover impairments in the speech of aphasic patients which are not easily perceivable by the human ear.

Furthermore, the ERP studies now being conducted are proving to provide important evidence regarding both representation and processing. Hagoort and Brown's paper in this symposium provides an excellent overview of what is involved in such studies. They claim that this new technique provides a real-time neurophysiological measure of speech processing with temporal resolution superior to other imaging techniques

Their findings of different neurophysiological responses to semantic and syntactic processing replicates another study by Neville and co-researchers [33]

Using a different experimental task, Neville's group found that syntactically well-formed but semantically anomalous sentences produced a pattern of brain activity (ERPs) that is distinct in timing and distribution from the patterns elicited by syntactically deviant sentences, and further, that different types of syntactic

deviance produced distinct ERP patterns as illustrated in the examples below:

1. #The man admired Don's headache of the landscape.
2. *The man admired Don's of sketch the landscape..
3. *What$_i$ was [$_{NP}$ a sketch of t_i] admired by the man?

As in Hagoort and Brown's studies, the semantic anomalies sentences such as 1. produced a negative potential, N400, that was bilaterally distributed and was largest over posterior regions. The phrase structure violations such as in 2. and 3 enhanced the N125 response over anterior regions of the left hemisphere, and elicited a negative response (300-500 msec) over temporal and parietal regions of the left hemisphere. The specific types of syntactic violations such as specificity constraints, and subjacency constraints elicited distinct timing and distribution responses.

They conclude: "the distinct timing and distribution of these effects provide biological support for theories that distinguish between these types of grammatical rules and constraints and more generally for the proposal that semantic and grammatical processes are distinct subsystems within the language faculty."

CONCLUSIONS

The four papers presented in this session aim at illustrating the importance of the research on the brain /language /speech interface. Whether one uses the new technologies and experimental techniques to investigate the speech production and comprehension of normals or of aphasics we are beginning to gain a better understanding of the neurobiology of language and speech.

REFERENCES

[1] Benton, Arthur L. and Robert J. Joynt (1960) Early descriptions of aphasia. *Archives of neurology* 3.109/205-126/222

[2] Fromkin, V. A. (1966) Neuro-muscular specification of linguistics units. *Language and Speech* 15.219-242

[3} Chomsky, N. (1988) *Language and Problems of Knowledge: The Managua Lectures.* Cambridge, Mass. MIT Press.

[4] Breasted,J.H. (1930) *The Edwin Smith Surgical Papyrus.* Chicago: University of Chicago Press

[5] Clarke,E. and O'Malley, C.D. (1968) *The human brain and spinal cord: A Historical study illustrated by writings from antiquity to the twentieth century.* Oxford:Sanford Publications

[6] Broca, Paul (1861) Nouvelle observation d'aphemie produite par une lesion de la moitie posterieure des deuxieme et troisieme circumvolutions frontales. *Bulletin de la Societe Anatomique de Paris* 3.398-407

[7] Wernicke, C. (1874) The aphasic symptom complex:a psychological study on a neurological basis. Kohn and Weigert, Breslau reprinted in R.S. Cohen and M.W. Wartofsky (eds) *Boston Studies in the Philosophy of Science, vol 4.* Reidel, Boston

[8] Broadbent,W.H. (1879) A case of peculiar affection of speech, with commentary *Brain 1,* 484-503

[9] Bastian,C. (1887) On different kinds of aphasia with special reference to their classification and ultimate pathology. *British Medical Journal* 2, 931-6

[10] Pick,A. 1913 *Die Agrammatischen Sprachstorungen*. Springer, Berlin

[11] Salomon, E. 1914 Motorische Aphasie mit Agrammtismus und Sensorischagrammatischen Storungen. *Monatschrift fur Psychiatrie und Neurologie* 35, 181-208, 216-75

[12] Moutier, F. 1908 *L'aphasie de Broca* Paris:Steinheil

[13] Jackson, J.H. (1874) On the nature of the duality of the brain. in . J. Taylor (ed.) *Selectged Writings of John Hughlings Jackson/*. vol 2. New York: Basic Books

[14] Jackson, J.H. (1878) On affections of speech from disease of the brain. *Brian*. 1.304-30; 2. 203-22; 323-56

[15] Jakobson, R. (1940) *Kindersprache, Aphasie and allgemeine Lautgesetze* Almqvist u. Wilsells, Uppsala. Reprinted as *Child Language, Aphasia, and Phonological Universals* 1968. Mouton, The Hague.

[16] Jakobson, Roman (1955) Aphasia as a linguistic problem. in *On Expressive Language*. edited by H. Werner. Worcester, Mass: Clark University Press,. 69-81

[17] Jakobson, R. (1964) Towards a liguistic typlogy of aphasic impairments. *Disorders of language*. Edited by A.V.S.deReuck and M. O'Connor Boston: Little,Brown 21-41

[18] Jakobson, R. (1970) Toward a linguistic classification of aphasic imnpairments. *Selected Writings II*. The Hague: Mouton.

[19 Saussure, Ferdinand de. (1916) *Cours de linguistique generale*. Paris. Payot.

[20] Kean, Mary Louise 1977 The linguistic interpretation of aphasic syndromes: agrammatism in Broca's aphasia, an example *Cognition 5*, 9-46

[21] Kean, M-L. (Editor) (1985). *Agrammatism* Academic Press, New York

[22] Blumstein, S. (1973) *A Phonological Investigation of Aphasic Speech*. The Hague: Mouton

[23] Caplan, D. and Futter, C. (1986). Assignment of thematic roles by an agrammatic aphasic patient. *Brain and Language* 27. 111-134

[24] Caramazza,A. and E. Zurif (1976) Dissociation of algorithmic and heuristic processes in language comprehension: Evidence from aphasia *Brain and Language* 3, 572-82

[25] Grodzinsky,Y. (1990) *Theoretical perspectives on Language Deficits*. Cambridge, Mass: MIT Press

[26] Linebarger, M., M.F. Schwartz, and E.M. Saffran. (1983) Sensitivity to grammatical structure in so- called agrammatic aphasics.*Cognition*13.361-92

[27] Schwartz, M.R., Linebarger,M.C. and Saffran, E.M. 1985. The status of the syntactic deficit theory of agrammatism in *Agrammatism*. edited by M-L. Kean editor. NY: Academic Press. 83-124.

[28] Zurif,E.B. and Grodzinsky, Y. 1983 Sensitivity to grammatical structure in agrammatism: A reply to Linegarger et al. *Cognition* 15: 207-213

[29] Poizner, H.E., Klima, E. and U. Bellugi. (1987) *What the Hands Reveal About the Brain*. Cambridge, Mass: MIT Press

[30] Damasio,H. & Damasio, A.R. (1989) *Lesion Analysis in Neuropsychology*. Oxford, NY: Oxford University Press

[31] Poeppel, D. (1993) A critical Review of PET Studies of Language. Dept. of Brain and Cognitive Sciences, MIT. Unpublished ms.

[32] Sapir, E. (1925) Sound patterns in language. *Language*. 1. 37-51

[33] Helen Neville, Janet L. Nicol, Andrew Barss, Kenneth I. Forster, and Merrill F. Garrett. 1991. Syntactically Based Sentence Processing Classes: Evidence from Event-Related Brain Potentials, *J of Cognitive Neuroscience*. Vol. 3/.2/ 151-165) in experiments involving event-related potentials or ERP's. Such

ROOTS: 5 NOTES ON THE HISTORY OF NEUROLINGUISTICS

Harry A. Whitaker

Psychologie, Université du Québec à Montréal, Montréal, Québec, Canada

Prolegomena 1: using historical studies

Norman Geschwind's work in the 1960's [1,2] was influential in kindling an interest among contemporary neuropsychologists in the historical contributions of late 19th century neuroscientists, notably that of Wernicke, Liepmann and Dejerine whose models of brain function had, over the intervening half-century, slipped from their former prominence. Geschwind's interest in history was straightforward: the connectionist model of brain function which he had come to believe in, had clear origins in the work of these earlier scientists; it was good scholarship to recognize that indebtedness as well as interesting and entertaining, all of which we should consider a first 'use' of history: finding the roots of scientific concepts, the background of contemporary ideas.

Concomitant with searching out roots is the more difficult task of placing historical contributions in their proper context, to understand what might have been dictated by necessity, what might have been a limiting factor because of the then-dominant scientific paradigms, where there were true breaks with tradition and where not, what knowledge was built up incrementally through one or more trends and what accidents of popularity, influence, social pressure and the like might have led to one event rather than another. Consider, for example, why Broca focused on disorders of speech production (*langage articulé*): at least part of the reason is that before the 1860's, one did not talk about the <comprehension> of language. Language was *constructed* as speech output and the rest, what would have been or could have been discussed under the notion of comprehension, fell under the notion of <mind> which at the time was in the province of philosophy, religion and nascent psychiatry (the alienists). Since the 1820's the medical, and as well the phrenological, journals had been filled with case reports of expressive language impairments arising from stroke and trauma, with and without autopsy evidence of the involvement or lack of involvement of the frontal lobes [3]. The work of Bouillaud, Lallemand, Broussais, Dax and Lordat are the more familiar names, however there were dozens of obscure medical practitioners publishing these reports; Alexander Hood was one such lesser known researcher [4], about whom more will be said below. Beginning at least as early as 1866 with the publication by Theodor Meynert of a case of receptive aphasia with jargon [5], followed by Bastian's symmetric model of language input and output in 1869, and Schmidt's case of receptive aphasia in 1871, the stage was set for the *re-construction* of language --by the neuroscience community-- to include comprehension, seen in the work of Hughlings-Jackson and Wernicke starting

in 1874. Two exemplary models of historical analysis of the trends in the neurosciences as well as insightful expositions of the varying milieu are to be found in Harrington [6] and Clarke & Jacyna [7]; a third and monumental compendium of the origins of neuroscience from the earliest written records --nearly a third of the chapters bear directly upon neurolinguistic issues-- is Finger's recent text [8].

Prolegomena 2: pitfalls in historical research

Establishing priority is, in a word, fun. It is intellectually entertaining to learn that Roberts Bartholow, in 1874, was the first person to electrically stimulate the human brain with an electrode directly inserted into the cortex (when the electrode was pushed deeper into the cerebrum, possibly in the basal ganglia, possibly in the thalamus, most certainly sub-cortical, the subject of this experiment did cry out; otherwise, he did not elicit speech and did not test to see if his electrical stimulation interrupted speech), regardless of what one may think of the ethics of this experiment (he was publicly berated in a British medical journal) [9]. However, establishing priority is typically a very tricky enterprise. Consider the following quote from David Caplan [10, p. 46]: "The 1861 paper by Broca is the first truly scientific paper on language-brain relationships." Caplan supports this conclusion (here, as elsewhere, Caplan successfully integrates historical with contemporary research) with three claims, viz. that Broca presents a detailed case history with "excellent gross anatomical findings at autopsy", that Broca has the insight that the gross brain convolutions are constant anatomical features that may be related to particular psychological functions, and that Broca's primary conclusion that expressive speech depends upon a small part of the inferior frontal gyrus is a good first approximation which we generally accept today. Clearly, <priority> in this example is a matter of scholarly judgment. What then to make of the fact that Alexander Hood, in 1824, did a better job of analyzing expressive language functions and correlating them to frontal lobe anatomy? Hood had postulated a lexical-phonological level, a phonological-articulatory level and a motoric level for expressive speech, based upon the speech and language impairments which he observed in stroke patients. The oddity is, he used the phrenological model of Gall & Spurzheim [4, 11, 12]. What then to make of the fact that excellent clinical-pathological -- autopsy-- studies of aphasic cases may be found in the 17th century studies of Wepfer [13], studies that are so good one may verify the left hemisphere localization of language from them, or, the fact that Lallemand and Bouillaud in 1824 and 1825 published dozens of autopsy reports of patients with aphasia? What then to make of the fact that the classical neuroanatomists of the late 19th and early 20th century virtually abandoned the possibility of systematically describing gyral geography because of its evident variability, a variability that currently plagues PET researchers who need to co-register sites of PET activation with MRI images in order to cross-subject compare results? And, finally, what to make of the fact that the autopsy of Broca's 1861 patient actually demonstrated a very large left hemisphere lesion encompassing frontal, parietal and temporal cortex?

Broca "inferred" that the 3rd inferior frontal part of this large lesion was the one responsible for the patient's aphemia, by estimating the degree of necrosis and trying to back-correlate that with the patient's medical history. What is important to appreciate here is that it is not a question of disputing the facts but a question of how one chooses to interpret the historical record. A view which I prefer is (a) that Broca inherited a tradition of clinical-pathological correlation that already presupposed that different brain regions had different functions, (b) that Broca was theoretically constrained by a construct of language that placed psychological pre-eminence on speech production, (c) that Broca was immediately challenged and certainly intrigued by the debates (involving many famous members of the French scientific community, e.g. Gratiolet, Bouillaud, Auburtin, Flourens, et alia) concerning the role of the frontal lobe in speech and therefore was predisposed to see the age of that lesion as having a significant frontal component, and, most important of all, (d) that Broca had the position, power and prestige to take advantage of a serendipitous clinical observation.

Prolegomena 3: what not to do

Although some historical "facts" are subject to interpretation as we've just seen, some are just plain right or wrong, and, it behooves us to get it right. It is quite another matter, however, to commit the unpardonable historical sin of *presentism*. Consider the following quotation from John Morton [14, p. 40]: "We have a number of lessons to learn from history. If we are lucky we can avoid making the same mistakes as thinkers in the past." The mistake made by the "diagram-makers", according to Morton, was to confuse the goal of representing the elements of language processing in the brain with the goal of determining the localization in the brain of these elements. Needless to say, Morton assures us that "the same mistake will not be made again..." [14, p. 61], leaving this reader with the clear impression that his logogen model has, at last, revealed the truth about language (*veritatem patefacere* - Cicero). There is a fine line between science and religion and Morton's rhetorical style makes it hard to tell if the line has been crossed. That, however, is of less concern at present than the notion of an historical "mistake". Following the scientific paradigm of the day is simply not a mistake; to evaluate an earlier paradigm using the principles of one's own paradigm is *presentism* --to judge the past by today's standard. Most historians do not regard this as very productive. On the other hand, scientists do make mistakes, past and present company included, and some of the historical errors in brain-language relationships are quite interesting. Consider Franz Joseph Gall's localization of language functions (*sprachsinn und wortsinn*) in the anterior, inferior frontal lobe. The craniological method of relating skull protuberance (the "bumps") to hypertrophy of the underlying brain region, in turn due to above-normal development of the faculty which is expressed by that same brain region, is an unexceptional scientific method. We may find it humorous but it is a clear and falsifiable hypothesis. And in fact, one could argue that Flourens' experiments which demonstrated that animals whose cerebellums he had lesioned still exhibited copulatory behavior, which thus provided evidence against Gall's localization of the reproductive faculty in the cerebellum, was one of the principle reasons why many scientists rejected craniology, later phrenology; Gall refused to accept Flourens' evidence -- the scientific community, particularly Bouillaud, accepted it. On the other hand, Gall's argument that a well-developed language faculty, particularly verbal memory, would cause a protuberance of the inferior, anterior frontal lobe, which in turn would make the eye sockets shallow --thus, folks with high verbal skills were said to have "cow's eyes"-- was not successfully challenged by the scientific community. Rather, Bouillaud not only accepted *this* localization but championed it unceasingly right up to 1861 when Broca's publication seemed to vindicate Gall's model. What is curious is that it was well known at the time that the backside of the eye sockets do not abut the frontal lobe --a great deal of sinus cavity lies between the two. It is virtually impossible that a frontal brain bump could impinge upon the eye sockets. Evidently it was the accumulating evidence that frontal lesions typically led to speech disturbances, documented by Lallemand, Bouillaud and others from the 1820's on, that kept the phrenological language model alive until the great paradigm shift of the 1870's. Another error, not fully appreciated until recently, was committed by Lichtheim, one of the diagram-makers discussed in Morton's chapter. Laubstein [15] has elegantly shown that this "paradigmatic diagram-maker" had produced a neurolinguistic model that is ambiguous with respect to some predictions of language disorders, that fails to predict some language disorders that had already been described, that is internally inconsistent and, finally, that cannot be falsified, all in terms of the 19th century paradigm within which Lichtheim operated. This is the kind of analysis of the diagram makers diagrams that goes to the heart of the basic model-making assumption of that period and of our own: the correlation between aphasic language data and the components of the processing model of language used to account for such data.

Prolegomena 4: psychological vs neurological modeling

Having argued that what Morton says is the diagram-makers "mistake" should not be considered a mistake, let us examine the actual claim that Morton makes: did the diagram-makers confuse the psychological (processing elements) with the neurological (localization of elements)goals of their neurolinguistic enterprise, as Morton asserts? Baginsky (1871), the first diagram-maker discussed by Morton, believed he was basing his model on the "physiology of speech formation"; he did not stipulate specific anatomic sites for each of his language "centers", maintaining that "we do not yet have a precise conceptualization" of this relationship [16]. Kussmaul, the sixth diagram-maker discussed by Morton, claimed that his colleagues, particularly Wernicke, were mistaken in trying to localize the various speech centers to specific regions of the brain. Kussmaul was "acutely aware of the limitations of the localizationist approach to linguistic processes" [16, p. 509]; "extraordinarily removed from strictly anatomical and physiological considerations, Kussmaul the physician achieves an understanding

of the psychology of language in terms of the concepts which constitute the core of present day models, e.g. the distinction between various levels of representation ....their respective autonomy yet interconnection, and the notion of linguistic processes" [16, p. 497]. The same may be said of Elder [17] and Grasset [18] both mentioned by Morton, as well as of Bastian [19] who, though not discussed by Morton, was one of the best-known of the British diagram-makers of the period. As Paul Eling [19] remarked: "In general, characterizing the work of these classical aphasiologists with a few short statements and adjectives does not do justice to the careful analytic description and argumentation of these scientists." In fairness it ought to be noted that Morton's questionable analysis of the model-theoretic assumptions of Baginsky, Kussmaul, Elder, and Grasset, may not be entirely his fault; he relied on Moutier's 1908 dissertation as his secondary source material. Moutier was the student of Pierre Marie, notorious for his antipathy toward anyone who fractionated language into its component elements and thus anyone who believed that there were several different types of aphasia, obviously the main tenet of the diagram-makers. Ironically, and history sometimes has a penchant for the ironic, it was Pierre Marie who proposed that the insula (Island of Reil) was a functional component of expressive language (Marie's quadrilateral). George Ojemann and myself, about two decades ago, established that the insula can be language cortex (electrical stimulation of the insula elicited naming errors) and recent work by Nina Dronkers, using the lesion-overlap technique, suggests that insular lesions lead to apraxia of speech, a view quite consistent with Marie's view as Dronkers has noted. To return to the question of psychological Vs neurological modeling, Marie's fundamental objection to the localizationists/diagram-makers was a psychological one, viz. the dictum *l'aphasie est une.* In Marie's view the re-construction of language in the 1870's to include comprehension now became re-constructed again so that comprehension (understanding, the lexicon, etc.) now *was* language and speech production was relegated to the status of motoric output. To this day neurolinguistics has wrestled with the motor component of the expressive aphasias. In the 1960's and 1970's this was one of the major theoretical disputes between the Mayo School (Darley, Aronson, Brown et alia) and the Boston School (Geschwind, Goodglass, Benson et alia), a dispute which, with the benefit of two decades of hindsight, squarely addressed and never resolved the different demands of a psychological Vs a neurological model of language.

Prolegomena 5: gaps in the story (1600-1900)

As entertaining as it may be to learn that Pharonic medicine circa 3000 B.C. recognized temporal lobe injuries as leading to aphasia [8], the knowledge was not passed on to later cultures. Comments in the Hippocratic texts --which do have historical continuity with the present through the reintroduction of Greek texts via Arabic in the early Renaissance-- refer to what we would likely label dysarthria or aphasia and additionally to right-sided paralyses sometimes associated with speech disorders; it is debatable that these neurolinguistic observations were ever

systematically understood [20], although several recent historians have argued that the epilepsy commentaries indicate that the Hippocratic physicians did understand the connection between unilateral brain lesions and symptoms to the contralateral side of the body. From the period of Plato and Aristotle through the time of Galen and up to the Renaissance, many observations on the loss of speech and language associated with either intrinsic brain disease or traumatic injury, were written. However, based as they were for the most part upon theories of meningeal or ventricular function, these accounts differ substantially from our concepts of brain function, as is very well documented in O'Neill's scholarly text [20]. Benton & Joynt [21] pointed out that most of the "classical" aphasias had been described ("observed" would be more apt terminology) by 1800. O'Neill demonstrated that at least through the Renaissance (beginning of the 17th century) these observations were hardly part and parcel of any general, coherent theoretical model of brain-language relationships [20].

The dominant brain function model before the Renaissance was "ventricular theory", derived from Galen and elegantly modified by Descartes among others; basically, this was a model based on fluids and fluid flow for the obvious reason that thoughtful early scientists realized that something in the brain must move in order to be responsible for functions -- something passes from sense organs to effectors and the animal spirits were as good a candidate as any available. One marvels at medieval and early-Renaissance discussions of memory disorders following damage to the 4th ventricle which are models of clinical-pathological correlation despite the casual disregard of the actual neuroanatomy. However, by the end of the Renaissance period, ventricular theory had been disproven by, for example, the cases reported by Johannes Schenck (1530-1598) in 1584, cases with 4th ventricle damage in which memory was spared and cases with damage to the cortical substance in which the 4th ventricle was intact but memory was impaired [13]. Ynez O'Neill's summary of early neurolinguistics ends at the 17th century, leaving us with a number of gaps in the story from the Renaissance to the 20th century, gaps that are only partly filled in by current research on persons who have actually made substantial contributions to the development of neuropsychology and neurolinguistics.

Little has been written [13] about the 17th century brain scientist Johannes Jakob Wepfer (1620-1695); for our interests here, his posthumously published book, *Medical-practical observations of affections inside and outside the head* (1727) is most relevant. In it Wepfer discusses 13 well-described cases of aphasia, often noting *paralisys in dextri lateris, cum loquelae impedimentum* and yet never drawing the self-evident conclusion that left hemisphere lesions and right sided paralysis were associated. Perhaps the fever from which he died, overtook him before he completed his work; perhaps the reason for his silence on the matter of laterality was that his contemporaries, particularly those in the church, might have viewed such localization as too materialistic. Galileo was "processed" less than 30 years before

and the squares of Europe still smelled of the stakes of the Inquisition [13].

Although David Hartley (1705-1757), an early 18th century village doctor practicing without benefit of a medical degree, did not write on language nor did he study patients with brain damage, he was one of the first to explicitly propose a brain-based model of psychological functions [22] (Thomas Willis, a contemporary of Wepfer, had proposed the rudiments of such a model in the century before). His psychological theory was the associationism of Locke and was later destined to be the dominant neuropsychological model of the 19th century. His physiological theory was based on elements he called vibratiuncles (analogous to Willis' corpuscles which were in gentle vibration and directly borrowed from Isaac Newton [8]) which allowed him to account for the transmission of sensory images into the brain, motor operations out of the brain, attentional and memory mechanisms in the brain and, presaging neuroscientists of the 19th century, his vibration theory led him to a concept of domain specific localization of function.

Much has been written about Franz Joseph Gall's (1758-1828) contribution to neuroscience [3, 4, 6, 7, 8, 11, 12, 19 and references in these studies] but not a lot is known about the roots of his ideas. Christine Grou, in her unpublished doctoral dissertation, demonstrated a close parallel between the faculty psychology of Thomas Reid (1710-1796) and the faculties of Gall & Spurzheim and also a commonality between the many (hundreds of) physiognomic characteristics proposed by Johann Kaspar Lavater (1741-1801) and the phrenological faculties. Gall's idea that growth patterns of the cortex, i.e. hypertrophy or atrophy, would impress themselves on the inner table of the skull and thus be "readable" as bumps on the skull, was directly borrowed from Lavater. The great 18th century naturalist Charles Bonnet (1720-1792) proposed a vibration-based theory of memory reminiscent of Hartley; Bonnet also proposed a doctrine of localization of function in the brain that clearly had influenced Gall as the latter cites the former in several of his books. However, the details of Gall's indebtedness to these 18th century scientists remain to be elucidated. On the other end, we have worked out a few of the connections between craniology-phrenology and the development of neuropsychology in the period from 1820-1860 [3, 11] and we have also begun an analysis of how the early phrenologists helped to found the doctrine of clinico-pathological correlation of language impairments [4]; little is known about phrenology's contribution to other aspects of neuropsychology and psychiatry. Craniology-phrenology was quite clearly an early personality theory, cf. its roots in physiognomy; whether and in what respects it may have influenced the development of personality theory in modern psychology as well as psychiatry are not well worked out.

Historical analyses can help us realize that our neurolinguistic models (a) have precursors, (b) are contextually influenced by the scientific milieu and (c) are relative to the assumptions and constraints of the paradigms we happen to currently accept. And they can amuse.

References

[1] Geschwind, N. (1963), "Carl Wernicke, the Breslau school and the history of aphasia", in Carterette, E.C.[Ed.], *Brain Function, Vol III*, Berkeley: University of California

[2] Geschwind, N. (1965),"Disconnexion syndromes in animals and man", *Brain*, 88, pp. 237-294, 585-644.

[3] Whitaker, H.A. and Grou, C. (1993), "From craniology to neurology in 19th century France: how the localization of language became the test case", *Psychologie canadienne* 34.2a, p. 435.

[4] Whitaker, H.A. and Grou, C. (1991), "Spurzheim's legacy: the case of Adam M'Conochie (1824)", *Neurology* 41, 239

[5] Whitaker, H.A. and Etlinger, S.C. (1993), "Theodor Meynert's contribution to classical 19th century aphasia studies", *Brain and Language* 45.4, pp. 560-571.

[6] Harrington, A. (1987), *Medicine, Mind and the Double Brain*, Princeton: Princeton University Press

[7] Clarke, E. and Jacyna, L.S. (1987), *Nineteenth-Century Origins of Neuroscientific Concepts*, Berkeley: University of California

[8] Finger, S. (1994), *Origins of Neuroscience*, New York: Oxford UP

[9] Whitaker, H.A. and Ojemann, G.A. (in preparation), "The early history of electrical stimulation of the human brain: from Bartholow (1874)to Penfield (1928)

[10] Caplan, D. (1987), *Neurolinguistics and Linguistic Aphasiology*, Cambridge: Cambridge University Press.

[11] Grou, C. et Whitaker, H.A. (1992), "Le Cerveau: Petite histoire de la localisation des fonctions", *Interface* 13.5, pp. 14-21.

[12] Zola-Morgan, S. (1995), "Localization of brain function: the legacy of Franz Joseph Gall (1758-1828)", *Ann Review of Neuroscience*, 18, 359-383

[13] Luzzatti, C. and Whitaker, H.A. (in press), "Johannes Schenck and Johannes Jakob Wepfer: Clinical and Anatomical Observations in the Prehistory of Aphasia and Cognitive Disorders", *Journal of Neurolinguistics*

[14] Morton, J. (1984), "Brain-based and non-brain-based models of language", in D.Caplan, A.R.Lecours & A.Smith [Eds.], *Biological Perspectives on Language*, Cambridge: MIT Press. pp 40-64.

[15] Laubstein, A.S. (1993), "Inconsistency and ambiguity in Lichtheim's model", *Brain and Language* 45.4, pp. 588-603.

[16] Jarema, G. (1993), "In sensu non in situ: the prodromic cognitivism of Kussmaul", *Brain and Language* 45.4, pp. 495-510.

[17] Whitaker, H.A. (1988), "William Elder (1864-1931): Diagram Maker and Experimentalist, in L.Hyman and C.Li [Eds.], *Language, Speech and Mind*, London: Routledge. pp 163-174.

[18] Dos Santos, G., Nespoulous, J-L. and Whitaker, H.A. (in preparation) "Grasset's Polygon"

[19] Eling, P. [Ed.] (1994), *Reader in the History of Aphasia*, Amsterdam: John Benjamins.

[20] O'Neill, Y. (1980), *Speech and Speech Disorders in Western Thought Before 1600*, Westport: Greenwood Pr

[21] Benton, A.L. and Joynt, R. (1960), "Early descriptions of aphasia", *Archives of Neurology* 3, pp 205-221.

[22] Aubert, D. and Whitaker, H.A. (in preparation) "David Hartley's model of vibratiuncles seen as a contribution to the localization theory of brain function"

ELECTROPHYSIOLOGICAL INSIGHTS INTO LANGUAGE AND SPEECH PROCESSING

P. Hagoort and C.M. Brown
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

ABSTRACT

Event related brain potentials (ERPs) have been used to study language and speech processing. Two distinct ERP-effects will be discussed: (1) The *N400-effect*. This effect is related to the integration of word meaning into an utterance level representation; (2) The *Syntactic Positive Shift* (SPS). The SPS is related to syntactic processing. Both N400 and SPS were originally observed in reading, but they can also be observed with speech, with some changes in their latency and distribution.

EVENT RELATED BRAIN POTENTIALS

Cognitive electrophysiology provides a record of various perceptual and cognitive processes as they unfold in real time. The basis for this record are the voltage fluctuations recorded with the help of electrodes placed on the scalp, known as the electro-encephalogram (EEG).

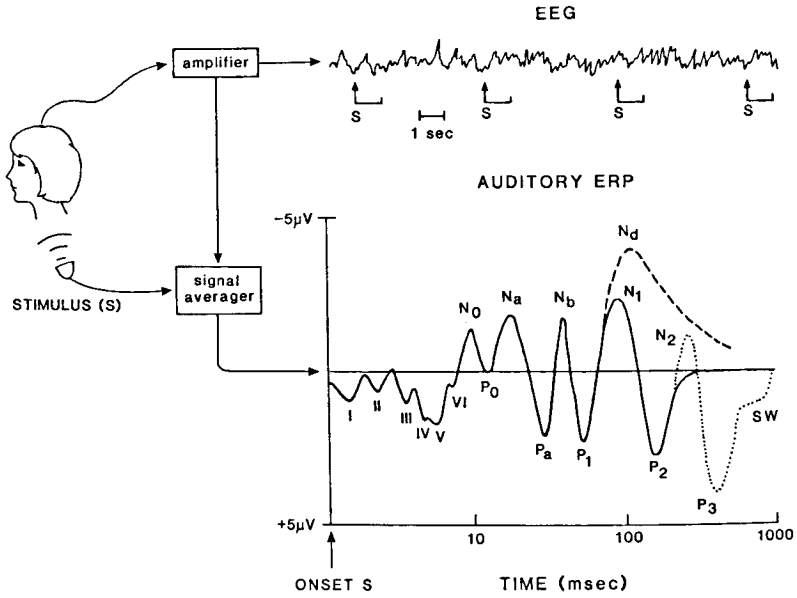


Figure 1 (after [1]): Idealized waveform of a series of ERP components that become visible after averaging the EEG to repeated presentations of a short auditory stimulus. Usually, averaging over a number of stimulus tokens is required to get an adequate signal-to-noise ratio. Along the logarithmic time axis the early brainstem potentials (Waves I-VI), the midlatency components (No, Po, Na, Pa, Nb), the largely exogenous components (P1, N1, P2), and the endogenous, cognitive ERP components (Nd, N2, P300, Slow Wave) are shown. The components with a negative polarity are plotted upwards, the components with a positive polarity are plotted downwards.

Under the appropriate stimulation conditions, one can derive so-called event related brain potentials (ERPs) from the EEG. Scalp-recorded ERPs reflect the summation of synchronous post-synaptic activity of a large number of neurons. ERPs differ from background EEG in that they reflect brain electrical activity time-locked to particular stimulus events. Establishing a reliable ERP trace normally requires averaging over a series of ERP recordings to tokens of the same stimulus type. The resulting average waveform typically includes a number of positive and negative peaks, often referred to as ERP-components (see Figure 1). Usually, the peaks in the ERP waveform are labelled according to their polarity (*N* for negative, *P* for positive) and their average latency in milliseconds relative to the onset of stimulus presentation (e.g., *N400*, *P300*). In some cases, the ERP peaks get a functionally defined label (*SPS* for syntactic positive shift; *ERN* for error-related negativity). ERPs are recorded from a number of electrodes distributed over the scalp. Often they have a characteristic distribution, showing larger amplitudes at some sites than at others. These distributional characteristics can be helpful in identifying a certain component.

For the purposes of psycholinguistically oriented ERP research, the most informative ERPs belong to the class of so-called "endogenous" components. Endogenous components are relatively insensitive to variations in physical stimulus parameters (e.g., size, intensity), but highly responsive to the cognitive processing consequences of the stimulus events. The modulations in amplitude or latency of an endogenous ERP as a consequence of some experimental manipulation, usually form the basis for making inferences about the nature of the underlying cognitive processing events.

For research on language and speech processing, particular two characteristics of ERPs are of relevance. The first is the *multidimensional* nature of the ERP waveform. ERPs can vary along a number of dimensions: specifically, the latency at which an ERP component occurs relative to simulus onset, its polarity, its amplitude, and its amplitude distribution over the recording sites. On the basis of these characteristics it is reasonable to assume that different types of ERP peaks (e.g. positive peaks vs. negative peaks) are generated by, at least in part, non-overlapping neuronal populations. Insofar as the involvement of different neuronal ensembles implies qualitatively different processing events, in principle these processing events can show up as qualitatively different in the overall ERP waveform. This characteristic makes ERPs a useful addition to the recording of unidimensional measures, such as reaction times. For instance, if in sentence processing the electro-physiological signatures of semantic integration processes and parsing operations turn out to be qualitatively different, ERPs might provide us with a crucial tool for testing how and at what moments in time the process of assigning a structure to the incoming string of words, and interpreting this string semantically, influence each other.

The second important characteristic of ERPs is that they provide a *continuous, real-time* measure. This high temporal resolution of ERPs is unmatched by other brain imaging techniques such as PET and fMRI. Like speeded reaction-time (RT) measures in the more classical psycholinguistic tasks, such as naming, lexical decision, and word or phoneme monitoring, ERPs are tightly linked to the temporal organization of ongoing language processing events. But in

contrast to RT measures, ERPs provide a continuous record throughout the total processing epoch and beyond. Therefore, it is possible to monitor not only the immediate consequences of a particular experimental manipulation (e.g., a syntactic or semantic violation), but also its processing consequences further downstream. This feature enabled us to show that the impossibility of assigning the preferred structure to an incoming string of words has consequences for lexical-semantic integration processes further downstream in the sentence (see below) [2].

N400-EFFECTS

The N400 was first reported in a paper by Kutas & Hillyard (1980) [3]. These authors presented subjects with a variety of sentences either ending in a word that was semantically congruous with the sentence context (e.g., "He shaved off his mustache and beard") or ending in a semantic anomaly ("I take coffee with cream and dog"). The semantically anomalous words elicited a negative component with a centro-parietal maximum on the scalp, and a latency that peaked around 400 ms. This component has since become known as the N400, and the difference between the N400 amplitude in the experimental and the control conditions has become known as the N400 effect.

Since its discovery, it has become clear that N400 effects are not elicited by only semantic *violations*. This can be illustrated by the following result from one of our studies [4]. We presented subjects with sentences that were identical, with the exception of a highly expected word in sentence-medial position (e.g., "Jenny put the sweet in her *mouth* after the lesson) versus a word that made perfect sense but was less expected in this position (e.g., "Jenny put the sweet in her *pocket* after the lesson"). Figure 2 shows the ERP waveforms to the more and less expected words, preceded and followed by one word. As can be seen in this figure, the N400 to the less expected word 'pocket' is larger than to the word 'mouth' which is the more expected continuation of the context. This probably reflects the different degree to which these words can be readily integrated within the higher order representation of their preceding sentential-semantic context.

Across many N400 studies, the following general characteristics are known to hold for the N400: (a) Each open-class word elicits an N400. (b) The amplitude of the N400 is inversely related to the cloze probability of a word in sentence context. The better the semantic fit between a word and its context, the more reduced the amplitude of the N400. (c) The amplitude of the N400 varies with word position, such that the first content word in a sentence produces a larger negativity than content words in later positions. This amplitude reduction is most likely due to the increasing semantic constraints throughout the sentence. (d) N400 effects are obtained in sign language, but not with violations of contextual constraints in music.

Importantly, N400 effects are not only observed with visual language input, but also with speech input. The most important difference between N400 effects to written and spoken words is their onset latency. Whereas the N400 effect in the visual modality usually onsets at about 250 ms, with spoken input the onset can be up to 200 ms earlier. This means that on average, N400 effects to speech start to emerge well before the end of a word.

Recent research suggests that the amplitude of the N400 is related to lexical-semantic integration processes [5]. That is, once a word has been accessed in the mental lexicon, its meaning has to be integrated into an overall representation of the current

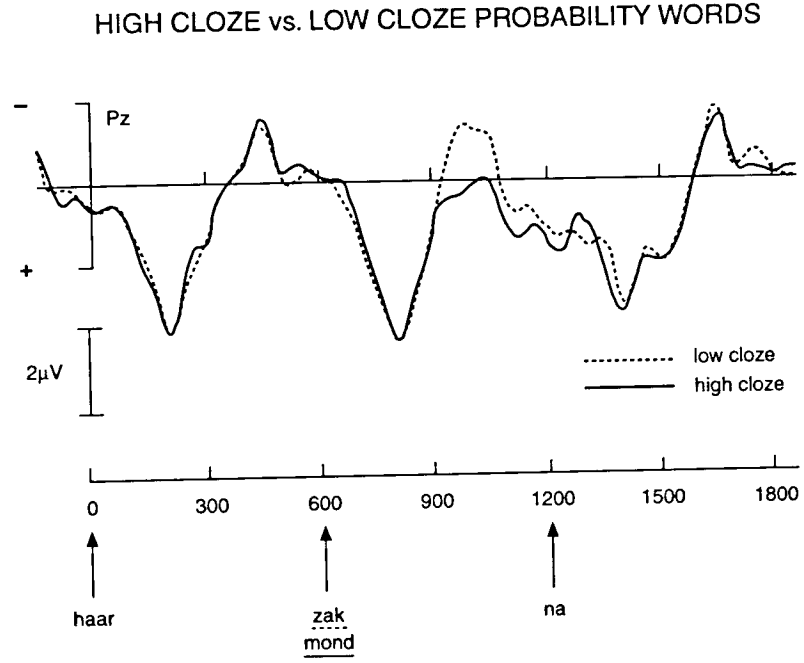


Figure 2: Grand-average waveform for electrode site Pz, for sentence-medial words with a high cloze probability and a low cloze probability. Sentences were presented word by word on the center of a computer screen at a rate of one word per 600 ms. The cloze target is preceded and followed by one word. The translation of the Dutch example sentence is "Jenny put the sweet in her pocket/mouth after the lesson." The waveforms represent the part of the sentence that is underlined.

word or sentence context. The easier this integration process is, the smaller the a amplitude of the N400 becomes.

The early onset of N400 effects in speech attests to the immediacy of lexical-semantic integration processes in this modality.

THE SYNTACTIC POSITIVE SHIFT

In recent years a number of ERP studies on syntactic processing have clearly shown that the ERP responses to violations of syntactic preferences are qualitatively different from the classical N400 effect [2][6].

In one of our studies, we had subjects read sentences that violated the agreement between the subject nounphrase and the finite verb, as in the following example sentences (literal translation in English between brackets; the word that renders the sentence ungrammatical (the Critical Word [CW] and its counterpart are italicized):

"Het verwende kind *gooit* het speelgoed op de grond."
(The spoiled child *throws* the toys on the floor.)
* "Het verwende kind *gooien* het speelgoed op de grond."
(The spoiled child *throw* the toys on the floor.)

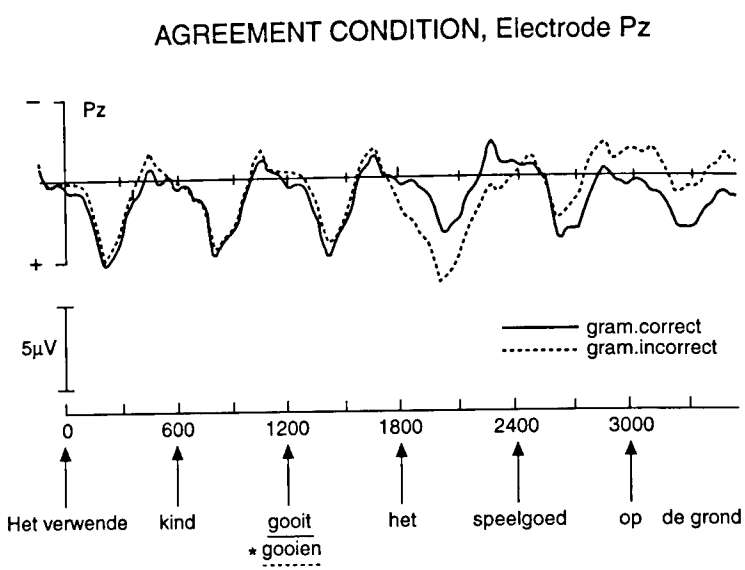


Figure 3: Grand-average waveform for electrode site Pz, for the grammatically correct and incorrect Critical Words (CW). The CW is preceded by two and followed by three words. Sentences were presented word by word on the center of a computer screen at a rate of one word per 600 ms. The translation of the example sentence is "The spoilt child throws/throw the toy on the ground."

The basic pattern of results that we observed is shown in Figure 3 for a posterior midline electrode (Pz). The CW is preceded by two words and followed by three words.

As can be seen, the ERP waveform to the incorrect CW shows a positive shift in comparison with its correct counterpart. This positive shift is widely distributed over the recording sites and has a centro-parietal maximum. Based on its sensitivity to syntactic aspects of a sentence, we have labeled this effect the SPS (i.e., Syntactic Positive Shift). The onset of the SPS is at about 500 ms after presentation of the incorrect CW. A similar pattern of results is obtained for a number of other syntactic violations in both Dutch and English.

As can be seen, the SPS is replaced by a negative shift on word positions following the CW. These are N400 effects, indicating the increased difficulty of integrating words into the sentence context following a syntactic violation.

As holds for the N400, the SPS is not elicited by only syntactic violations. In general, an SPS can be observed when a syntactic preference can no longer be maintained. That is, the word in the sentence that renders the preferred syntactic structure impossible, elicits an SPS. An example in case are so-called syntactically ambiguous sentences. Very often part of a sentence can be assigned more than one syntactic structure. For instance, in the utterance "The pope greets the priest and the monk...", the noun 'monk' can go together with 'priest' to form the object of the sentences (e.g., "The pope greets the priest and the monk at the annual meeting"). Alternatively it can start a new clause (e.g., "The pope greets the priest and the monk welcomes the cardinal"). For reasons of processing economy, the first (conjoined-NP) reading is preferred over the second (Sentence conjunction) reading. The verb that renders the preferred syntactic assignment impossible ('welcomes' in the example), therefore, elicits an SPS. In short, the SPS seems to signal that the initially assigned syntactic structure can no longer be maintained, and that some form of reanalysis has to be initiated.

To date, only two studies have tested for the occurrence of an SPS to syntactic violations in spoken sentences. Osterhout and Holcomb [7] report a somewhat earlier onset of the SPS during the perception of continuous speech. In our own study, however, the onset of the SPS in continuous speech was quite similar to that in the visual modality. Although more research needs to be done with continuous speech input, current results tentatively suggest that the signal for reanalysis is relatively insensitive to the rate at which words are presented.

CONCLUSIONS

From these results the following general conclusions can be drawn:

(1) Electrophysiological recordings provide a real-time neurophysiological measure of language and speech processing with a temporal resolution that is far superior in comparison with other brain imaging techniques such as PET and fMRI. These latter methods, however, have a much better spatial resolution than ERPs. For a full understanding of the neurobiological basis of language and speech, we have to rely on the combined use of different brain imaging techniques.

(2) The existence of different ERP responses to aspects of semantic and syntactic processing suggests different underlying brain states for semantic integration and parsing. To the degree to which the SPS and the N400 individuate different sets of neural generators, and to the degree to which these different sets of neural generators (directly or indirectly) correspond to different cognitive states, it can be concluded that the processing mechanism for the computation of syntactic structures is different from that for the computation of the meaning of an utterance. In other words, the brain honours the distinction between syntax and semantics.

(3) The observed ERP effects are independent of modality. That is, both with written and spoken input N400 effects and SPS are observed. However, the effects seem to be earlier in continuous speech, especially for the N400. This attests to the speed at which speech has to be processed.

(4) The findings of qualitatively different neurophysiological responses to semantic and syntactic processing, suggests that in further research additional ERP effects sensitive to, for instance, early phonological processing might be obtained. Some recent findings are suggestive of this possibility [8][9].

ACKNOWLEDGEMENT

The research reported in this paper was supported by a grant from the Dutch Science Foundation (NWO), with grant number 400-56-384.

REFERENCES

[1] Hillyard, S.A., & Kutas, M. (1983), "Electrophysiology of cognitive processing." *Ann. Rev. Psychol.*, vol. 34, pp. 33-61.

[2] Hagoort, P., Brown, C.M., & Groothusen, J. (1993), "The syntactic positive shift (SPS) as an ERP-measure of syntactic processing." *Language and Cognitive Processes*, vol. 8, pp. 439-483.

[3] Kutas, M., & Hillyard, S.A. (1980), "Reading senseless sentences: Brain

potentials reflect semantic incongruity." *Science*, vol. 207, pp. 203-205.
[4] Hagoort, P., & Brown, C.M. (1994), "Brain responses to lexical ambiguity resolution and parsing" In: Ch. Clifton Jr., L. Frazier, & K. Rayner (Eds.), *Perspectives on sentence processing.* Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 45-80.
[5] Brown, C.M., & Hagoort, P. (1993), "The processing nature of the N400: Evidence from masked priming." *Journal of Cognitive Neuroscience*, vol. 5, pp. 34-44.
[6] Osterhout, L., & Holcomb, P.J. (1992), "Event-related brain potentials elicited by syntactic anomaly." *Journal of Memory and Language*, vol. 31, pp. 785-806.
[7] Osterhout, L., & Holcomb, P.J. (1993), "Event-related potentials and syntactic anomaly: Evidence of anomaly detection during the perception of continuous speech." *Language and Cognitive Processes*, vol. 8, pp. 413-437.
[8] O. Conolly J.F., & Phillips, N.A. (1994), "Event-related potential components reflect phonological and semantic processing of the terminal words of spoken sentences." *Journal of Cognitive Neuroscience*, vol. 6, pp. 256-266.
[9] Praamstra, P., Meyer, A.S., Levelt, W.J.M. (1994), "Neurophysiological manifestations of phonological processing: Latency variation of a negative ERP component timelocked to phonological mismatch." *Journal of Cognitive Neuroscience*, vol. 6, pp. 204-219.

ON THE NEUROBIOLOGY OF THE SOUND STRUCTURE OF LANGUAGE: EVIDENCE FROM APHASIA

S.E. Blumstein

Brown University, Providence, RI 02912, USA

ABSTRACT

The patterns of both speech production and speech perception deficits in aphasia suggest that the disorders reflect impairments to the processes involved in accessing the sound structure rather than selective impairments to the sound properties of speech or to their representations. Speech production deficits occur at both the phonological level, reflecting selection or access impairments, as well as at the phonetic level, reflecting articulatory implementation impairments. Phonological deficits emerge regardless of lesion site, whereas phonetic deficits emerge with damage to specific neuroanatomical structures. Speech perception impairments reflect misperceptions of phonetic features rather than deficits in extracting the acoustic patterns associated with these features. Such impairments emerge particularly as the sound properties of speech contact the lexicon.

One of the most challenging issues in the study of the neurobiology of speech is understanding the neural basis of speech production and speech perception mechanisms. This domain of inquiry has largely focused on investigations of adult aphasic patients, exploring the clinical and behavioral manifestations of their disorder and the accompanying lesion localization. Just as the study of the sound structure of language has been guided by considerations of the structural properties of the speech-language system in normals, both phonological and phonetic, so has the study of speech production and speech perception deficits in aphasia. Two major questions have shaped research in the field. The first centers on whether speech production and speech perception deficits reflect selective impairments to the sound properties of speech and their representations, or alternatively, impairments to the processes involved in accessing these representations. The second centers on whether speech production and speech perception impairments reflect deficits that are primarily phonological in nature, affecting the structural properties of language, or whether they are phonetic in nature, affecting, on the one hand, articulatory implementation in speech production, or on the other hand, acoustic decoding in speech perception. It is these two issues which are the primary focus of this paper.

SPEECH PRODUCTION

In order to produce a word or an utterance, the speaker must select the word candidate(s) from the lexicon including its phonological form (selection), and then encode the abstract phonological representation of the word in terms of the articulatory parameters required for realizing the phonological properties in the particular context in which they appear (articulatory planning). Subsequent to the selection of a lexical candidate or candidates and the articulatory planning of the utterance, the phonetic string is ultimately converted into a set of motor commands or motor programs to the articulatory system. This set of 'instructions' to the articulators relates to the phonetic implementation of speech. The final speech output must conform to the phonological rules of the language including the correct production of the sound segments in their phonetic environment, the appropriate stress pattern of the word, and in larger contexts the appropriate prosodic structure of the utterance including both stress and intonation.

Linguistic theory makes a distinction between phonology and phonetics, and the facts of aphasia support such a distinction. In some cases, patients may produce a wrong sound segment, but its phonetic implementation is correct, i.e. for 'teams' the patient says 'keams'. In other cases, patients may produce the correct sound segment but its phonetic implementation is distorted, i.e. for 'teams' the patient produces an initial /t/ that is overly aspirated.

Nearly all aphasic patients, regardless of the aphasia syndrome and underlying neuropathology, display speech production impairments that implicate a deficit at the phonological level. The patterns of impairment are similar across patients, suggesting that a common mechanism is impaired. The mechanism relating to this phonological impairment most likely relates to the selection and/or organization of the features comprising the candidate lexical entries.

The evidence for this comes from investigations of the patterns of speech production errors produced by aphasic patients [1]. For example, most sound substitution errors, e.g. 'teams' -> 'keams', involve a change in value of a single phonetic feature. This pattern of errors is consistent with the view that the incorrect phonetic feature has been selected or activated, but has been correctly implemented by the articulatory system. It also supports the view that phonetic features are organized in terms of tiers. Tiers have been defined in phonological theory to reflect the fact that phonetic features correspond to independent articulatory gestures (and consequently acoustic events) such as tongue placement and movement, lip movement, laryngeal activity, and height of the velum. Phoneme substitution errors in aphasia rarely involve more than one tier at a time, with feature changes typically relating to place of articulation, e.g. 'teams' -> 'keams', voicing, e.g. 'toy' -> 'doy', nasality, e.g. 'nut' -> 'dut', and manner of articulation, e.g. 'sun' -> 'tun' [2].

Phonological errors also suggest that the nature of the syllable structure of the lexical candidate constrains the type and extent of errors made during the selection process [1,2]. Phoneme substitution errors are more likely to occur in singleton consonants than in clusters, e.g. [f] is more likely to undergo a phoneme substitution error in the word 'feet' than in 'fleet'. Simplification and addition errors are more likely to result in the canonical syllable structure, CV, e.g. a consonant is more likely to be deleted in a cluster, 'sky' -> 'ky', and is more likely to be added in a word beginning with a vowel, 'army' -> 'jarmy'. Finally, assimilation errors across word boundaries preserve the syllable structure relations of the lexical candidates, e.g. 'history books' -> 'bistory books' and 'roast beef' -> 'roaf beef'. These results show that the syllable structure of a word is part of its lexical representation, and this information is used in the planning buffer for sentence production. If this were not the case, the syllable constraints shown in the assimilation errors would not occur across word boundaries.

While phonological patterns are similar across aphasic patients, phonetic deficits seem to be more selective. A long-held observation is that aphasic patients with anterior brain-damage produce phonetic errors. The implied basis for these errors is one of articulatory implementation: that is, commands to the articulators to encode words are incorrect, poorly timed, and so forth. A number of studies have explored these phonetic patterns of speech by investigating the acoustic properties or the articulatory parameters underlying the production of particular phonetic dimensions. These studies have shown that anterior patients have difficulty producing phonetic dimensions that require the timing of two independent articulators, e.g. voicing (i.e. voice-onset time) and nasality (i.e. the timing of the release of the closure in the oral cavity and the velum opening) [3,4,5,6]. In particular, there is considerable overlap between the target productions in the area of the phonetic boundary. Similar patterns emerge across different languages, occurring not only in English and Japanese for which voice-

onset time serves to distinguish two categories of voicing - voiced and voiceless - but also in Thai for which voice-onset time serves to distinguish three categories of voicing in stop consonants - pre-voiced, voiced, and voiceless aspirated.

That the phonetic output disorder of these patients likely reflects an articulatory implementation deficit rather than a failure to encode appropriately the phonetic feature such as voicing comes from acoustic analyses of the production of vowel length as a cue to final stop consonant voicing. Results indicate that while anterior aphasic patients show an impairment in the implementation of the phonetic dimension of voicing using voice-onset time, they maintain the distinction between voiced and voiceless stops on the basis of the duration of the preceding vowel [7,8].

Kent and Rosenbek [9] have suggested that the timing problem found for individual segments and their underlying features is a manifestation of a broader impairment in the integration of articulatory movements from one phonetic segment to another. Nonetheless, investigations of coarticulation effects in anterior aphasics show that they produce relatively normal anticipatory coarticulation [10]. For example, in producing the syllable [su], anterior aphasics anticipate the rounded vowel [u] in the production of the preceding [s]. Nonetheless, they may show a delay in the time it takes to produce these effects, and they may show some deficiencies in their productions [11, 12]. These results suggest that phonological planning is relatively intact, but the timing or coordination of the implementation of the articulatory movements is impaired.

While still premature, results exploring the neuroanatomical basis of these phonetic patterns of speech suggest the involvement of specific neuroanatomical substrates. These areas include Broca's area, the lower motor cortex regions for larynx, tongue, and face, and some white matter structures as well [7].

Several conclusions can be made concerning the nature of the phonetic disorders and their ultimate underlying mechanisms. The impairment is selective for patients with specific underlying neuropathology. The deficit is not a linguistic one affecting the implementation of a particular phonetic feature. Moreover, the patients have not lost the representation for implementation nor the knowledge base for how to implement sounds in context, but rather particular maneuvers relating to timing of articulators seem to be impaired.

Interestingly, posterior patients display a subtle phonetic deficit showing increased variability in the implementation of a number of phonetic parameters including vowel formant frequencies [13] and vowel duration [12, 13, 14]. Because these phonetic impairments are not clinically perceptible but emerge only upon acoustic analysis, they are considered to be subclinical. These results suggest that both anterior and posterior brain structures ultimately contribute to the speech production process.

SPEECH PERCEPTION

Current views of auditory language comprehension and specifically the auditory perception of words suggest that contact with the lexicon (and ultimately meaning) requires the encoding of the auditory input into a spectral representation based on the extraction of more generalized auditory patterns of properties from the acoustic waveform, the conversion of this spectral representation to a more abstract feature/phonological representation, and then the selection of a word candidate from a set of potential word candidates sharing phonological properties with the target word.

Most studies exploring the role of speech perception deficits in auditory comprehension impairments have focused on the ability of aphasic patients to perceive phonemic or segmental contrasts. Results show that nearly all aphasic patients show some problems in discriminating phonological contrasts [15, 16, 17] or in labeling consonants presented in a consonant-vowel context [18, 19]. The overall patterns of performance are similar across patients and essentially mirror the patterns found in the analysis of phonological errors in speech production. Namely, subjects are more likely to make perception errors when the test stimuli contrast by a single phonetic feature than when they contrast by two or more features, and the perception of place of articulation is particularly vulnerable [15, 17, 20].

What is not clear from these studies is whether the failure to perceive segmental contrasts reflects an impairment in the perception of phonetic features or alternatively an impairment relating to the extraction of the acoustic patterns associated with these features. To investigate this issue, several studies have explored categorical perception (both labeling and discrimination) of the acoustic parameters associated with voicing [18, 19, 20,21] and place of articulation in stop consonants [22]. Results showed that if aphasic patients could successfully complete one of the two labeling or discrimination tasks, it was discrimination. Most importantly, the shape of the discrimination functions and the locus of the phonetic boundary were comparable to those of normals, even in those patients who could only discriminate the stimuli. The fact that no perceptual shifts were obtained for the discrimination and labeling functions compared to normals, and that the discrimination functions remained stable even in those patients who could not label the stimuli suggests that aphasic patients do not have a deficit specific to the extraction of the acoustic patterns corresponding to the phonetic categories of speech. Rather, their deficit seems to relate to the threshold of activation of the phonetic/phonological representation itself or to its ultimate contact with the lexicon.

Several recent studies have suggested that speech perception problems manifest themselves most strongly when the sound properties of speech contact the lexicon. For example, nonwords, e.g. 'gat', phonologically related to real words, e.g. 'cat', do not seem to access the lexicon in Broca's aphasics as they do in normals [23]. In contrast, for Wernicke's aphasics, such nonwords seem to activate the lexicon more so than they do in normals. Similarly, the lexical status of a word affects differentially how aphasic patients perform phonetic categorization. Broca's aphasics show a larger than normal lexical effect, seeming to place a greater reliance on the lexical status of the stimulus in making their phonetic decisions than on the perceptual information in the stimulus. In contrast, Wernicke's aphasics do not show a lexical effect, suggesting that top-down information does not influence phonetic categorization, and may even fail to guide their language performance [24].

Overall, the findings from speech perception studies of aphasic patients suggest that the neural basis for speech reception is broadly represented, and includes far greater neural involvement than the primary auditory areas and auditory association areas in the temporal lobe. In fact, anterior as well as posterior brain structures are implicated in the auditory processing of speech. Although the number of neurophysiological and electrophysiological studies focusing particularly on speech reception are few, they provide converging evidence consistent with this view [25, 26, 27, 28, 29].

ACKNOWLEDGMENTS

This research was supported in part by NIH Grant DC00314 to Brown University.

REFERENCES

[1] Blumstein, S.E. (1973), *A Phonological Investigation of Aphasic Speech*, The Hague: Mouton.

[2] Blumstein, S.E. (1990), Phonological deficits in aphasia: Theoretical perspectives, In A. Caramazza (Ed.), *Neurolinguistics: Advances in Models of Cognitive Function andCognitive Neuropsychology and Impairment*, Hillsdale: Lawrence Erlbaum.

[3] Blumstein, S.E., Cooper, W.E., Goodglass, H., Statlender, S., and Gottlieb, J. (1980), Production deficits in aphasia: A voice-onset time analysis, *Brain and Language, 9*, 153-170.

[4] Gandour, J. and Dardarananda, R. (1984), Voice-onset time in aphasia: Thai, II: Production, *Brain and Language, 18*, 389-410.

[5] Itoh, M., Sasanuma, S., Tatsumi, I., Murakami, S., Fukusako, Y., and Suzuki, T. (1982), Voice onset time characteristics

in apraxia of speech, *Brain and Language, 17,* 193-210.
[6] Shewan, C.M., Leeper, H., and Booth, J. (1984), An analysis of voice onset time (VOT) in aphasic and normal subjects, In J. Rosenbek, M. McNeill, and A. Aronson (Eds.), *Apraxia of Speech,* San Diego, Ca.:College-Hill Press.
[7] Baum, S.R., Blumstein, S.E., Naeser, M.A., and Palumbo, C.L. (1990), Temporal dimensions of consonant and vowel production: An acoustic and CT scan analysis of aphasic speech, *Brain and Language, 39,* 33-56.
[8] Duffy, J. and Gawle, C. (1984), Apraxic speakers' vowel duration in consonant-vowel-consonant syllables, In J. Rosenbek, M. McNeil, and A. Aronson (Eds.), *Apraxia of Speech,* San Diego, Ca.: College-Hill Press.
[9] Kent, R. and Rosenbek, J. (1983), Acoustic patterns of apraxia of speech, *Journal of Speech and Hearing Research, 26,* 231-248.
[10] Katz, W.F. (1988), Anticipatory coarticulation in aphasia: Acoustic and perceptual data, *Brain and Language, 35,* 340-368.
[11] Ziegler, W., and von Cramon, D. (1985), Anticipatory coarticulation in a patient with apraxia of speech, *Brain and Language, 26,* 117-130.
[12] Tuller, B. (1984), On categorizing aphasic speech errors. *Neuropsychologia, 22,* 547-557.
[13] Ryalls, J. (1986), An acoustic study of vowel production in aphasia, *Brain and Language, 29,* 48-67.
[14] Gandour, J., Ponglorpisit, S., Khunadorn, F., Dechongkit, S., Boongird, P., and Boonklam, R. (1992), Timing characteristics of speech after brain damage: Vowel length in Thai, *Brain and Language, 42,* 337-345.
[15] Blumstein, S.E., Baker, E., and Goodglass, H. (1977), Phonological factors in auditory comprehension in aphasia, *Neuropsychologia, 15,* 19-30.
[16] Jauhiainen, T., and Nuutila, A. (1977), Auditory perception of speech and speech sounds in recent and recovered aphasia, *Brain and Language, 4,* 572-579.
[17] Miceli, G., Caltagirone, C., Gainotti, G., and Payer-Rigo, P. (1978), Discrimination of voice versus place contrasts in aphasia, *Brain and Language, 2,* 434-450.
[18] Basso, A., Casati, G., and Vignolo, L.A. (1977), Phonemic identification defects in aphasia, *Cortex, 13,* 84-95.
[19] Blumstein, S.E., Cooper, W.E., Zurif, E.B., and Caramazza, A. (1977), The perception and production of voice-onset time in aphasia, *Neuropsychologia, 15,* 371-383.
[20] Baker, E., Blumstein, S.E., and Goodglass, H. (1981), Interaction between phonological and semantic factors in auditory comprehension, *Neuropsychologia, 19,* 1-16.
[21] Gandour, J. and Dardarananda, R. (1982), Voice onset time in aphasia: Thai, I. Perception, *Brain and Language, 17,* 24-33.
[22] Blumstein, S.E., Tartter, V.C., Nigro, G., and Statlender, S. (1984), Acoustic cues for the perception of place of articulation in aphasia, *Brain and Language, 22,* 128-149.
[23] Milberg, W., Blumstein, S.E., and Dworetzky, B. (1988), Phonological processing and lexical access in aphasia, *Brain and Language, 34,* 279-293.
[24] Blumstein, S.E., Burton, M., Baum, S., Waldstein, R., and Katz, D. (1993), The role of lexical status on the phonetic categorization of speech in aphasia, *Brain and Language, 46,* 181-197.
[25] Lauter, J., Herscovitch, P., Formby, C., and Raichle, M.E. (1985), Tonotopic organization in human auditory cortex revealed by positron emission tomography, *Hearing Research, 20,* 199-205. 1985.
[26] Petersen, S.E., Fox, P.T., Posner, M.I., Mintun, M., and Raichle, M.E. (1988), Positron emission tomographic studies of the cortical anatomy of single-word processing, *Nature, 331,* 585-589.
[27] Petersen, S.E., Fox, P.T., Posner, M.I., Mintun, M., and Raichle, M.E. (1989), Positron emission tomographic studies of the processing of single words, *Journal of Cognitive Neuroscience, 1,* 153-170.
[28] Zatorre, R.J., Evans, A.C., Meyer, E., and Gjedde, A. (1992), Lateralization of phonetic and pitch discrimination in speech processing, *Science, 256,* 846-849.

[29] Ojemann, G.A. (1983), Brain organization for language from the perspective of electrical stimulation mapping, *Behavioral and Brain Sciences, 6,* 189-230.

PHONOLOGICAL STRUCTURE AND THE ANALYSIS OF PHONEMIC PARAPHASIAS

Mary-Louise Kean
University of California, Irvine, California

ABSTRACT

Errors in the realization of consonants and vowels by substitution, deletion, and epenthesis are typical of the speech of aphasic patients. The characterization of the structural pattern of such phonemic paraphasias requires appeal to a variety of levels of phonological representation. Critical to any analysis of phonemic paraphasias is consideration of the markedness of the underlying, lexical, and surface repre-sentations of segmental and syllabic structure.

INTRODUCTION

Phonemic paraphasias, errors involving the substitution, deletion, or addition of a vocalic or consonantal segment, are a general characteristic of the speech of aphasic patients. These errors have long been a focus of consideration in the characterization of aphasic speech [1, 2, 3]. Traditionally analyses have been based on two central assumptions: (a) segments are represented as a set of binary distinctive features, and (b) phonological representations are linear, i.e., are ordered strings of segmental repre-sentations. Though the exact content of the featural representations of segments has evolved since the pioneering work of Jakobson, Fant, and Halle [4], the first assumption is still maintained. However, as will be discussed below, the content of featural representations varies in specificity at different levels of phonological representation [5, 6, 7]. The second assumption has, in contrast, been abandoned in favor of multi-dimensional representations with hierarchical structure. Three levels of morpheme/word representation are postulated, an underlying level, a lexical level, and a surface level. The multidimensionality of phonological representations arises from the character-ization of strings on three "planes", the melodic, which provides featural repre-sentations of segments, the syllabic, and a skeleton linking the melodic and syllabic.

At the underlying level of representation (UR) strings are not syllabified and segments have the minimal featural repre-sentations necessary to uniquely individuate segments At the lexical level of representation (LR) strings are syllabified in accordance with the specific algorithm of a language. Featural matrices of segments are further specified at LR by redundancy rules. At the surface level of representation (SR) full featural matrices are assigned to each segment

The theory of markedness has played a role in phonology since Trubetzkoy [8] first introduced the concept in his analysis of segmental structure. For Trubetzkoy, a segment was more or less marked depending on its closeness to the neutral (breathing) configuration of the vocal tract. This notion has long since been supplanted by more abstract characterizations which appeal to the intuition that some segments and structures are more highly favored linguistically than others. The redundancy rules which specify the featural representations at LR can be viewed as markedness rules which assign unmarked values to features which are not specified in UR. Thus, the phoneme /t/ is not specified for voice at UR, since voicelessness is unmarked, while /d/ must be specified as [+voice] at UR. In addition to using markedness in the characterization of segments, markedness considerations also apply in the analysis of syllable structure [9].

Jakobson [10] was the first to argue that

phonological theory could be used to explain the pattern of phonemic paraphasias. He proposed that the pattern of phonemic paraphasias reflected a general tendency for unmarked values to replace the marked values of segments; Blumstein [1] likewise claimed that the substitution errors of aphasic patients involve replacement of marked segments with unmarked ones. In this and considerable subsequent work, it was generally assumed that the pattern of phonemic paraphasias was consistent across aphasic populations. Recent work (e.g., [11, 12, 13]) raises significant questions about this assumption. If the uniformity of the pattern of paraphasias across aphasiological types is in question, then that raises further questions about the claim that markedness considerations play a critical role in accounting for the pattern errors. In what follows, it will be argued that the current approach to phonology provides a means for distinguishing among classes of aphasics and supports, at least in part, the claim that phonemic paraphasias can be explained on the basis of markedness.

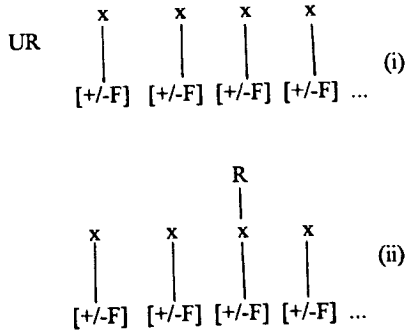
THE APHASIAS

Aphasias, acquired impairments of linguistic capacity, arise typically from damage to the left cerebral hemisphere. There are a variety of distinct symptom complexes in aphasia. In the literature on phonemic paraphasias, three types of aphasia figure most prominently: Broca's aphasia, conduction aphasia, and Wernicke's aphasia. Broca's aphasia and conduction aphasia are nonfluent aphasias, patients exhibiting relatively spared comprehension and word-finding difficulties. These two types of aphasia are distinguished on the basis of limited and/or agrammatic speech output in Broca's aphasia frequently accompanied by arthric disorders, as opposed to an absence of arthric disorders and production of numerous phonemic paraphasias in conduction aphasia. Wernicke's aphasia is characterized as a fluent aphasia. Speech typically contains semantic paraphasias, phonemic paraphasias, and neologisms. Each of these types of aphasia will be considered separately, and it will be proposed that the phonemic paraphasias of conduction aphasia involve UR and LR, with markedness considerations playing a significant role, while the segmental paraphasias of Broca's patients involve SR and motor planning and the errors of Wernicke's aphasics involve retrieval of URs.

Conduction Aphasia

Beland et al. [11] have provided the most detailed analysis of phonemic paraphasias in conduction aphasia. In an extensive single case study of a French speaking patient, they offer a detailed taxonomy of the pattern of simple paraphasic errors, e.g., substitutions, deletions, and additions of single segments, in the context of current phonological theory.

At UR, as was noted above, represen-tations consist of underspecified segmental representations on the melodic plane and a skeleton; syllable structure is assigned in the derivation of LR (Figure 1). The first step in syllable formation is the identification of the rime (R), the vocalic nucleus of the syllable. Next, in this simplified presentation, comes "sigma formation" which establishes a syllable on each rime. Every position to the left of the rime which the syllabification alogorithm provides for is attached to sigma creating an onset. Finally, through coda formation unattached segments to the right of the rime are attached to it.



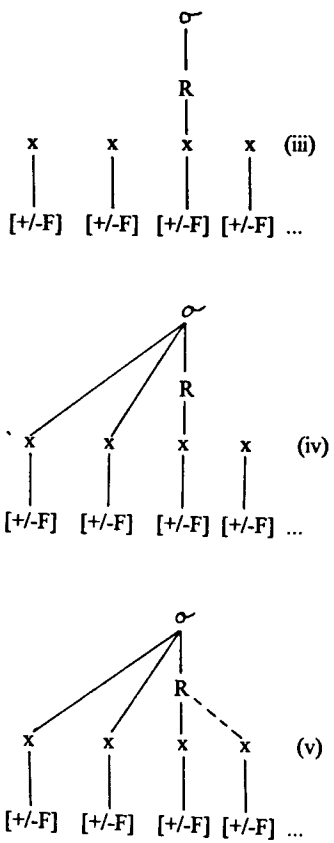


Figure 1. Assignment of syllable structure.

Syllable structure plays an important role in the description of phonemic paraphasias in conduction aphasics. Vocalic omissions can occur where there are adjacent rimes (naive, /na$iv/> [niv]). As adjacent rimes are marked, omission of one vowel reduces markedness. The markedness of adjacent rimes is also reduced by consonant epenthesis (ahuri, /aYRi/ > [alYRi]). Such epenthetic consonants are liquids, glides, or copies of another segment in the string. Syllable structure markedness is also reduced by vowel epenthesis in internal closed syllables (chemise, /ʃmiz/ > [ʃœ$miz], final closed syllables (fleur,/floer/ > [floe$roe], and in complex word onsets (strie, /stri/ > [sœ$tri]). Both vocalic and consonantal deletions occur at the extreme positions in words (kilo, /kilo/ > [kil] but never /kilo/ > [klo]; talc /talk/ > [tal]). Omissions at the extremes are insensitive to syllable structure, hence should be attributed to UR processes, whereas the other errors just described reflect difficulties in syllabification, that is the mapping from UR to LR.

In addition to omissions and deletions, substitutions are also encountered. At UR, where segments are not fully specified, only two types of errors are possible: some feature(s) may be "lost", or the specification on a feature might be "changed". There is no mechanism which allows for there to be any addition of features. Word-internal consonant omissions are found in the structure VC$CV where it is always the first consonant which is omitted, yielding V:$CV (admis, /ad$mi/ > [a:$mi]). Analysis of such errors involves the loss of the melodic plane associated with the first consonant which leaves an empty skeletal position that is automatically associated with the preceding vowel by the syllabification algorithm giving rise to compensatory lengthening. Given an incomplete melodic representation, there are three options available: (a) at extreme skeletal positions, the segment may be lost, (b) the un(der)specified slot is specified through the redundancy rules (diet, /djɛt/ > [tjɛt]; niaise, /njɛz/ > [njɛs]), or (c) has its full speci-fication realized through a process of copying (valse, /vals/ > [valv]). The former two processes must be associated with UR, while the latter process can be accounted for at LR. There is, in fact, a considerable range of errors which are to be accounted for at LR. For example, various consonant deletions, which also serve in the reduction of syllable structure markedness, can only be interpreted with respect to syllabified strings. In onset sequences obstruent + liquid (+glide), the onset may be simplified by the deletion of the

liquid, a deletion which enhance the sonority contrast between the onset and nucleus (droit, /drwa/ > [dwa]). By the same token, codas may also be simplified by consonantal omission; the most sonorous segment in the sequence is deleted which enhances sonority contrast (film, /film/ > [fim]).

Analyses such as that proposed by Beland et al. [11] illustrate the richness of phonological theory and its ability to provide a structure for accounting for a disparate range of errors. At the same time, the theory precludes, in principle, certain classes of conceivable errors. Markedness theory provides further constraints on the range of errors. Thus, for example, vocalic additions never create adjacent rimes, and errors in specification of the melodic plane at UR feed the redundancy rules which assign unmarked specifications to features. The analysis of the error corpus strongly suggests that in conduction aphasia the phonological compromise involves UR and LR. This attribution of the errors of conduction aphasics to a premotoric "phonemic" level is supported by Nespoulous et al. [12].

Broca's Aphasia

The most striking feature of speech in Broca's aphasia is the relative paucity of output. Agrammatism, the tendency to omit function words and various inflectional morphemes, is the most frequently noted realization of this deficit; however, there are also Broca's aphasics who speak in short relatively well-formed phrases and/or sentences. Like other aphasic patients, Broca's aphasics do make phonemic paraphasias, though these are not as ubiquitous as in the speech of conduction aphasics. Consequently, there has not been extensive research focused specifically on the pattern of phonemic paraphasias in this population. However, a consistent view emerges from a review of the literature (e.g.,[2, 3, 12, 14): The phonemic paraphasias of Broca's aphasics reflect simplifications at the levels of SR and motor planning and control.

The phonemic paraphasias of Broca's aphasics typically involve a change of a single feature in segmental realization [14]. Given this, the "misreading" of feature specifications must occur after segments have been fully specified. A prominent aspect of the consonant substitutions of these patients is a tendency to replace a voiced segment with its voiceless counterpart. MacNeilage [2] has argued that this devoicing is a simplification which is a function of the slow rate of speech of Broca's aphasics. Indeed, the devoicing is, at least in part, a reflection of the perceiver rather than the speaker: Phonetic analysis indicates that segments are only partially devoiced [14]. That the segmental disorder of Broca's aphasia is to be analyzed at the level of motor planning and the temporal control of speech production is further supported by the finding that voice onset time can be disturbed in Broca's aphasia [15, 16].

Nespoulous et al. [12] provide a contrastive analysis of the phonemic paraphasias of conduction and Broca's aphasics. They observe, for example, that in conduction aphasia errors in segmental realization are frequently contextually conditioned, but this is not a prominent feature of such errors in the speech of Broca's aphasics. Their data not only replicate the earlier finding regarding the tendency to substitute voiceless consonants for voiced ones, but also indicate a tendency for errors with voiceless consonants to involve a change in place of articulation (e.g., /k/ > [t]; /k/ > [t]). This general finding is also reflected in the data from a variety of studies which is analyzed by Beland and Favreau [17]. Such errors, like shifts in voicing, are readily accounted for at SR

Wernicke's Aphasia

The phonemic paraphasias of Broca's and conduction aphasics illustrate distinctive impairments in phonology, the latter involving the underlying and lexical representations of morphemes/words and the former the surface representation and mechanisms of motor

planning and control. If words have an abstract phonological representation, UR, which must be retrieved, then a third source of possible error is in the retrieval of UR's or in the loss of URs. Wernicke's aphasia, with its characteristic neologisms, is a candidate for such a disorder. Kohn and her colleagues [13, 18] argue for just such a position. Following previous authors, Kohn et al. [13] recognize two types of neologisms: (a) those which are "target-based" and (b) those which have no apparent lexical motivation.

Target-based errors reflect some access to the phonological lexicon. Some target based errors involve simple phonemic paraphasias. In cases where there is a substitution (e.g., bride > blide), Kohn et al. [13] argue that the UR melodic specification of [r] is "lost" and the patient reconstructs the string by analogy to other lexical entries. However, as we have already seen such errors can also be explained on the basis of the application of redundancy rules to an underspecified melodic representation. As the appeal to an analogical repair process involving lexical search is not independently motivated, an analysis utilizing standard phonological theory is preferable.

A second class of errors discussed involves the addition of a syllable; for example, one subject realized "umbrella" as [ʌmrɪpɛlə]. In such an example, there is an apparent relation to the correct lexical entry, but it is impossible to determine a clear source for the error. Descriptively, the error would appear to involve metathesis of /b/ and /r/, devoicing of /b/, and vowel epenthesis, or deletion of the skeletal position of /b/, addition of a consonant (perhaps with labiality copied from /m/), and vowel epenthesis, or deletion of the skeletal position of /r/, etc.. Kohn et al. [13] suggest that a case such as this can be accounted for because "[s]yllables can be 'randomly' added when the process of phonological reconstruction overcompensates for missing information." Again appeal is made to an otherwise unmotivated mechanism of reconstruction. An alternative approach would rely on basic phonological processes and markedness. First, it is assumed, that the shift of /b/ to /p/ arises from a loss of specification of the feature [voice]. This phenomenon has already been recognized in the analysis of phonemic paraphasias in conduction aphasia. Metathesis such as seen in this case can be viewed as a consequence of the breakdown in normal skeletal-melodic linkages in UR. The consequence of the metathesis is the sequence VmrpV which cannot be grammatically syllabified in English - *Vmr$pV, *Vm$rpV. The minimally marked locus for an epenthetic vowel which is consistent with English syllable structure is after the /r/, yielding VmrVpV. That is, once the stop and the /r/ are metathesized, it follows automatically that there will be an epenthetic vowel in a specific location given the syllable structure algorithm and markedness theory. Other target-based errors in their corpus are amenable to similar accounts which assume a disruption of representations at UR or of efficient retrieval of UR representations which are then "repaired" by normal phonological processes. In order to determine the adequacy of such a UR based approach to target-based neologisms, it would be necessary to carry out an analysis of comparable detail to that of Beland et al. [11].

In contrast to the target-based errors are the true neologisms, errors in naming, reading, or discourse which bear no decipherable relation to targets. Kohn et al. [13] provide several such examples, e.g., pig > [batɛntuvɛtu], elephant > /kʌlʌnsʌr/. They argue that cases such as these should be accounted for in terms of a loss of URs and not simply a deficit in the full retrieval of URs. The latter possibility is rejected since, under their theory, it would entail access to analogical reconstruction and the product of such reconstruction should be, at least in some degree, relatable to the target. However, if, as is suggested here, there is no mechanism of analogical reconstruction, then there is no basis for deciding between a retrieval impairment and a loss of URs. Neither hypothesis, it should be clear,

provides a clear basis for explaining neologisms in the context of normal phonological processes. Thus, it would appear that neologisms may well involve more than extreme distortions of phonological representations.

CONCLUSION

It has been argued here, following the tradition laid down by Jakobson [10], that the phonemic paraphasias of aphasic patients can be accounted for in terms of phonological theory. Departing from earlier analyses which were based on the assumption that phonemic paraphasias do not differ across different aphasic syndromes, recent work has provided evidence that the phonemic paraphasias of different types of aphasics are different in character. An overview of recent studies which have looked at the paraphasias of different aphasic populations has illustrated how modern phonological theory can contribute to our understanding of phonological and phonetic disruptions consequent to brain damage. There has, however, not been sufficiently detailed work on a broad enough class of cases to provide convincing evidence that the distinctions among aphasic populations suggested here generalize across patients within syndromes.

ACKNOWLEDGEMENT

This work has been supported in part by NIH-NIDCD grant no. DC02166.

REFERENCES

[1] Blumstein, S. (1973), *A phonological investigation of aphasic speech*, The Hague: Mouton.

[2] MacNeilage, P.F. (1982), "Speech production mechanisms in aphasia", in S. Grillner, B. Lindblom, J. Lubker, and A. Perrsson (eds.) *Speech motor control*, Oxford: Pergamon Press.

[3] Trost, J.E. nd Canter, G.J. (1974), "Apraxia of speech in patients with Broca's aphasia: A study of phoneme production accuracy and error patterns", *Brain and Language*, vol. 1, pp. 63-79.

[4] Jakobson, R., Fant, G., and Halle, M. (1963), *Preliminaries to speech analysis*, Cambridge, MA: MIT Press.

[5] Kiparsky, P. (1982), "From cyclic phonology to lexical phonology", in H.vdH. and N. Smitth (eds.) *The structure of phonological representations*, Dordrecht: Foris.

[6] Pulleyblank, D. (1986), *Tone in lexical phonology*, Dordrecht: Reidel.

[7] Archangeli, D.B. (1984), *Underspecification in Yawelmani phonology and morphology*, unpublished doctoral dissertation, MIT.

[8] Trubetzkoy, N.S. (1969), *Principles of phonology*, LosAngeles: University of California Press.

[9] Kaye, J. and Lowenstam, J. (1981), "Syllable structure and markedness theory", in R. Bandi, A. Belletti, and L. Rizzi (eds.) *The theory of markedness in generative grammar*, Pisa: Estratto.

[10] Jakobson, R. (1942) *Kindersprache, aphasie und allgemeine lautgesetze*, Uppsala University Arsskr., 9.

[11] Beland, R., Caplan, D. and Nespoulous, J-L (1990), "The role of abstract phonological representations in word production: Evidence from phonemic paraphasias", *J. Neurolinguistics*, vol. 5, pp. 125-164.

[12] Nespoulous, J-L, Joanette, Y., Beland, R., Caplan, D. and Lecours, A.R. (1984) "Phonological disturbances in aphasia: Is there a 'markedness effect' in aphasic phonetic errors?", in F.C. Rose (ed.) *Advances in neurology, Vol. 42: Progress in aphasiology*, New York: Raven Press.

[13] Kohn, S.E., Smith, K.L., and Alexander, M. (in press), "Differential recovery from impairment to the phonological lexicon", *Brain and Language*.

[14] Lecours, A.R. and Nespoulous, J-L (1990), "Playing with your speech organ: An unfinished play in five acts", *J. Neurolinguistics*, vol. 5.

[15] Blumstein, S., Cooper, W., Zurif, E. and A. Caramazza (1977), "The perception and

production of voice-onset time in aaphasia", *Neuropsychologia*, vol. 15, pp. 371-383.
[16] Blumstein, S., Cooper, W., Goodglass, H., Statlender, S. and Gottlieb, J. (1980) "Production deficits in aphasia: A voice-onset time analysis", *Brain and Language*, vol. 9, pp. 153-170.
[17] Beland, R. and Favreau, Y. (1991), "On the special status of coronals in aphasia", *Phonetics and Phonology*, vol. 2, pp. 201-221.
[18] Kohn, S.E. and Smith, K.L. (1994) "Distinguishing two phonological output deficits: Activation of stored phonological representations vs. construction of phonemic representations", *Applied Psycholinguistics*, vol. 15, pp. 75-95.

CINEMATIC ACOUSTIC-TO-GEOMETRIC MAPPING

J. Schoentgen * *and S. Ciocea* †
Institute of Modern Languages and Phonetics , CP 110, Université Libre de Bruxelles, 50 Avenue F.-D. Roosevelt, 1050 Bruxelles, Belgium.
* *National Fund for Scientific Research, Belgium,* † *Grant, U.L.B.*

ABSTRACT

The article describes a method of cinematic acoustic-to-geometric mapping. It directly calculates the cross-section areas of a vocal tract model from observed formant frequencies. The map is cinematic since it relates the time-derivatives of area function parameters and formant frequencies.

INTRODUCTION

The acoustic-to-geometric map relates eigenfrequencies of a vocal tract model to its area function. The area function is the link between the areas of the tract cross-sections and their distances from the glottis. Generally speaking, three approaches to acoustic-to-area mapping exist: inversion by table-look-up, inversion by means of optimization and inversion by means of linear prediction coefficients.

We here propose an alternative method. It uses analytical expressions of the time derivatives of the area function parameters to compute iteratively the parameter trajectories. The input data are experimentally obtained formant frequency trajectories. The output are lengths and cross-sections of the tubelets of an n-tubelet vocal tract model. Both tubelet lengths and cross-sections can vary with time. Inversion is mathematically separate from the choice, by means of additional constraints, of a unique area function. This, together with the iterative calculation of the parameter trajectories by means of time derivatives, guaranties that the trajectories are maximally smooth and the agreement between desired and generated formant frequencies is better than $0.01Hz$.

ANALYTIC ACOUSTIC-TO-AREA MAPPING

Conventionally, when the vocal tract shape is approximated by means of a concatenation of uniform tubelets, the link between resonance frequencies ω and tubelet cross sections S_i and lengths l_i is mathematically expressed by means of an algebraic equation $F(\omega, l_i, S_i) = 0$ [1].

It is assumed that formant frequencies ω_j have been experimentally obtained at time coordinates t_k, $t_a \leq t_k \leq t_b$. The acoustic-to-area mapping problem then is to determine the evolution with time of tubelet cross-sections and lengths so that the n-tubelet area function model generates eigenfrequencies $\omega_j(t_k)$ for $t_a \leq t_k \leq t_b$.

Hereafter, the set of area function parameters $\{S_i, l_i\}$ is designated by X_i. Provided that time interval $(t_{k+1} - t_k)$ is small, the link between parameters $(X_i)_{t_{k+1}}$ and $(X_i)_{t_k}$ can be formulated by means of their Taylor expansion.

$$(X_i)_{t_{k+1}} \simeq (X_i)_{t_k} + \left(\frac{dX_i}{dt}\right)_{t_k} (t_{k+1} - t_k). \quad (1)$$

Given X_i and derivatives $\frac{dX_i}{dt}$ at time coordinate t_a, area function parameter trajectories $(X_i)_{t_k}$ can be calculated by iteratively applying relation (1). Time derivatives $\frac{dX_i}{dt}$ can be obtained by means of equation $F(X_i, \omega_j) = 0$. Indeed, the so-called chain rule establishes a link (2) between time derivatives $\frac{dX_i}{dt}$ and $\frac{d\omega_j}{dt}$.

$$\sum_{i=1}^{n} \left(\frac{\partial F}{\partial X_i}\right)\left(\frac{dX_i}{dt}\right) + \left(\frac{\partial F}{\partial \omega}\right)\left(\frac{d\omega_j}{dt}\right) = 0. \quad (2)$$

Index n is the number of area function parameters, ω is the formant frequency variable and ω_j is the frequency of the first, second, third ... formant. The values of $\frac{d\omega_j}{dt}$ are arrived at by numerically derivating observed formant frequency trajectories. Expressions $\left(\frac{\partial F}{\partial X_i}\right)$ and $\left(\frac{\partial F}{\partial \omega}\right)$ are analytically obtained by means of equation $F(X_i, \omega) = 0$. The number of expressions (2) is equal to the number, m, of observed formants. These expressions are turned into a system of $m \times n$ linear algebraic equations by inserting the values of formant frequencies ω_j and area function parameters X_i of time coordinate t_k into analytic expressions $\frac{\partial F}{\partial X_i}$ and $\frac{\partial F}{\partial \omega}$.

$$\sum_{i=1}^{n} a_{ji} x_i = y_j, \quad j = 1, m. \quad (3)$$

Here $x_i = \left(\frac{dX_i}{dt}\right)_{t_k}$, $y_j = -\left(\frac{\partial F}{\partial \omega}\right)_{t_k,\omega_j}\left(\frac{d\omega_j}{dt}\right)_{t_k}$ and $a_{ji} = \left(\frac{\partial F}{\partial X_i}\right)_{(t_k,\omega_j)}$. Singular value decomposition delivers the general solution vector $\overline{x}$, even when $n \geq m$ [2].

$$\overline{x} = \overline{d^*} + \sum_{j=1}^{n-m} \overline{d_j}\lambda_j. \quad (4)$$

Parameters λ_j may take any real value. Vector $\overline{d^*}$ is a particular solution of system (3) and vectors $\overline{d_j}$ are columns of a matrix which singular value decomposition splits off from matrix $\{a_{ji}\}$. The selection of a unique solution is carried out by means of additional constraints. A possible constraint is the requirement that the generalized potential energy of a given area function is a minimum. The generalized potential energy is the greater the farther away an area function is from the "neutral" area function. The definition is the following.

$$E_p = \frac{1}{2}\sum_{i=1}^{n} k_i (X_i - X_{i0})^2. \quad (5)$$

Coefficients k_i are pseudo spring constants which are, for sake of convenience, put equal to 1. X_{i0} are the "neutral" area function parameters.

Inserting solution (4) into generalized potential energy (5) and combining with Taylor expansion (1), the potential energy at time coordinate t_{k+1} becomes the following.

$$E_p = \frac{1}{2}\sum_{i=1}^{n}[(X_i)_{t_k} + (d_i^* + \sum_{j=1}^{n-m} d_{ij}\lambda_j)(t_{k+1} - t_k) - X_{i0}]^2. \quad (6)$$

Computing the extremum condition $\frac{\partial E_p}{\partial \lambda_j} = 0$ leads to a system of $n - m$ linear algebraic equations with $n - m$ unknowns λ_j, system which can be solved by conventional means.

Finally, once the optimal λ-values have been determined, solution (4) yields the values of $\frac{dX_i}{dt}$ at time coordinate t_k, which, inserted into Taylor expansion (1), is used to compute area function parameters X_i at time coordinate t_{k+1}. The procedure then starts all over again with the estimation of $\frac{dX_i}{dt}$ at time coordinate t_{k+1}.

ERROR CORRECTION

Iteratively applied estimation steps accumulate small errors over time. In order to keep error accumulation to a minimum, we applied several additional stratagems, namely *a*) initialization by means of reference area functions, *b*) linearization and *c*) iterative correction of parameters X_i. *a*) Initial area function parameters $(X_i)_{t_0}$ were those that generated the first observed m-tuplet of formant frequencies. Reference formant frequencies were frequencies for which the matching area functions were known. Possible reference frequency m-tuplets were the formant frequencies of the French vowels [a], [e], [i], [o], [u]. Smooth trajectories were constructed by means of interpolation between reference and initially observed formant frequencies. Thus, the inversion procedure started with known reference formant frequencies and area function parameters and iteratively calculated parameters X_i along

interpolated formant trajectories till the initial area function parameter values $(X_i)_{t_0}$ were obtained. From then on, the method followed observed formant trajectories. *b*) The purpose of linearization was to improve the quality of Taylor approximation (1). Several authors have drawn attention to the fact that the relation between the logarithm of the area function and the logarithm of the eigenfrequencies is nearly linear around the uniform area function [3]. A change of variables $X_i \rightarrow \ln X_i$ and $\omega_j \rightarrow \ln \omega_j$ was therefore performed in equation $F = 0$ and in Taylor expansion (1). *c*) The purpose of the iterative error correction presented here was to adjust area function parameters X_i so as to suppress any remaining small disagreements between observed and generated formant frequencies. When parameters X_i at time coordinate t_k were known, relation (1) gave an estimate $(X_i)^{(1)}$ of parameters X_i at time coordinate t_{k+1}. But, formant frequencies $(\omega_j)^{(1)}_{t_{k+1}}$ generated by vocal tract model $(X_i)^{(1)}_{t_{k+1}}$ were generally slightly different from observed formant frequencies $(\omega_j)_{t_{k+1}}$. Therefore, advantage was taken of the fact that quantities $\Delta(\omega_j)^{(1)}_{t_{k+1}} = (\omega_j)^{(1)}_{t_{k+1}} - (\omega_j)_{t_{k+1}}$ and $\Delta(X_i)^{(1)}_{t_{k+1}} = (X_i)^{(1)}_{t_{k+1}} - (X_i)_{t_{k+1}}$ were related by a formula similar to formula (2). As a consequence, given $\Delta\omega_j$, solving an algebraic system analogue to (2) yielded corrections $\Delta(X_i)^{(1)}_{t_{k+1}}$ which were added to the previously calculated $(X_i)^{(1)}_{t_{k+1}}$. The procedure was repeated till the difference $\Delta(\omega_j)^{(p)}_{t_{k+1}}$ was as small as desired (i.e. 0.01 Hz). Typically, p was equal to 2 or 3.

METHODS

Our cinematic acoustic-to-area mapping method was applied to a Kelly-Lochbaum model which consisted of a concatenation of 6 uniform equal-length tubelets. The section of tubelet S_1 at the glottis was fixed at 2.5 cm^2. The cross-sections of the other five tubelets were variable and total length L depended on the area of lip tubelet S_6 ($L\ (cm) = 22 - S_6\ (cm^2)$).

The method was tested on 1170 transitions in an [iVi] context of the first three formants of vowels [a] and [e]. Vowels [a] and [e] were stressed or unstressed and the French carrier sentence was produced by a male speaker at 10 speaking rates controlled by a metronome. Each combination of vowel, stress pattern and rate was repeated 30 times.

RESULTS

The objectives of the test were the following. Firstly, to check that acoustic-to-area mapping did not, in addition to noise stemming from the formant frequency measurements, introduce noise into cross-section trajectories. Secondly, to determine the kind and quantity of gross errors. A gross error occurred when measured and generated formant frequencies differed by more than 0.01 Hz after 10 error correction iterations. Thirdly, to study the dependence of area function parameters on speaking rate and accent pattern. Indeed, a qualitative agreement between dependencies obtained by inversion and reported elsewhere would lend further support to the inversion method presented here.

The results were the following. (1) Shimmer values were computed for the first three formant frequency trajectories and for the trajectories of areas S_2 and S_5. Average shimmer was not greater in the area function trajectories than in the formant trajectories. (2) The number of experimentally obtained formant frequency triplets was equal to 44 879. Gross errors occured 273, 233 and 189 times for the first, second and third formant respectively. (3) Figure 1, for example, represents the trajectories of sections S_i for stressed vowels [a] and [e] in an [iVi] context produced at the fastest (120 beats/sec) and at the slowest (48 beats/sec) rate. It is seen that for vowel [a], "fast" and "slow" trajectories differ considerably more than for vowel [e]. A possible explanation is that, since vowels [a] and [i] differ both on frontness and closeness whereas vowels [e] and [i] only differ on closeness, contrast was easier to implement for the vowel [e] in the case of fast rates. The asymmetry of the trajectory of lip-tubelet cross-section S_6 is possibly related to the fact that, in the carrier sentence, $[i]_2$ was followed by [m].

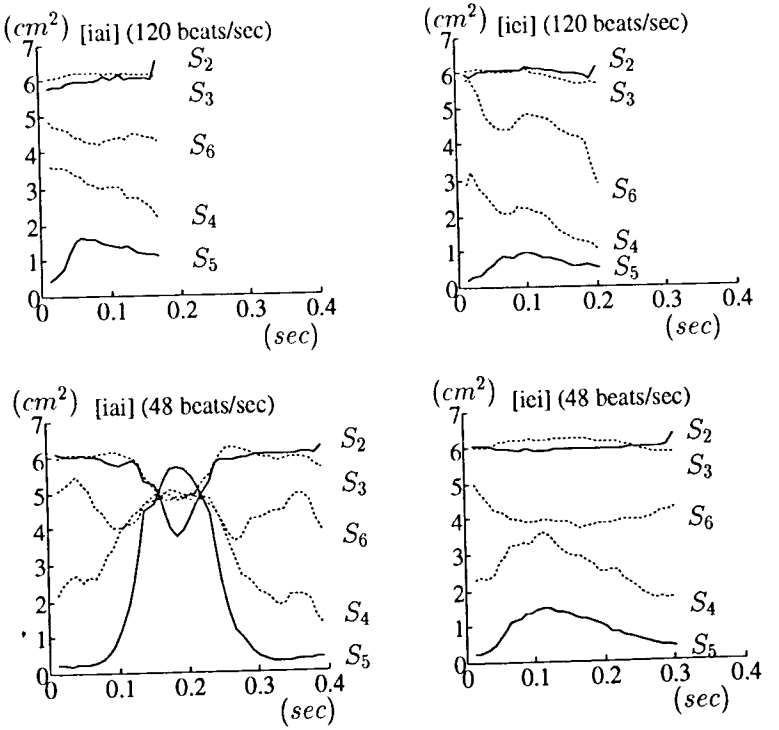


Figure 1.

References

[1] L.J. Bonder. The n-tube formula and some of its consequences. *Acustica*, 52:216–226, 1983.

[2] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. *Numerical Recipes - The Art of Scientific Computing*. Cambridge University Press, New York, 1987.

[3] M.R. Schroeder. Determination of the geometry of the human vocal tract by acoustic measurements. *Journal of Acoustic Society of America*, 41:1002–1010, 1967.

AN OBJECT-ORIENTED STRUCTURE FOR A SPEECH SYNTHESIS SYSTEM

Corine Bickley (1),(2) and Eric Carlson (1)
(1) Sensimetrics Corp., Cambridge, MA 02139, USA
(2) Research Lab of Electronics, MIT, Cambridge MA 02139, USA

ABSTRACT

A new approach to phonetic-string-to-speech synthesis is described. This approach uses an object-oriented framework to structure a speech synthesis system around acoustic landmarks. A landmark-based representation is used in the generation of high-level synthesis parameters. Constraints based on speech production and on the spectral and temporal properties of the speech waveform guarantee that the speech synthesized corresponds to a signal that could have been produced by a human speaker. An example of the use of the system and a discussion of its design are presented.

1. INTRODUCTION

A structure for a phonetic-string-to-speech synthesis system is presented along with examples of phonemes in various phonetic environments. An object-oriented approach was chosen because such systems have been shown to be particularly well-suited for implementing and working with systems consisting of large sets of inter-related constraints [1]. Generating high-quality synthetic speech is more a problem of constraint satisfaction than of specifying parameter tracks for a synthesizer. Synthesis-by-rule systems have traditionally used either ad hoc rule implementations [2] or have been closely modeled on standard phonological notations [3], and have typically been plagued with problems of providing a flexible environment for maintenance and enhancement of rules [4]. An appreciation for such problems has led us to alternative programming methodologies to apply to developing a speech synthesizer.

2. CONSTRAINT-BASED SYNTHESIS

Our approach to speech synthesis is based on a view of speech production as controlling the characteristics of the sound while balancing linguistic goals with speech production constraints. These constraints include, for example, limits on the rates of movement of the articulators, values of resonances of the vocal tract in the vicinity of constrictions, and conditions involving the pressure in the oral cavity and airflows through the orifices of the vocal tract as well as the spectral and temporal properties of the resulting synthesized sounds [5]. The synthesis problem is similar in some ways to the problem solved by a talker attempting to organize the sequence of articulatory movements needed to implement a series of phonemes.

The process of conversion from an input phonetic transcription (with syllabic prosodic specification) to audio waveform proceeds in several phases. The starting point for the synthesis process we describe here is a series of feature markings of the sort proposed in [6]. The transformation from phonetic and prosodic symbols to the series of feature markings is not treated in this paper. Initially this stage will be performed by hand; we plan to draw on the literature for methods of generating a series of feature markings from phonetic and prosodic information. The focus of this paper is the part of the system that generates high-level (HL) parameter tracks [7] – parameters that are based on a speech production model – from feature markings that are clustered together into landmarks. The notion of landmarks [8] is central to the implementation of the synthesis system described in this paper. Landmarks are the anchors around which the constraints are structured, and as such are the principal objects of the system. Currently, the locations in time of the landmarks are determined by hand. Another phase converts HL parameters to low-level (LL) parameters [7]. The resulting LL parameters are then used as input to a Klatt-type formant synthesizer as the final phase, which yields an audio waveform as output.

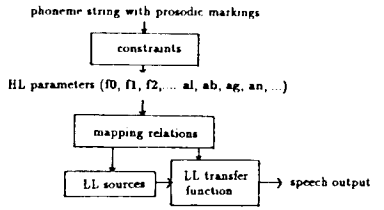


Figure 1: *Process of transformation from string of phonemes into synthesized speech.*

Figure 1 schematizes this transformation from a string of phonemes with prosodic markings into a series of landmarks, and from landmarks into sets of HL parameters. The mapping relations that transform the HL parameters into the low-level parameters that control the sources and transfer functions of a synthesizer such as KLSYN88 [9] have been described previously (for instance, [7], [10]).

A key feature of the kind of system proposed here is the characterization of the synthesis problem as one of constraint satisfaction. The synthesis program encodes constraints that must all be satisfied, and the software finds a solution automatically if possible, and if not, a reason is indicated and modifications to either the specification of the landmarks or to the rules of the system can be made by the user. The opportunity for the user to explore the relationships among the entities (objects) is one important feature of this sort of system.

3. AN EXAMPLE

A short example serves to illustrate the elegance and utility of this sort of object-oriented approach to speech synthesis. The synthesis of consonants is illustrated in this example. Synthesis of sounds involving quickly changing characteristics (such as formant frequencies and bandwidths, and amplitudes of aspiration, frication, and voicing sources) that must be carefully timed relative to one another, has often presented a difficult problem for synthesis-by-rule programs, but yet, in our opinion, is essential to the generation of highly intelligible and natural-sounding synthetic speech.

Consider the phonetic string /wʌnsəpɔnə/ which has sequences containing [+consonantal] segments (/ʌnsə/ and /əpɔn/). One kind of constraint applies to [+consonantal] segments (the /ns/ and the /p/ in /ʌnsəpɔ/, for instance). In each case, a closure/opening gesture for a major articulator (or articulators) must be generated. The rates of closure and subsequent opening as well as the time required for reversal of direction depend on the particular articulator(s) and are based on limits of movement of the physical structure(s). For each [+consonantal] segment, constraints for secondary articulators (such as the glottis and the soft palate) apply. These constraints place limits on the rates of glottal abduction or adduction, rates of change of vocal-fold stiffness, rates of change of the velopharyngeal opening, and rate of change of vocal-tract volume expansion during obstruents.

Figure 2 shows parameter tracks for the individual lip and tongue blade articulators **al** and **ab** (in the top panel). The opening/closing gestures for the lips and tongue blade are similar. The parameter tracks for the secondary articulators of the velum and the glottis are shown in Fig. 3. The complete parameter tracks for the area of each orifice (shown in the bottom panel) are constructed so that the following constraints are all satisfied: closure and opening of the tongue-blade constriction aligned with the landmarks for /n/ and /s/; lip closure followed by opening (approximately 85 ms later) for the /p/; rate of closure/opening is 50 cm^2/sec for both the lips and the tongue blade; for a [+consonantal,−continuant] segment, the minimum area is zero; for [+consonantal,+continuant] segments, the minimum area is such that the amplitudes of the sources of the synthesized sound are appropriate in relation to the adjacent vowel; the area parameter tracks are constructed from individual opening/closing gestures that are derived from the landmarks – the complete track is constructed using the minimum value at each point in time; the values of **f2** and **f3** near each [+consonantal] land-

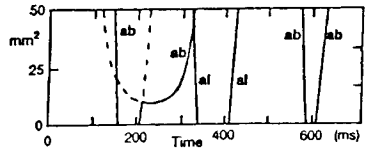


Figure 2: *Parameter tracks for lip* **al** *and tongue blade* **ab** *opening/closings.*

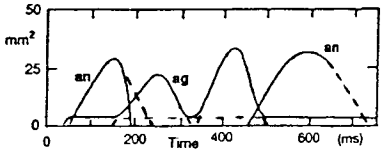


Figure 3: *Parameter tracks for the velum* **an** *and the glottis* **ag**.

mark are constrained to be appropriate for the corresponding oral-cavity constriction in relation to the features for the adjacent vowel; the parameter **ag** must reach a large enough value so that the airflow at the oral cavity constriction causes an appropriate amplitude of aspiraton to be calculated; for any [+consonantal,+nasal] segment, **an** reaches a peak at the first landmark of the pair of [+consonantal] landmarks, and the rate of opening and closing is about 10 cm²/sec. The vocal-tract resonances, target areas, and rates of change of area are adjusted automatically so that appropriate airflows and pressures are maintained to meet the constraints listed above. For any situation in which is it not possible to satisfy all of the specified constraints, the system flags the condition and an adjustment to either the landmark timing or the features markings can be made by the user.

4. SYSTEM DESIGN

Many complex systems may be modeled as interactions among a collection of less complex objects. Object-oriented systems are usually decomposed into a static component - the classes, types, or structures of objects - and a dynamic component - the methods or functions and the evaluation of interactions. The design of an object-oriented system consists of (1) choosing appropriate objects and (2) specifying how the objects interact.

In converting from the landmark representation to HL parameters, four major classes of objects are involved: landmarks, features, articulators, and parameters. Landmarks may be one of vocalic, glide, consonantal closure, or consonantal release, where the consonantal landmarks occur in pairs. The landmark objects combine a time with a list of features, which are binary (+/−) values and an associated label. The articulator-bound features add an associated articulator, which may be either the primary or secondary articulator. The output parameter objects are lists of control points, where control points are time/value pairs. There are ten HL parameters: **f0, f1, f2, f3, f4, ag, an, ab, al,** and **ue** [7]. For each of the three different types of landmarks (vocalic, glide, and consonantal) the object-oriented procedure uses one underlying, basic landmark object. This basic object is a *generic landmark*, one that captures the behaviors and attributes intrinsic to any landmark. The generic landmark consists of attributes such as phonetic-features, preceding-phonetic-context, and following-phonetic-context, as well as information representing the prosodic context.

When rules are stated using traditional phonological notations, such as "A − > B / X _ Y", evaluation takes the form of searching a database of rules for any which are applicable to the current context. In essence, we begin by turning this system inside-out. Each landmark object notifies its predecessor and successor of its features, and is likewise notified by them of their features. Feature messages bypass landmarks unconcerned with that particular feature, providing a natural means of expressing underspecification. Within a landmark object, the cumulative effect of the incoming messages is to apply any applicable rules, sometimes causing additional features messages to be sent to the surrounding landmarks. The rules may therefore be stated subjectively from the viewpoint of each type of landmark, allowing an object-oriented system to simplify much of the rule organization and selection.

The list in Section 3 describes a group of interrelated constraints. It is not an ordered list of calculations. Instead, the system satisfies all of the constraints in a demand-driven order, that is, each calculation is performed only when the value it defines is needed (demanded) by some other calculation or

user action. An important aspect of demand-driven evaluation is that there is no need for the user to specify the particular order in which the computations are to be performed, as this sequencing is done automatically. For instance, if the user requests to display the LL parameter track AF (amplitude of frication), then the sytem demands the values for the oral-cavity pressure P_m and for the minimum cross-sectional area of an oral cavity constriction (acx) as seen in the equation for AF [5]: $AF = 20\log[K_f P_m^{1.5} acx^{0.5}]$, where K_f is a scaling factor. The pressure P_m is calculated using a low-frequency equivalent circuit model of the vocal tract [7]. The minimum area acx is determined by comparing the areas of the glottis, the tongue-body constriction, the tongue-blade constriction, and the lip opening.

Because the input to the synthesis process is in terms of feature markings and not areas and pressures, a range of values for the areas is (usually) possible, and the system calculates the values of P_m and acx that result for each choice of area allowed within the range (in appropriate increments, such as 1 mm²). A rule for "best" configuration (such as "the amplitude of the noise at the constriction is closest to the amplitude of voicing") is used to select automatically the values of P_m and acx, and therefore AF. In this way, the demand-driven feature of the sytem mirrors in some ways the behavior of a talker who adjusts the vocal-tract configuration in order to produce the "best" production of a series of phonemes.

5. SUMMARY

The development of this system has only recently begun, so there remain many unasked as well as unanswered questions, and our understanding will no doubt evolve along with the system. Nonetheless, we believe that this approach represents a significant step toward taming the complexity of existing models of speech production, and it may provide fresh insights into that complexity. For example, we can at this time only mention the apparent connection between feature geometries and object-oriented systems. This approach to speech synthesis also demonstrates the efficacy of both the HL and acoustic landmark representations of speech.

ACKNOWLEDGEMENTS

We gratefully acknowledge K.N. Stevens for helpful discussion during the preparation of this paper. This work was supported in part by NIH Grant R43 MH52358-01.

REFERENCES

[1] Wagner, M. (1988), *Understanding the ICAD System.* Cambridge, MA: Concentra Corp.

[2] Klatt, D.H. (1987), "Review of text-to-speech conversion for English", J. Acoust. Soc. Am. **82**, 737-793.

[3] Hertz, S.R. (1990), "The Delta programming language: an integrated approach to nonlinear phonology, phonetics, and speech synthesis", in *Between the Grammar and Physics of Speech (Papers in Laboratory Phonology I).* Eds. J. Kingston and M.E. Beckman. Cambridge: Cambridge Univ. Press.

[4] Local, J. (personal communication)

[5] Stevens, K.N. (forthcoming) *Acoustic Phonetics.*

[6] Stevens, K.N. (1993) "Lexical access from features", in *Speech Technology for Man-Machine Interaction.* Eds. P.V.S. Rao and B.B. Kalia. New Dehli: Tata McGraw-Hill, 21-46.

[7] Stevens, K.N. and C.A. Bickley (1991), "Constraints among parameters simplify control of Klatt formant synthesizer", J. Phonetics, **19**, 161-174.

[8] Stevens, K.N., S.Y. Manuel, S. Shattuck-Hufnagel, and S. Liu (1992) "Implementation of a model for lexical access based on features", in *Proceedings of ICSLP*, October 12-16, Banff, Canada.

[9] Klatt, D.H. and L.C. Klatt (1990) "Analysis, synthesis, and perception of voice quality variations among female and male talkers", J. Acoust. Soc. Am., **87**:2, 820-857.

[10] Bickley, C.A., K.N. Stevens, and D.R. Williams (1994), "A framework for synthesis of segments based on articulatory parameters", in *Conference Proceedings of ESCA/IEEE Workshop on Speech Synthesis*, Sept. 12-15, 1994, New Paltz, New York.

A MODEL OF FRICATION NOISE SOURCE BASED ON DATA FROM FRICATIVE CONSONANTS IN VOWEL CONTEXT

P. Badin, K. Mawass and E. Castelli

Institut de la Communication Parlée, Grenoble, France

ABSTRACT

Aerodynamic and acoustic parameters have been measured for a subject uttering reiterant [pVFV] non-sense words at various loudness levels. Using *ensemble averaging* and *simplified inverse filtering* techniques, SPL and overall spectral tilt variations have been determined. A model of relations between these characteristics of the noise source and the pressure drop and minimum area at the constriction has been established by multilinear regression.

INTRODUCTION

The rapid growth of interest for articulatory synthesis calls for good articulatory *plants*, i.e. models capable of faithfully reproducing the articulatory, aerodynamic and acoustic behaviour of the human speech apparatus. In this domain, knowledge on noise excitation sources is still badly needed.

A recent study [1] has established relationships between noise source characteristics (SPL, overall spectral tilt) and aerodynamic characteristics (pressure drop and minimal area at the oral constriction), from data acquired for a subject sustaining voiceless fricatives. The aim of the present study was thus to extend this noise source model to fricatives in vowel context.

EXPERIMENTAL SET-UP AND CORPUS

Since the constriction area can be determined from the volume velocity and the pressure drop across the constriction, the set-up included a circumferentially vented wire-screen pneumotachograph, known as Rothenberg Mask to measure the volume velocity. In addition, the mask was equipped with a small polyethylene tube inserted through the lips at the mouth corner, and connected to a pressure transducer. The tube was running on one side of the mouth between the cheek and the gum, up to about 1 cm downstream the limit between the soft and the hard palates, in order to allow measurements of the pressure upstream a possible palatal-alveolar constriction. An electret microphone was placed at a distance of approximately 10 cm from the lips, and at an angle of 45° from the subject sagittal plane. The three signals delivered by the microphone, the pressure transducer, and the volume velocity transducer, were directly digitised at 12 kHz and stored on a computer disk.

The corpus consisted of reiterant [pVFV] non-sense words repeated about 10 times on one expiratory breath by one subject. The fricatives [f s ʃ] were combined with three symmetric vowel contexts [a i u], leading to 9 different items. Following the strategy already used in [1], each item was uttered at 18 different loudness levels, resulting in a SPL range of approximately 30 dB.

DATA PROCESSING

Segmentation

For each [pVFV] repetition, VTT (Voice Termination Time) and VOT (Voice Onset Time) instants were manually detected. The aim was to extract the largest possible portion of the fricative, including both transitions in and out of the fricative, but taking care to exclude any portion where voicing could be present, in order to ensure that the spectral characteristics of the signal would not be modified by the voice source. This resulted in nine fairly similar signals for each set of repetitions (the first repetition was systematically discarded). The length of the fricative segments thus defined varied between 190 and 300 ms over the whole corpus, with smaller variations for each context. For each of the nine repetitions, 20 measurement windows of 10 ms length were then uniformly distributed over the whole fricative segment, with some overlap if necessary, in order to allow the time alignment of the nine repetitions of each item.

Pressure drop and cross-sectional area at the constriction

The pressure drop Δp across the constriction is assumed to be very close to the Intra-Oral Pressure, and the volume velocity at the lips very close to that at the constriction. The Δp and volume velocity signals were first low-pass filtered at 80 Hz with a zero phase filter, and then used to determine the cross-sectional area A_c by means of the *orifice equation* [2]. Finally, Δp and A_c were averaged in each of the 20 measurement windows defined above. Fig. 1 gives an example of Δp and A_c trajectories obtained for [paʃa].

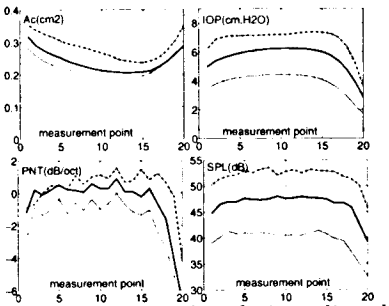


Fig. 1 – Example of time-aligned trajectories (solid line: medium; dotted line: soft ; dashed line: loud).
Top left: A_c; top right: Δp; bottom left: TLT; bottom right: SPL.

SPL and overall spectral tilt variations

In order to study the spectral variations of the fricatives in vowel context, two techniques have been combined: *ensemble averaging* [3] and *simplified inverse filtering* [1].

The spectrum of fricative sounds can be reliably determined by time averaging spectra computed from consecutive time windows throughout the duration of the fricative, if a segment of sufficient duration is available (about 100 ms). However, this technique can not be applied to the rapidly changing spectra in the transitions in and out of the fricative. Therefore, the alternative *ensemble averaging method* has been used (cf. [3]). In this case, averaging was done not *over* time, but at the *same* time across the ensemble of nine repetitions of the same fricative segment. Finally, for each item, 360 spectra (18 levels × 20 measurement points) were obtained by ensemble averaging, cumulating a time duration of 20 × 10 ms. These spectra consist of 61 frequency bins of 100 Hz width.

As measuring directly vocal tract noise sources is impossible, indirect techniques have to be employed. Inverse filtering is widely used to determine voice source for vowel configurations, but such a technique is difficult to implement – and not very reliable – for fricatives [4]. Therefore, a procedure of *simplified inverse filtering* was used (cf. [1]). This procedure does not yield absolute spectral characteristics of the source spectrum, but provides an estimation of the variation of the overall spectral tilt of this spectrum as a function of speech effort. We have verified by simulation that small variations of the constriction area have little influence on the acoustic transfer function overall spectral tilt: the assumption that the variation of spectral tilt of the radiated sound is due to that of the source spectrum seems thus valid.

Simplified inverse filtering has been applied to each of the 9 items. First, third-octave spectra have been determined from the linear spectra mentioned above. The average of the corresponding 360 spectra has been computed for each item, and differential spectra have been estimated as the difference between each spectrum and the average. These spectra are deemed to represent the effects of the variations of the source spectrum, whereas the average spectrum represents the cumulated contribution of the average source spectrum, of the vocal tract transfer function and of the radiation characteristics at the lips. They have been found approximately flat between 200 and 5000 Hz. Overall spectral tilts (henceforth TLT) have finally been estimated as the slopes, expressed in dB/Oct., of the regression lines fitting these differential spectra. Thirteen third-octave bins were retained: from 250 Hz (224-280 Hz) to 4000 Hz (3550-4500 Hz).

On the other hand, overall SPL was computed as the RMS average of the sound pressure in each of the 20 measurement windows. The resulting trajectories of SPL and TLT are exemplified in Fig. 1.

Note that the validity of the simplified inverse filtering method has been verified

[1], including the effect of the mask upon the radiated sound.

RESULTS

Aerodynamic variables

An example of trajectories of Δp and A_c is displayed in Fig. 1. The minimum of each A_c trajectory has been determined for each item, each level and each context. Fig. 2 shows the span of this minimum for the different context. These data fit well with similar data measured on the same subject (PB) for the same corpus, but at a medium level only [5]. The variance of the present data is higher, likely due to the variation of loudness level.

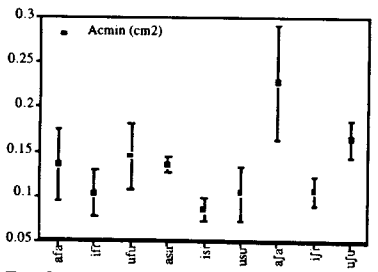


Fig. 2 – Minimum constriction area

In the previous study, a relatively strong correlation had been found between Δp and A_c. In the present corpus, these parameters are globally more independent, even though local correlations (at the level of one repetition) can be observed, as exemplified in Fig. 3 (V-shaped trajectories).

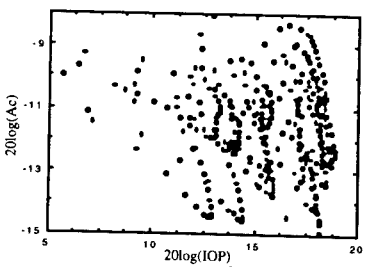


Fig. 3 – A_c vs. Δp for [paʃa]

Acoustic variables

A correlation analysis has shown a high correlation between SPL and Δp, and a smaller, though not negligible, correlation between SPL and A_c. Similar results were found for the spectral tilts, but with lower correlation coefficients. On the average, these correlations are lower than those found for the sustained fricatives in [1].

SOURCE VARIATION MODEL

Badin et al. [1] assumed that SPL and TLT can be expressed as:

$$SPL = k_1.\Delta p^p.A_c^q \quad (1)$$

$$TLT = k_2.\Delta p^r.A_c^s \quad (2)$$

where the exponents p, q, r and s, and the coefficients k_1 and k_2, can be determined by applying a multiple linear regression analysis to SPL and TLT as dependent variables, using $20\log_{10}(\Delta p)$ and $20\log_{10}(A_c)$ as independent variables. In order to extend the results on sustained fricatives to fricatives in vowel context, a similar analysis was performed for each of the nine contexts of the present corpus. Results are given in Fig. 4.

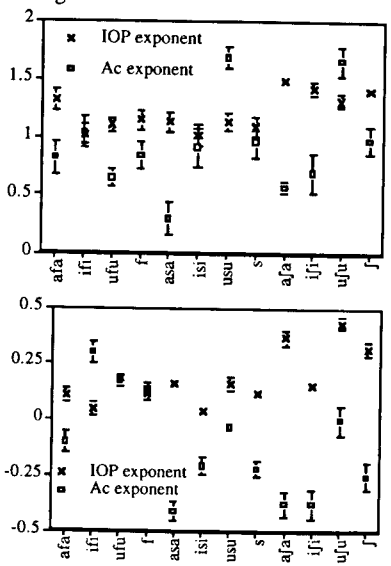


Fig. 4 – Δp and A_c exponents for SPL (top) and TLT (bottom), with the 95% confidence intervals

For each fricative, the exponents display a fairly large dispersion as a function of context. Computing exponents for each fricative, with the three vowel contexts pooled, did not lead to good results, because of unrealistic compensations between the influence of Δp and A_c. Instead, a single set of exponents has been derived for each fricative, by averaging the exponents obtained for the three vowel contexts, and then determining the constant as the average of the residues, context by context, in order to take into account the level differences related to the three vowels. The resulting fit is less optimal, but presents the advantage of a unique model for each fricative. The coefficients obtained are shown in Table I.

	p	q	r	s
[f]	1.15	0.84	0.12	0.13
[s]	1.10	0.97	0.12	-0.22
[ʃ]	1.41	0.98	0.32	-0.25

Table I – Model exponents for the three fricatives

When comparing these exponents with those obtained for the sustained fricatives, the first striking observation is the relatively strong influence of A_c upon SPL (q close to 1). This may be ascribed to the fact that the span of A_c in the present data is larger, due to transitions in and out of the fricatives. It is also worth observing the negative correlation between TLT and A_c for [s ʃ] whereas [f] exhibits a negative correlation (in fact for [ifi] and [ufu] only, cf. Fig. 3).

ASSESSMENT OF THE MODEL

It was first verified that the absolute error between the predicted and measured SPL values was about 2-3 dB. The error between TLT's is relatively larger, in proportion, about 0.5 dB/Oct. As for statistical models in general, we have observed that the prediction was less accurate for the data far from the centre of the distribution, i.e. the low values of Δp and the high values of A_c (up to 6 dB for SPL and up to 2 dB/Oct. for TLT).

Third-octave spectra have been finally re-synthesised as a function of Δp and A_c. For each single measurement point, SPL and TLT were first estimated, and then synthetic third-octave spectra were computed as the sum of the average spectrum established for corresponding context, and of the SPL and TLT variations. Fig. 5 displays the measured and synthetic spectra at 3 measurement points for [paʃa]. The analysis of the results has also shown that the mean absolute error (i.e. the average of the absolute differences of the third-octave spectra in dB) between measured and synthetic spectra was about 2-3 dB, with occasional peaks up to 6-7 dB.

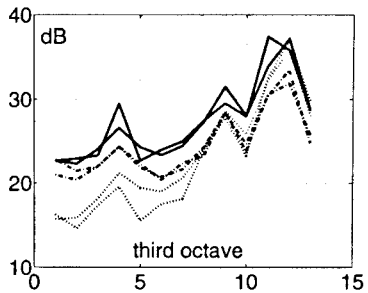


Fig. 5 – Example of 3 measured and re-synthesised spectra for [paʃa].

CONCLUSIONS

We have presented a model of variation of noise source spectrum based on a corpus of fricatives in vowel context. The model predicts the variations of SPL and overall spectral tilt of the source spectrum as a function of the aerodynamic state in the vocal tract. This model will next be implemented in an articulatory synthesiser and thus tested in a complete articulatory synthesis scheme.

ACKNOWLEDGEMENTS

This work has been partially funded by the EC ESPRIT/BR project *Speech Maps*. We are indebted to Denis Beautemps for his suggestions on statistical processing.

REFERENCES

[1] Badin, P., Shadle, C.H., Pham Thi Ngoc, Y., Carter, J.N., Chiu, W., Scully, C., & Stromberg, K. (1994). Frication and aspiration noise sources: contribution of experimental data to articulatory synthesis. *ICSLP*, Yokohama, Japan, Vol.1, 163-166.

[2] Scully, C. (1986). Speech production simulated with a functional model of the larynx and the vocal tract. *Journal of Phonetics*, 14, 407-414.

[3] Shadle, C.H., Dobelke, C.U., and Scully, C. (1992). Spectral analysis of fricatives in vowel context. *J. de Physique* IV, Coll. C1, supp. au J. de Phys. III, vol.2, April 1992. Pages C1-295 to C1-298.

[4] Badin, P. (1991). Fricative consonants: acoustic and X-ray measurements, *Journal of Phonetics* 19, 397-408.

[5] Stromberg, K., Scully C., Badin, P., & Shadle, C.H. (1994). Aerodynamic patterns as indicators of articulation and acoustic sources for fricatives produced by different speakers. *Proc. of the Institute of Acoustics*, Vol. 16, Pt 5, pp 325-333.

ASSESSMENT METHODS OF SPEECH SYNTHESIS SYSTEMS FOR CHINESE

Jialu Zhang, Shiqian Qi and Ge Yu
Institute of Acoustics, Academia Sinica
Beijing P.O.Box2712, China

ABSTRACT

A national assessment of the performance of speech synthesis systems for Chinese has been carried out yearly since 1994. The synthetic speech quality of five different systems were evaluated and diagnosed by using speech intelligibility tests. 16 college students (8 male, 8 female) with no experience with speech synthesis were the listeners, they were asked to do open response task by pencil-paper. In addition, speech naturalness was measured by Mean Opinion Score(MOS) on a ten Point scale. The perceptual confusion matrices of consonants were analyzed in order to give some diagnostic information of the synthesis systems at segmental level. And the statistical relations between speech intelligibilities at different levels of synthetic speech were compared with that of natural speech to explore some deficiencies in prosody. It is shown that stable and rational quality index and diagnostic information can be obtained in this method.

The evaluation methods of prosody have not been become available.

INTRODUCTION

With the growing use of synthetic speech in information-retrieval and man-machine communication systems in China, speech synthesis systems including text-to-speech systems are developed rapidly. Regarding to the COCOSDA work[1,2,3,4] in this field a national evaluation of the performance of synthesis systems for Chinese is to be needed in order to promote the development and the enhancement of speech synthesizers and to compare them with the systems for different languages. The Intelligent Computers Section of National Project 863 and the National Natural Science Foundation have supported the annual assessment activity on speech synthesis and recognition systems for Chinese since 1994.

So far more of available text-to-speech systems (TTS) have a phoneme intelligibility score close to that for natural speech and none of the present TTS for Chinese equipped with complete prosodic rule systems. At first stage we still have to pay more attention to speech intelligibility. Speech intelligibility test method for Chinese was developed at the beginning of 60s and it became one of the national standards.

The goal of the test is to evaluate the speech quality of different synthesis systems and to give some diagnostic information for each system individually.

Chinese is a tone language with multi-tone system and there are some distinctions at both phonetic level and syntactic level. For instance, there are only about 400 mono-tonal syllables used in real speech, almost no affixes exit at word level and the principle of sentence structures is almost the same with the phrase structures. So we have to carefully design the testing material according to the phonetic and linguistic characteristics of Chinese. The testing materials are described in section II and section III explains the testing methods. Then the results and discussion are presented in section IV and V.

TEST MATERIALS

The syllable lists

The syllable lists KXY1-10 are phonetically balanced in initials, finals and lexical tones. Each basic syllable list has 75 syllables divided into 25 trisyllablic groups. The basic syllable lists totaled 10 (KXY1-10). A test item is randomly composed of three syllables, so that a lot of testing syllable lists can be multiplies greatly from the basic lists. And there is no memory effect in the lists. Every test item---a trisyllabic group is added to a carrier sentence, such as "Dì yī zǔ shì ×××" (The first group is ×××.)to present to the listeners.

The word list

The word lists KXC1-10 are phonetically balanced too. There are 100 words in each word list in which the occurrence frequency of the word with different lengths is nearly the same with daily speech. There are 10 basic word lists composed of 1,000 common words, which were elected from "A list of 3,000 word of standard Chinese"[5] by 30 people.

Each list was divided into 25 four-word groups when it is presented to the listeners. A four-word group is a semantically unpredicted sentence, such as "Xīguā chī hēisède tàiyáng" (The watermelon eats the black sun).

By changing the four-word combinations a lot of different testing word lists can be deduced.

The sentence lists

Each sentence list is composed of 25 sentences which were selected from newspapers. Generally the sentence length is less than 7 words, and the content of the sentence lists covers a variety of domains. A strict criterion for judging the sentence intelligibility adapted is that any one key word i.e. the content word in the sentence is mistaken then this sentence is considered to be wrong. So that none of the poor systems can easily get 100% score with sentence lists

TEST METHODS

Listeners

16 college students (8 male, 8 female) who speak standard Chinese---"*putonghua*" with normal hearing were selected as the listeners. They have no experience with synthetic speech. Some instructions and training were given before test work. Especially, it should be explained that the tests aim only at to evaluate the synthesis systems not at the listeners, and the scores are only of the indices of the system performances. They were asked to do an open response task to write down the syllables, the words or the sentences on a special form after the stimuli presented.

Training

The listeners were trained by doing speech intelligibility tests with natural speech and with degraded speech (narrow bands and lower S/N), for 2-4 hours before evaluating the synthesis systems. And the test materials read by two speakers(one male and one female), which were included in the CAS speech database, were presented at first during test sessions, in order to set a reference of making naturalness assessment.

Testing

Two syllable lists, two word lists and two sentence lists were presented to the listeners sequentially, and the listeners were asked to make qualified judgments after each word list and each sentence list on a 10 point scale of naturalness. The categorical judgment on the 10 point scale is like that : excellent -9~10, good-7~8, fair-5~6, bad-3~4, very bad-1~2.

It took about 35 min. to evaluate a system and then a 5 min. break was given for listener rest and tested system preparation.

The listeners wrote down their open response in Chinese characters or in "*Pinyin*"(Chinese phonetic alphabet).

RESULTS

The speech intelligibility and naturalness of five systems with normal speech rates (3~4 syllables/s) are given in Table 1 and 2.

1. Speech intelligibility

Two lines of data of intelligibility scores S, W and J for each system were resulted from using two different test lists respectively, $\bar{x}$ stands for the average value over all listens and σ the standard deviation. In Table 1 the fact that the differences in intelligibility S and W between two test lists in the same category are less than the deviation among listeners means that the test lists are well equivalent. As for the sentence intelligibility J the deviations are rather high.

It is worth to notice that the sentence intelligibility J of natural speech generally is higher than the word intelligibility W. For synthetic speech of the systems tested the

relation is inverse except system 4#. That means on the one side all the systems evaluated are not so good in prosodic processing and the judgment criterions of sentence intelligibility we adopted are sensitive and practical on the other.

Table 1, Intelligibility scores of speech synthesis systems for Chinese

Interlligibility		Natural Speech		1#		2#		3#		4#		7#	
		$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$	σ
Consonant	C			86.7		80.0		80.8		72.5		96.5	
Syllable	S	95.6	3.1	80.7	6.5	73.6	4.3	74.5	6.0	65.6	6.5	94.9	2.3
		94.0	3.7	83.9	5.7	76.0	3.8	77.6	5.6	66.2	9.1	93.4	2.8
Word	W	98.1	1.6	94.3	3.2	87.0	4.6	76.3	8.0	81.8	6.5	95.1	1.5
		97.6	2.4	92.9	5.1	79.8	5.0	73.4	8.1	81.8	7.2	96.3	1.8
Sentence	J	99.9	0.7	57.0	13.5	54.1	14.6	61.5	13.8	83.3	10.1	89.5	7.4
		100	0	74.5	10.7	75.5	12.7	60.0	7.7	86.1	9.0	95.0	5.4

Note: 1# - PCM waveform edited; 2# - CELP concatenative; 3# - VQ-LPC concatenative; 4# - formant; 7# - PSOLA.

2. Naturalness

The naturalness in MOS of the synthetic speech of five different synthesis systems is listed in Table 2.

Table 2. The naturalness in MOS of five synthesis systems for Chinese .

	Systems				
	1#	2#	3#	4#	7#
Naturalness	4.6	3.2	3.4	5.6	7.8

It can be seen from Table 1 and 2 that the naturalness of synthetic speech is still not so high although some systems, such as 7#, can get quite high intelligibility. The ,speech quality of naturalness 7.8 in MOS, the highest in the five systems, is nearly equivalent to that of 8k bps VSELP (Victor Sum Excitation Linear Prediction) telephone system.

3. Comparison between natural speech and synthetic speech

Fig.1 shows the statistical relation between syllable intelligibility and word intelligibility and Fig. 2 the statistical relation between syllable and sentence intelligibities, they were established for natural speech transmitted under different conditions. And the data of synthetic speech of five systems were also drawn on Fig.1 and Fig.2.

From Fig.1 it can be seen that the statistical relation between syllable intelligibility and word intelligibility is almost the same for natural speech and synthetic speech. As for

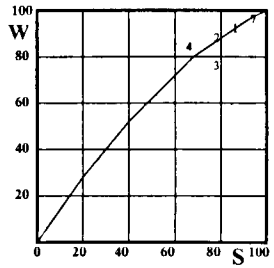


Fig.1 The statistical relation between syllable intelligibility S and word intelligibility W. (Solid line: natual speech; Numeric: synthetic speech.)

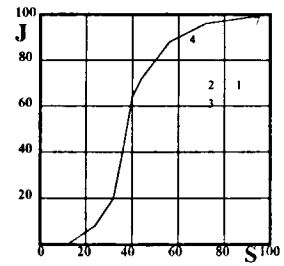


Fig.2 The statistical relation between the syllable intelligibility S and sentence intelligibility J. (Solid line: natural speech; Numeric: synthetic speech.)

the statistical relation between syllable intelligibility and sentence intelligibility in Fig.2, however, the sentence intelligibility of synthetic speech is always lower than that of natural speech. System 1# a text-to-speech system with bad prosodic rule system got lower sentence intelligibility and lower naturalness but higher syllable and word intelligibilities. On the contrary system4# a text-to-speech system using formant synthesizer with better prosodic rule system got higher word and sentence intelligibilities but lower syllable intelligibility.

4. The perceptual confusion of consonants

The perceptual confusion metrices of consonants were analyzed to get the diagnostic information.

In systems 1# and 7# the most frequent confusion was occurred in between voiceless fricatives and aspirated plosives and affricates. This phenomenon makes clear that the waveform edited systems, both PCM(1#) and PSOLA (7#), have some defects in voiceless fricative processing. As for the formant synthesizer, system 4#, the unaspirated plosives and affricates got the lowest segmental intelligibility, that means it is difficult to control the parameters of voiceless plosives and affricates by using formant synthesizers.

DISCUSSIONS

Speech intelligibility tests used as a assessment method of speech synthesis systems can give very much useful information about the system performance at phonetic level. It is more important at the developing stage in the laboratory. The perceptual confusion matrices of consonants give you a clear pattern of the defects the synthesis systems have and it can be a useful diagnostic method.

Almost all synthesis systems developed in China are syllable-based as there are only about 1200 syllables with tonal patterns being used in real speech. The tone intelligibility is high, and the major perceptual confusion of tones was in between rising tone and dipping tone. The same confusion pattern of tone perception was observed in natural speech[6] , too. Maybe this is the psychological basis of the distinct tone sandhi rule --- the dipping tone should be changed into rising tone when it is followed another dipping tone in spoken Chinese.

There is another lexical tone --- light tone at word level in spoken Chinese besides the four lexical tones at syllable level. The light tone syllable is really atonic from the acoustical point of view, but it was treated as a toneme in traditional phonology of Chinese. From Fig 1 it can be seen that the word intelligibility is closely related to the syllable intelligibility, and the word intelligibility is not so strongly effected by the prosodic features. At sentence level, however, thing are different.

REMARK

There is no doubt that prosodic features, especially the interaction between tones and intonation, are more important for speech naturalness of Chinese but the evaluation methods have not become available. A lot of work have to be done in this field.

REFERENCES

[1] Pols, L.C.W. and SAM-partners, "Multi-lingual synthesis evaluation methods", ICSLP'92(Banff), Vol.1, 181-184, 1992.

[2] "Section Two: Synthesis Assessment" in Proc. 1992 Workshop of COCOSDA (Banff), Ed. by Kate Jones and Joseph Mariani, PP.II:1-II:4, 1992.

[3] "Synthesis"in Report on the COCOSDA Workshop (Berlin), PP.25-28, 1993.

[4] "Synthesis" in 1994 International workshop on International Coordination of Speech Databases and Speech I/O Systems Assessment, PP.15-21, 1994(Yokohama).

[5] "A list of 3,000 words of Standard Chinese" Ed. by Research and Spread Section, China Charater Reform Committee, Character Reform Publishing House, 1959 (Beijing).

[6] Yang, Y. and Fang, Z., "Study of Tone Perception of Standard Chinese", in Proc. 5th International Symposium on Chinese Language - Cognition Science, (Science Publishing House, Beijing), 1992.

THE REALIZATION OF PLOSIVES IN NASAL/LATERAL ENVIRONMENTS IN SPONTANEOUS SPEECH IN GERMAN

K. J. Kohler
IPDS, Kiel, Germany

ABSTRACT

This paper deals with the realization, within words and across word boundaries, of plosives as glottal stops or through creak in surrounding sonorants in German spontaneous speech. It also discusses the wider spectrum of phonetic processes affecting plosive production in the frame 'nasal + plosive + /@/ + nasal'.

INTRODUCTION

It was shown in [1] on the basis of limited data from read speech that German plosives within nasal environments may be realized in a way that the velum remains in a lowered position and that instead of a velic closure for a plosive there is a break in the type of phonation from voice to creak and possibly back to voice.

The starting point of this investigation is the occurrence of the notations /ptkbdg/-/Qq/ (in modified SAMPA transcription [2,3]) in the Kiel Corpus of Spontaneous Speech [4]. Here /Q/ refers to a segmentally delimitable glottal stop, /q/ to the componential feature of glottalization, i.e. an irregular vocal fold vibration of low frequency. "-" symbolizes replacement, "-:" deletion, "=-" insertion of a symbol in relation to a canonical citation form pronunciation of a lexical item. Thus, e.g., /t-q/ indicates the replacement of a plosive by glottalization. However, in keeping with its componential reference, it is not given a duration, but only put on the same time mark as the following segment (for further details see [2,3]). The data will be broken down into three classes: instances in word-final '/@/ + sonorant' syllables, in non-final positions and across word boundaries. These groupings will be supplemented by an enquiry into other phonetic processes for the canonical transcription sequences of 'nasal + /ptkbdg/ + /@/ + nasal', namely nasalization of plosives and the addition of other voice qualities than creak.

GLOTTALIZATION FOR CANONICAL PLOSIVES

The label files and variants lexica of the corpus [4] provide 179 instances either of plosives being replaced by a glottal stop, or of neighbouring nasals and laterals being characterized by glottalization. The latter process is by far the most common. Of this total

- 156 apply to word-final syllables of the type 'plosive + /@n/', e.g. *können*, *halten*, *einverstanden*, *hatten*,
- 6 occur word-internally, viz. *eigentlich* (/Qq'aIg-:@-:=-~n-:t-qlIC/, with glotalization in addition to the replacement of canonical /g@n/ by nasalization of the preceding diphthong, and /Q-:q'aIg-N@-:n-:t-qlIC/, with nasalization of /g/ and glottalization of the sonorants), *schrecklich*, *bedeutendsten* (/b@d'OYt-q@-:nt-:sth@n/ and /b@dh'OYt-Q@-:nt-:st-:@-:n/), *besonders* (/b@z'Ond-q6s/),
- 17 across word boundaries, e.g. *bedanke mich* /bh@d'aNk-q@-: mIC+/, *und dann* /Q-:U-:nt-q+ d-na=-~n-:+/.

Of the 156 word-final syllable glottalizations

- the vast majority, 127, are tied to the canonical structures 'nasal/lateral + plosive + /@/ + nasal'
- only 29 occur without a preceding 'nasal/lateral'; they are all of the type /t@n/.

Of the 17 glottalizations across word boundaries

- 13 involve a nasal at the beginning of the following word, also including nasalized initial lenis plosives, as in *und dann* with continuous nasality all through the two words,
- 3 occur before the approximants /v/ and /j/;
- in one case final /t/ is realized as /Q/ before initial /d/.

In 5 instances of word-final /nt@n/ the sequence becomes monosyllabic, with glottalization running into an initial nasal (resulting from assimilated /d/, /v/) or vowel of the next word, e.g. *zehnten das* /ts'e:nt-Q@-:n-: d-nas+/, *könnten wir uns* /kh9=-~n-:t-q@-:n-m+ v-mi:6-:+ Q-:Uns+/.

THE FRAME 'SONORANT + PLOSIVE + /@/ + SONORANT'

There are 237 instances of this structure in the corpus with only 5 examples of /l/ in the second 'sonorant' slot (the proper name *Schindel*) and 43 in the first (all /lt/, *gelten*, *halten* and compounds, *sollten*, *wollten*). /@/ is kept in only 5 cases; 3 occur in the second of two successive /@n/ syllables: in *folgenden*, either realized as /g@-:n-Nd-n@n/, or as /g-N@-:n-:d-n@n/ with lenis plosive nasalization and nasality throughout the last two syllables, but with an oral opening in the second. Of this total of 232 /@/ elisions

- 127 have glottalization or glottal stop for the plosive,
- 105 either have a 'plosive-nasal/lateral' sequence or, in 9 cases, plosive nasalization; the latter occurs in one lenis plosive structure - *einverstanden* (2) - and in 7 fortis ones, all of which are either in unstressed syllables or non-accented function words - *fünfzehnten* (2), *sechzehnten*, *siebzehnten*, *achtzehnten*, *könnten*, *sollten*.

Of the 127 glottalizations

- 114 involve fortis stops,
- 13 lenis ones.

SIGNAL ASPECTS OF GLOTTALIZATION

The time course of glottalization and its synchronization with supraglottal articulation in the frame 'nasal + plosive + nasal' varies a great deal. Figure 1 illustrates 6 different realizations in the word *könnten*.

(a) comes closest to a segmental sequence of 'voice - creak - voice', corresponding to two voiced nasals, separated by creak for a stretch in time where the plosive is located in the canonical structure. But this case is still treated componentially in view of a consistent treatment of the variability encountered in (a) - (f), and because the auditory syllable boundary occurs at the beginning of the 3rd creak pulse. The SAMPA notation is /kh9nt-q@-:n+/; /t-q/ and /n+/ have the same time, at the syllable boundary.

(b) and (c) represent two-phase cases with either voice or creak coming first. They are disyllabic; in (b) the second syllable begins after the 1st creak pulse, in (c) with the 3rd, according to auditory assessment. There is strong nasalization of the preceding vowel. The SAMPA notation and the alignment are the same as for (a) in both cases. The vowel nasalization being treated as conditioned by /n/ is not marked.

(d) takes us one step further from (c): it is auditorily monosyllabic, the vowel is absorbed in the aspiration of the word-initial plosive, and the final voiced nasal is very weak before the labiodental /v/ of *wir* (which is also shown in the spectrogram, fused with *uns*). The two creak pulses are sufficient to signal the word *könnten*, rather than *können* to the listener. The SAMPA notation is /kh9n-:t-q@-:n+/; /t-q/ and /n+/ have the same time, at the 1st creak pulse.

(e) is one-phased, with creak only, but it is auditorily disyllabic, the boundary occurs at the 4th creak pulse. The SAMPA notation is /kh9nt-q@-:n+/; /t-q/ and /n+/ have the same time, at the boundary.

(f) is again one-phased, but auditorily monosyllabic, with strong nasalization of the preceding vowel, which replaces a missing voiced /n/. The SAMPA notation is /kh9=-~n-:t-q@-:n+/; /t-q/ and /n+/ have the same time, at the beginning of glottalization (the spectrogram also shows the following *wir un(s)*). In this

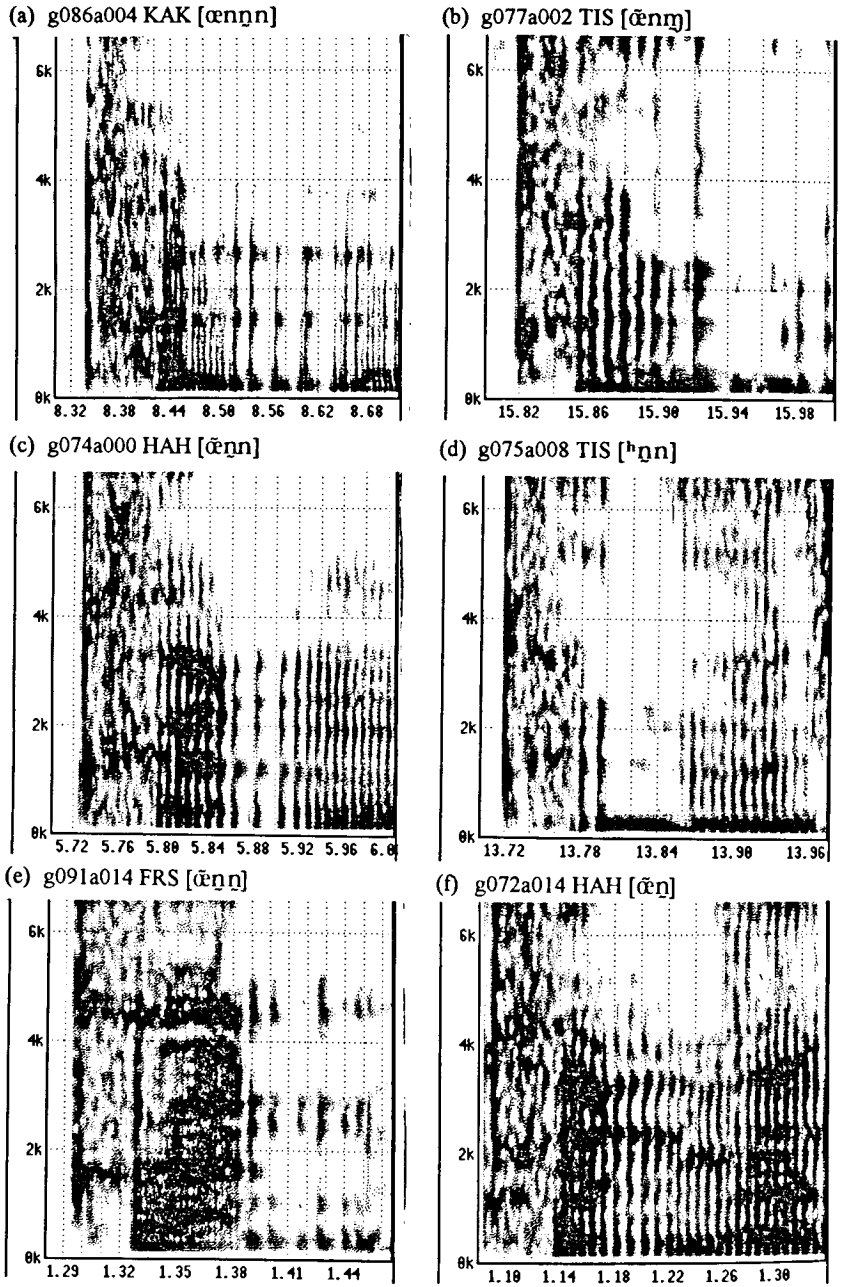


Figure 1. Different manifestations of glottalaziation in "könnten" from the Kiel Corpus of Spontaneous Speech. 4 speakers, 1 female (FRS)

case the nasalization of the vowel has to be marked specially by the insertion of /=~/ because it would otherwise not be implied by the notation.

PHONETIC EXPLANATION

All the glottalizations in the 'nasal-plosive-nasal' frame can be subsumed under the principle of economy of effort: the velum remains lowered during the whole sequence, and further adjustments to ensure a differentiation between the presence or absence of stop articulation in, e.g., *könnten* and *können*, are transferred to the glottis for the signalling of a break in the nasality feature. As long as this break is there its timing can be quite variable, ranging from the vowel to the final nasal, spanning syllable and word boundaries. This phonetic process represents a gestural reorganization and is extremely common in spontaneous German speech since more than half the potential cases in the corpus exhibit it. It seems to affect lenis and fortis plosives alike, although the frequency of potential frames for the former is very much reduced in the data base. The phonetic exponency is comparable to the Danish stød [5], even if the historical genesis is quite different.

The timing of velum lowering is variable as well: it may precede, or more or less coincide with, the formation of the oral occlusion for the sequence. In the case of lenis plosives there is also the possibility of velum lowering for the complete sequence without glottalization, i.e. a simple change in the timing of the velic gestures. The shorter closure phase, compared with fortis plosives [6], supports this process. If fortis plosives are shortened in unstressed position they can go the same way. Again the corpus is not big enough for a more exhaustive analysis of this aspect. Similarly, the question as to whether, in glottalization, there is complete coalescence of fortis and lenis, e.g. in the timing parameters of vowel and sonorant durations, cannot be answered by this data base because of its restricted size and contextual heterogeneity.

A break in the sequence can also be effected in a different way, which this corpus, however, does not provide any examples of. For fortis stops the glottis may open, which together with a lowering of the velum during the entire sequence results in a voiceless nasal. In the case of lenis stops the vocal fold vibration may change to breathy voice, which combined with velic lowering produces a voiced but breathy nasal as a break within the continuous nasality. These features have been noticed in such words as *gebunden*, *vierzehnten*, with the latter having either voiceless or, because of the unstressed position, voiced nasal breath. The feature of breathy voice replacing lenis stops in a nasal environment can again not be explained by simple changes in synchronization: it requires gestural reorganization, just like glottalization, for the function of creating a break in the nasal stream for a listener.

REFERENCES

[1] Kohler, K. J. (1994), "Glottal stops and glottalization in German. Data and theory of connected speech processes", *Phonetica*, vol. 51, pp 38-51.

[2] Kohler, K.J., *Lexica of the Kiel PHONDAT Corpus. Read Speech*, vols.I, II, *AIPUK 27, 28*, Kiel: IPDS.

[3] Kohler, K.J., Pätzold, M., Simpson, A. (1994), *Handbuch zur Segmentation und Etikettierung von Spontansprache - 2.3.* VERBMOBIL Technisches Dokument Nr. 16, Kiel: IPDS.

[4] IPDS (1995), *CD-ROM#2: The Kiel Corpus of Spontaneous Speech*, vol. I, Kiel: IPDS.

[5] Fischer-Jørgensen, E. (1989), "Phonetic analysis of the stød in Standard Danish", *Phonetica*, vol. 46, pp. 1-59

[6] Kohler, K. J. (1984), "Phonetic explanation in phonology: the feature fortis/lenis", *Phonetica*, vol. 41, pp. 150-174.

VARIATION IN SCHWA + /r/ IN GERMAN

W. J. Barry,
Institut of Phonetics University of the Saarland, Germany

ABSTRACT

The possible psychological reality of an analysis of the German vowel [ɐ] as /ə/+/r/ is examined. Firstly, the sensitivity of [ɐ] to contextual factors is compared to [ə]. Secondly, the vowelised realisation of assumed /ə/+/r/ sequences is examined for a dialect with an apical /r/ variant. The plausibility of interpreting [ɐ] as a vocalic variant of /r/ in terms of reduced articulatory gestures is considered in the light of the results.

INTRODUCTION

The German vowel [ɐ] can be understood as the phonetic realisation of an underlying phonological segmental sequence /ə/+/r/. Examination of the acoustic structure of uvular variants of /r/ point to an articulatory continuum ranging from a uvular fricative [ʁ] to a half-open, central-to-back vocoid. Thus [ɐ] in consonant+[ɐ] sequences (e.g. Kupfer - [kʊpfɐ], bitter - [bɪtɐ], Bäcker - [bɛkɐ]) may be analysed as the syllabic equivalent to the nonsyllabic off-glide [ɐ] found in vowel + vowelised /r/ variants (e.g. Bier - [bi:ɐ], Kur - [ku:ɐ]). In terms of the phonological representation, the same modifications in the production processes may be evoked as for [ə]+nasal or [ə]+lateral realised as syllabic nasal (e.g. bitten - [bɪtn̩], Schuppen - [ʃʊpm̩], backen - [bakŋ̍]) and syllabic lateral (Mittel - [mɪtl̩]) respectively. That is, in terms of segmental structure, the schwa is elided, and the sonorant takes over the syllabic function. The vocalic nature of the resulting sonorant in the case of /r/ parallels the vowelised /l/ in some varieties of British English (e.g. bottle - [bɔtʊ], milk - [mɪʊk]) cf. [1].

While, articulatorily, the alternation between a contoid and a vocoid realisation of the underlying liquid consonant is easily explained as a case of target undershoot, a feature-based phonological representation is stuck with an unmotivated alternation of the general class feature [consonant]. A gestural phonological account, on the other hand [2, 3], captures the variation as a phonetic continuum from fricative (or trill) to vocoid depending on the degree of overlap between the vowel- and /r/-gesture (both being tongue-body gestures). If a part of the task of phonological theory is to capture the sound structure of linguistic signs in a manner which can be plausibly related to their production, i.e. to the underlying articulatory plan *for*, the articulatory patterns involved *in*, and the acoustic forms resulting *from* their use in speech, then a gestural approach would appear more adequate in this case at least.

If we exploit the presence of two free and articulatorily radically different variants of German /r/ (uvular and apical), a gestural phonological account allows a number of predictions to be made relating to the general interpretation of German [ɐ] as a vowelised realisation of a phonologically real /r/ (in the above sense) rather than a second unstressed vowel derived historically from post-vocalic /r/[1]:

1. In a speaker with an apical rather than a uvular /r/, the quality of the vocalic realisation of the assumed /ər/ sequence should relate differently to /ə/, since the retraction gesture of the tongue body towards the uvular target is replaced by a tongue-tip gesture which has much less effect on the tongue body.

2. As the phonetic reflex of a constant consonantal target overlapping a preceding schwa, the [ɐ] should vary less in equivalent flanking vowel conditions than schwa alone in that position. However, this expected difference in variance should be much more marked with uvular /r/ than apical /r/ speakers since the constant tongue-tip gesture constrains the tongue body less in its move from the preceding full vowel via (underlying) schwa to the following full vowel.

3. In a /ə/#/r/ sequence, the quality of /ə/ should approach that of [ɐ], presumably lying between /ə/ and [ɐ] as a result of reduced articulatory overlap between the final /ə/ and the initial /r/. In [ɐ]#/r/ sequences, the effect of the following initial /r/ should be less.

The following data analysis aims to address these predicitons.

[1] Of course, morphophonological alternations such as *unser - unsere*: [ʊnzɐ] - [ʊnzərə] argue at a different level for the underlying /r/ interpretation.

EXPERIMENT

Two native speakers (1M. 1F) of standard German, one (GF, M) with a uvular /r/ and a slight North German accent, the other (JB, F) with an apical /r/ and a mild West Bavarian accent, recorded the following corpus under quiet studio conditions:

1. Short sentences in the form of article + trochaic noun + trochaic verb-form, or pronoun + verb + noun. The first lexical form contained /i:/, /u:/ or /a:/ and ended with /tə/ or /tər/ (realised as [tɐ]); the second lexical form had either an initial bilabial stop or an initial /r/ followed by /i:/, /u:/ or /a/. e.g.:

Ich biete Pasta Der Dieter pustet
Ich biete Ruten Der Dieter ruhte

The 36 items (2 unstressed vowels x 2 initial consonant classes x 3 vowels x 3 vowels) were read 5 times in quasi-randomized order (180 tokens per speaker) with both lexical items accented. These are referred to as "context" items.

This condition was selected to reveal the degree of spectral variation in schwa and schwa+/r/ as a product of the flanking extreme vowels /i:, a(:), u:/.

2. Three-syllable phonological words (lexical words or minimal syntagms) stressed on the second syllable with either first-syllable /ə/ or /ər/ or third-syllable /ə/ or /ər/. These are referred to as "pretonic" and "post-tonic" schwa items, respectively. E.g:

Ich picke, Der Dieter
Gebieter, Verbieten

For the pretonic items 15 words were selected for each category to cover the stressable monophthong phonemes of German; each item was read three times in quasi-random order (2 schwas x 3 repetitions x 15 vowels = 90 tokens per speaker). For the post-tonic items, 90 words were selected to give each stressed vowel a bilabial, an alveolar and a velar stop as postvocalic context (2 schwas x 3 consonants x 15 vowels = 90 items per speaker).

These data were collected to provide a stressed-vowel frame of reference for the two speakers, within which to locate the quality of the schwa and schwa+/r/ realisations. They also provide further data towards a definition of the unstressed vowel qualities under two different positional and segmental context conditions.

The recordings were digitised at 10 kHz using the PC-based Kay Computer Speech Lab (CSL) facilities. Duration measures were made on the sound-pressure waveform linked to a broad-band (293 Hz) digital spetrogram. Formants were measured on a 12-pole LPC spectrum calculated over a 25ms window (reduced to 15 ms for very short schwa realisations) located in the middle of the segment.

RESULTS

Figures 1a and 1b show the relative positions of the unstressed and stressed

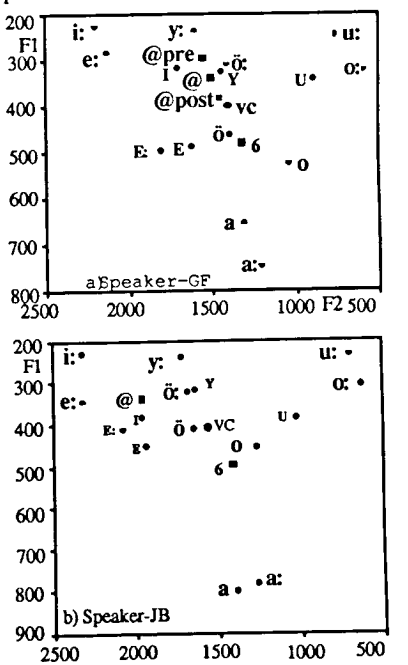


Fig. 1 Schwa values in relation to stressed vowels and centroid

vowels on an F1/F2 vowel chart for the combined pre- and post-tonic condition.

Let us consider the first prediction derived from a gestural phonological approach. The North German speaker (GF), with uvular /ʀ/, has an average /ə/ value which is considerably closer and slightly more "fronted" than the centroid of the stressed vowels (mean F1, mean F2). This is almost completely the product of the very short (therefore less open) pre-tonic /ə/ realisations. Post-tonically, the /ə/ is very close to the centroid value, conforming to the assumption that the schwa is phonologically targetless and therefore tends towards the relaxation position of the vowel articulators (tongue body, jaw and lips) [4, 5]. The average [ɐ] value conforms to the pattern found in a previous analysis of a standard German speaker with uvular /ʀ/ [5], lying centrally between /ɛ/ and /ɔ/ and could be plausibly attributed to merging the neutral vocalic element with a retraction gesture of the tongue body in the direction of a uvular target.

The Bavarian speaker (JB), on the other hand, has an extremely fronted /ə/, very close to /ɪ/ and well separated from the stressed vowel centroid. Her average value for [ɐ], however, is in a very similar position to that of the North German speaker, relative to her other stressed vowels, namely midway between (and slightly more open than) /ɛ/ and /ɔ/. In both cases, these data call for a different explanation from the one offered for standard German. On the one hand they suggest a definite target for /ə/ rather than a phonologically unspecified relaxation target. Auditorily, this is acceptable, since the unstressed <e> in Bavarian German in no way evokes the impression of a neutral central vowel.

On the other hand, the [ɐ] cannot be explained as an articulatory merger of schwa and /ʀ/, since there is nothing in the apical /ʀ/ gesture which would drag the tongue body away from the fronted, closer position. Here again, it would seem that JB's [ɐ] vowel, in contrast to GF, has a definite vocalic target.

If this interpretation is correct, there should also be a clear difference in the pattern of variability between the two speakers. According to prediction 2, the flanking vowels should exercise maximum influence on the phonologically undefined /ə/ tokens, but should be inhibited by the underlying /ʀ/ element in [ɐ] in the case of GF. Speaker JB, on the other hand, should have equal variability for /ə/ and [ɐ], since, according to the above data, they both appear to have a phonologically defined target.

Comparison of GF's /ə/ and [ɐ] in the context condition with following labial consonant (see fig 2a, each point represents 5 values for a given context condition) shows that under an identical set of context conditions, /ə/ varies considerably more than [ɐ] (F1: F = 2.51; F2: F = 3.17, in both cases df 89/89, and $p < 0.001$). JB does have different variability in F1 (F = 3.36, df 89/89, $p < 0.001$, see fig 2b), but it is [ɐ] which varies more; F2 variance does not appear to differ (F = 1.53, df 89/89, $p > 0.05$).

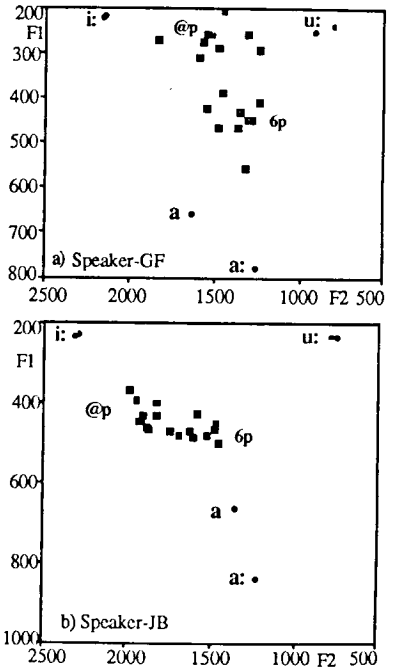


Fig 2 Vowel-context sensitivity of /ə/ and[ɐ] before bilabial consonant

Finally, a comparison of the influence of the two post-schwa(+/ʀ/) consonants (bilabial plosive and /ʀ/) on the unstressed vowel quality in the context condition

provides additional evidence in the question of a regional difference in the phonological status of /ə/ and [ɐ]. Figures 3a and 3b show the corner-vowel values and the pre-labial vs. pre-/ʀ/ values for /ə/ and [ɐ] in the context condition.

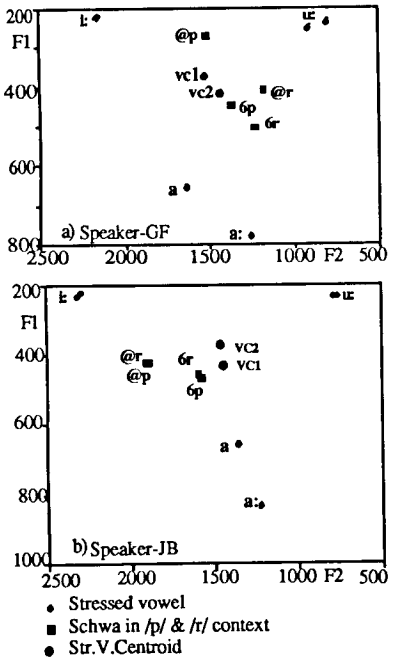


Fig. 3 Sensitivity of /ə/ and [ɐ] to labial and /ʀ/ post-context

Speaker GF shows a massive effect of the post-schwa /ʀ/-context; F1 increases and F2 decreases in comparison to /ə/ followed by /b/ or /p/ (one-way ANOVA, F1: F = 91.9; F2: F = 83.4; in both cases df 89, $p < 0.0001$). In other words, the same shift is observed in /ə/ before /ʀ/ as is found between /ə/ and [ɐ] in non-/ʀ/ contexts. A similar though smaller shift (but still highly significant F1: F = 18.1; F2: F = 31.8, df 89, $p < 0.001$) is observed for [ɐ] between the labial- and in the /ʀ/-context. This may be seen as an augmentation of the shift resulting from the effect of the assumed /ʀ/ behind the [ɐ] vowel.

Speaker JB, on the other hand, shows no contextual effects whatsoever for either /ə/ or [ɐ], indicating further that the difference between /ə/ and the [ɐ] vowel has nothing to do with an underlying /ʀ/ element (/ə/-F1: F = 0.00008, [ɐ]-F1: F = 0.52; /ə/-F2: F = 0.41, [ɐ]-F2: F = 0.41; df 89, $p > 0.1$ in all cases).

CONCLUSION

In the light of the results of the present analysis, we find support in the production patterns of speaker GF for the assumption that [ɐ] is represented as /əʀ/ in his articulatory plans. Firstly, variance for [ɐ] is less than for /ə/, indicating the "constraining" effect of an overlapping consonantal element; secondly, a surface /ʀ/ following /ə/ changes its quality massively in the direction of [ɐ].

For speaker JB, on the other hand, it would appear that [ɐ] is a separately encoded vocalic element, since it has a quality, relative to the stressed vowel system which is similar to the [ɐ] of a speaker with a uvular /ʀ/ and can therefore not be considered a merger of overlapping /ə/ and /ʀ/. It is seen that a following surface /ʀ/ (apical) has no appreciable effect on the quality of either /ə/ or [ɐ].

Finally, there is clear evidence that speaker JB has an established target quality for /ə/, whereas, at least for the durationally unconstrained post-tonic schwa, GF reveals a quality very close to the centroid of the stressed vowels, supporting the theory that the quality of /ə/ is phonologically undefined

REFERENCES

[1] Hardcastle, W.J. & Barry, W.J. (1989): Articulatory and perceptual factors in /l/ vocalisations in English. *J. Int. Phonetics Assoc.* 15(2), 3-17

[2] Browman, C.P. & Goldstein, L. (1992a): "Targetless" schwa: an articulatory analysis. In: G. J. Docherty and D. R. Ladd (eds.), *Gesture, Segment, Prosody. Papers in Laboratory Phonology II.* Chapter 2, pp. 26-55. Cambridge: University Press.

[3] Browman, C.P. & Goldstein, L. (1992b): Articulatory Phonology: An Overview. *Phonetica* 49, 155-180.

[4] Barry, W.J. (1992) Comments on Chapter 2: "Targetless" schwa: an articulatory analysis. In: Docherty and Ladd (eds.), *Gesture, Segment, Prosody. Papers in Laboratory Phono-logy II.* 65-67, Cambridge: University Press.

[5] Barry, W.J. (1995) Schwa vs. schwa + /ʀ/ in German. *Phonetica* 52

EFFECTS OF PROSODIC CONSTITUENT LENGTH ON PAUSE REALIZATION

Ludmila Menert
OTS (Research Institute for Languge and Speech),
Trans 10, 3512 JK Utrecht, The Netherlands

ABSTRACT

Does prosodic weight in the form of constituent-length affect restructuring of prosodic phrases, as predicted by the theory of Prosodic Phonology (Nespor & Vogel 1986)? As no evidence of such effect could be found in experiments on a Dutch sandhi process, additional testing involved realization of speech pauses in the same phonological environment.

INTRODUCTION

In previous experiments (Menert 1994) we investigated the effect of speech rate and prosodic constituent length on sandhi-processes in Dutch (voice assimilation, consonant degemination). The effect was expected to be mediated through restructuring of prosodic application domains of sandhi rules, which is typically rate and length dependent (cf. Nespor & Vogel 1986). Specifically, I assumed that voice assimilation occurs more frequently inside phi-domains than across phi-boundaries. According to N&V, phi-domains may be extended (allowing more assimilation) in case: a) speech rate is high; and b) the restructuring phis are relatively short.

Method

I expected to find evidence for the above assumptions in a series of experiments focussing on speech rate and voice assimilation. In the sentences used as stimulus material the segmental context allowing for assimilation was located in a fixed position, which was a potential phi-phrase boundary according to the N&V rules. The length of the last constituent was systematically varied, in order to vary the restructuring conditions for phi-phrases, (again following N&V), as is illustrated in (1) below:

(1) a. [Jan]$_{PHI}$[vond]$_{PHI}$[het paard]$_{PHI}$ [dapper]$_{PHI}$
'John considerd the horse brave'
b. [Jan]$_{PHI}$ [vond]$_{PHI}$ [het paard]$_{PHI}$ [danig dapper]$_{PHI}$
'John considerd the horse pretty brave'

According to N&V the branching constituent in b. blocks restructuring of the last two phis. If voice assimilation is a phi-bound process, as I assumed here, significantly less assimilations should occur in b. compared to a. Phi-restructuring and thereby wider occurrence of assimilation can be expected in fast speech rate compared to slow or normal rates.

Thirty sentence pairs were read aloud by five speakers at three speech rates (slow, normal, fast). These sentences were similar to those in (1), except for the first constituent, which contained more lexical material, as illustrated in (2).

(2) a. [Maar doktersassistente Ellen]$_{PHI}$ [vond]$_{PHI}$ [het paard]$_{PHI}$ [dapper]$_{PHI}$
'but medical receptionist Ellen considered the horse brave'
b. [Maar doktersassistente Ellen]$_{PHI}$ [vond]$_{PHI}$ [het paard]$_{PHI}$ [danig dapper]$_{PHI}$
'but medical receptionist Ellen considered the horse pretty brave'

The speech was recorded and served as material in perception experiments, in which assimilation occurrence was established by means of perceptual ambiguity of the stimuli.

The expectation was not borne out; while a rate effect on sandhi exists, there proved to be no effect of constituent length on sandhi occurrence. However, this does not yet entirely preclude a length effect on prosodic restructuring, since we may hypothesize that the relevant sandhi processes are relatively insensitive to phi-boundaries.

PAUSE REALIZATION

Introduction

An effect of constituent length on prosodic structuring in Dutch phrases can still be expected. This expectation is based on intuition and on literature, both regarding Dutch and other languages. Evidence for the psychological reality of phrase structure has been described in research literature since the 1960s. Later is has been shown, that the correspondence between the formal linguistic (syntactic) structure and the 'performance structure' (the phonetic organization of spoken material) is not always perfect. This was demonstrated by Grosjean et al. in studies of speech pause duration and distribution (Grosjean, Grosjean & Lane 1979, Grosjean 1980). In a series of production and parsing experiments Grosjean et al. found that pause placement and duration are affected not only by syntactic boundary strength (measured by depth of embedding in the syntactic tree), but also by the length of constituents. By checking their speech production findings in parsing tasks, the results of which were compared with predictions based on the production data, they established that sensitivity to constituent length is part of the internal phonological rule system of native speakers of English.

Vanderslice (1968) makes predictions regarding pauses on the basis of constituent length in his discussion of what I would call intuitive prosody. As for Dutch, there is evidence of such predicted length effects in studies by De Rooij (1979): effects of constituent length are found on realized phrase boundaries marked by intonation.

Method

I have attempted to establish a direct influence of constituent length on prosodic restructuring. The test was intended to score the realizations by the speaker of prosodic boundaries by means of the insertion of a speech pause. The stimulus material used in the sandhi experiments was inspected again, and speech pauses and prepausal lengthening were scored in different rate and constituent length conditions. For this purpose all the original recordings of the five speakers involved in the first experiment were listened to carefully, by myself, a trained phonetician, and a trained phonologist. No measurements were carried out, the occurrence of a pause was decided solely on the basis of auditory asessment. It is therefore very well possible that some instances of what was judged as a pause were in fact cases of perceptually conspicuous lengthening without a real pause.

If, in line with the assumptions and predictions presented in the introduction, the phi-prase restructuring is sensitive to constituent length, more pauses can be expected in the sentences with a branching constituent (as in b. in the above examples) compared to the non-branching ones (as in a. in the examples).

RESULTS

The results of the auditory examination are presented in Table 1.

Table 1. Number of speech pauses realized at the assimilation site in each experimental condition (N = 60 in each cell).

	slow	
	branching	non-branch.
speaker 1	22	31
speaker 2	48	51
speaker 3	22	13
speaker 4	59	56
speaker 5	21	21
mean %	57	57

	normal	
	branching	non-branch.
speaker 1	15	13
speaker 2	20	15
speaker 3	6	0
speaker 4	45	26
speaker 5	3	6
mean %	30	20

The results indicate that some sort of prosodic boundary is realized on the potential phi-boundary: in 57% of the slow rate realizations and in 25% of the normal rate realizations a speech pause was inserted at the predicted phi-boundary in my material. No speech pauses were perceived in the fast speech rate.

The data in Table 1 show that pause placement is not influenced by constituent branching in the stimulus sentence. Only in the normal speech rate condition are the mean pausing percentages slightly higher, but when the individual pausing behaviour is taken into account, only speaker 4 seems to show consistency with respect to branching (for detailed data cf. Menert 1994). This implies that the expected effect (i.e. more frequent pause insertions on boundaries of branching phi-phrases) did not occur, contradicting the theoretical claims in N&V.

Table 2. Number of speech pauses realized in other positions in the sentence, when no pause was realized at the assimilation site, total for both branching conditions (N=120 in each cell).

	slow	normal
speaker 1	16	14
speaker 2	33	58
speaker 3	16	22
speaker 4	6	33
speaker 5	53	23
mean %	21	25

It is obvious from the data that in the slow rate condition more pauses are inserted by the speakers than in the normal rate condition. This is not surprising; more interesting is the observation that the predicted spot for a phi-boundary in fact proves to be a natural position in the sentence for the speaker to insert a speech pause. This can be seen in Table 2: when just one speech pause is inserted, only in about 25% of all realizations the speakers do so in other positions in the sentence than at the predicted phi-boundary. I observed that the position directly preceding the verb was the other favorite for pause insertion. This position is also a predicted phi-boundary in a N&V analysis. As can be seen in the tables, the amount of pausing is very individiual. It can be concluded, however, that the preferred positions for pausing are more or less the same for all speakers (with different rankings of candidates, though), and that they co-occur with phi-phrase boudaries as predicted by the N&V rules.

Additional observation which results from the auditory examination is that the speech pauses were only very rarely combined with a pitch movement. If there was one, it was not a boundary marking tune, as described in IPO-system for Dutch intonation grammar ('t Hart, Collier & Cohen 1990).

The results can be summarized as follows:

- If a speaker of Dutch realizes a phrase boundary within the span of an intonation contour, that is in N&V terms, under the I-phrase level, he does so by means of a speech pause combined with syllable lengthening and no melodic marking (no pitch movement);
- the number of speech pauses of this sort realized by speakers decreases with higer speech rates;
- the realizations of such speech pauses are significantly more frequent om the hypothetical phi-boundary compared to other word-boundaries in the sentence.

DISCUSSION

The results can be interpreted as evidence for realization of phi-phrases. This supports my original assumptions regarding the prosodic structures of the stimulus sentences, particularly the existence of presumed phi-phrases in my spoken material. Similar observations of temporal adjustments, such as pauses and prepausal lengthening, that do not co-occur with an intonation boundary (e.g., marked by a pitch reset) are reported by Blaauw (1994), Bringmann (1991) and De Rooij (1979). These adjustments could be interpreted as markers of the Dutch phi-phrase, an intermediate level between the Intonational phrase and the Clitic group. However, if the realized pauses do demarcate the Dutch phi-phrase, I will have to conclude that these phi-phrases are not sensitive to to constituent branching, and that they do not restructure in the way predicted by the N&V rules.

An additional check of melodic markings on the presumed phi-boundaries in the stimulus material will be done, so as to confirm this result.

As for the problem of the (non)existence of the constituent length effect, there are, in my opinion, two possible areas of further investigation that should be explored before the idea of length-effects in Dutch is dismissed altogether. The first possibility is that such effects can be found in Dutch, if only the length variation is extreme enough. The second possible approach is to test phis at other positions in the sentences than I have done sofar. The choice for placing the supposedly restructuring phi at the end of the sentence was motivated by reasons concerning the application of the N&V phi-formation rules in Dutch, but it could have had some undesirable consequences. It is possible that the realization of the prosodic features towards the end of an utterance may differ from what can be predicted on the basis of of the prosodic behavior throughout the utterance. I would expect there to be a greater chance of finding phi-restructuring triggered by length when the required phi-structure is placed somewhere in the middle of a sentence or utterance.

REFERENCES

[1] Blaauw, E. (1994), The contribution of prosodic boundary markers to the perceptual difference between read and spontaneous speech, *Speech Communication* 14, pp. 359-375.

[2] Bringmann, E. (1991), *The distribution of Dutch reading pauses, the influence of prosodic phrases and punctuation on pause location and duration*, MA thesis, Utrecht University.

[3] 't Hart, J., R. Collier & A. Cohen (1990), *A Perceptual study of intonation: an experimental-phonetic approach to speech melody*, Cambridge University Press.

[3] Grosjean, F., Grosjean, L. & H. Lane (1979), The patterns of silence: performance structures in sentence production, *Cognitive Psychology* 11, pp. 58-81.

[4] Grosjean, F. (1980), Linguistic structures and Performance structures studies in pause distribution, *Temporal Variables in speech: Studies in Honour of Frieda Goldman-Eisler*, the Hague, Mouton, pp. 92-106.

[5] Menert, L. (1994), *Experiments on voice assimilation in Dutch: prosodic structures and tempo*, Dissertation, Utrecht University.

[6] Nespor, M. & I. Vogel (1986), *Prosodic Phonology*, Dordrecht, Foris Publications.

[7] Rooij, J.J. de (1979) *Speech punctuation, an acoustic and perceptual study of some aspects of speech prosody in Dutch*, Diss. Utrecht University.

[8] Vanderslice, R. (1968), Synthetic Elocution, *Working papers in phonetics* 8, University of California, Los Angeles.

HOW DETERMINABLE ARE INTONATION UNITS?

*A. Wichmann and *G. Knowles*
Department of Cultural Studies, University of Central Lancashire, Preston PR1 2HE
**Department of Linguistics, Lancaster University, LA1 4YT*

ABSTRACT

This paper examines some of the practical and theoretical issues surrounding the assignment of intonation units in the transcription of speech corpora. We suggest that current practice may obscure some aspects of the prosodic system, and that there is a need for caution when using such data.

INTRODUCTION

There is a widespread assumption that spoken texts can be segmented into discrete prosodic units all of the same status, and known variously as *intonation units (IU)*, *tone units*, or *tone groups*. This assumption has been applied in the annotation of spoken corpora including the London-Lund Corpus (LLC) [1], and the Lancaster/IBM Spoken English Corpus, (SEC) [2]. It has most recently been applied to the Corpus of Spoken American English, (CSAE) [3]. While there are eminently good practical reasons for this approach, the resulting transcription is open to misinterpretation, especially when it is used as data by non-specialists. There are two main considerations here: the accuracy of auditory transcriptions, and the theoretical assumptions which underlie them.

THE RELIABILITY OF TRANSCRIPTIONS

Transcriber accuracy

The concept of 'accuracy' in prosodic transcription begs many questions, not least because they are based on auditory perceptions. Human sensory perception relates to physical reality in complex ways, and is rarely discussed in terms of 'correctness', since this would presuppose that the physically measurable phenomenon in each case is criterial. It is more appropriate to consider the **consistency** of transcription across two dimensions. The first is consistency across transcribers, and the second is consistency within an agreed system of intonation.

The transcribers of the corpora mentioned above have undertaken various measures to ensure consistency across transcribers. The 50,000 words of the SEC were analysed by two transcribers working independently, but overlapping for about 9% of the corpus [2]. The CSAE is a much larger undertaking (200,000 words) and consequently employs many more transcribers. Consistency is sought first by intensive training of the transcribers and secondly by joint discussion of all doubtful cases [3]. In each case the transcribers are highly trained, and it is fair to assume that a high degree of internal consistency has been reached. With this degree of expertise it is possible to be relatively confident that the published spoken materials have been transcribed consistently, although this still does not make them necessarily comparable with one another.

Theoretical consistency

A more difficult issue than the reliability of transcribers is the reliability of the system itself. The biggest problem is to define the units of the system [5]. Phonological accounts differ, and different units are posited on different theoretical assumptions, see e.g. [6], [7], [8], [9], and [3]. British approaches to the IU are traditionally based on a structural definition in terms of onset or head, nucleus and tail. Other approaches assume a demarcative definition, in which the unit is defined in terms of its boundaries. Frequently adopted criteria are pause [10] and pitch contours [11].

Studies of the phonetic attributes of transcribed boundaries provide evidence for a wider range of attributes, some segmental, such as absence of assimilation and elision, others prosodic, such as syllable lengthening, changes in tempo and changes in voice quality. [12] [13]. These attributes co-occur in different ways, signalling boundaries of varying strengths, from very obvious ones to those hardly discernable. The inability of transcribers to agree on some boundaries is not evidence of lack of expertise, but an inherent problem in a phonological theory which does not allow for indeterminacy. Indeed, our experience of the analysis of prosodic data increasingly indicates that it would be desirable to incorporate such information, i.e. cases of disputed boundaries, in the transcription.

A further issue is that these various cues to boundaries may in fact belong to separate prosodic systems. We would argue that a distinction needs to be made between temporal criteria (including pause and final lengthening) and melodic criteria.

Melodic units and temporal discontinuities

The melodic and temporal systems appear to operate in different ways. Melodic patterns coincide with segments of text which can be called **melodic units**; temporal patterns divide one segment of text off from another and constitute **discontinuities**. These would also appear to have separate functions, and we would tentatively suggest that melodic units relate to the structure of the text, while temporal discontinuities are chiefly concerned with interaction management.

Melodic units have a complex structure which can be best represented by some kind of transition network [9]. The domain of these units corresponds to structural units of text. When sentences and clauses are marked off, they are given one or more melodic units. Relative clauses and appositives are marked by the repetition of melodic fragments, a pattern which is sometimes called *tone harmony*. Coordinated items and lists are assigned melodic fragments which combine to form a complete unit at a higher level.

Temporal breaks can be inserted into a tune for several reasons. The most obvious case is the hesitation pause, related to the speech planning process, but pauses can also be motivated by the text. (In the following examples (•) indicates a pause). They are used in reported speech, e.g. *'Yes'* (•) *she said* where the break separates the quotation from the reporting phrase. They can draw attention to a following item, e.g. *these are called* (•) *'formants'*, or correspond to scare quotes in writing, e.g. *Mary is his* (•) *'friend'*. Skilled readers of verse may use a temporal

discontinuity within a continuing melodic unit to accommodate conflicting demands of speech rhythm and verse metre. The set of uses is not necessarily fixed, and speakers may invent idiosyncratic uses of their own.

The fitting of tunes to a text, and the insertion of temporal breaks, are of course not mutually exclusive. Indeed, temporal discontinuities commonly co-occur with the end of a tune. Consequently, a tune is prototypically - but not necessarily - bounded by temporal breaks. On the other hand there are many temporal discontinuities that are not in any way related to melodic units.

In practice, transcribers of speech corpora have used a wide range of mostly demarcative criteria to identify just one kind of unit (or two when a distinction is made between 'major' and 'minor' units). In other words a boundary has been marked variously as a response to pause, and to contour change, and to other prosodic cues.

CONCLUSION

Prosodically annotated speech corpora are valuable sources of data for the investigation of the prosodic system or systems. Such investigations will normally be based on the transcription in conjunction with the original sound recording. Corpora are, however, also intended to 'provide materials for extensive studies of all aspects of spoken English grammar and lexicon' [3]. This kind of research tends to rely on the transcription alone, taking the annotations as given.

While the original transcribers may be well aware of the different clusters of phonetic attributes that are annotated as boundaries, the fact that they are marked in the same way makes it difficult for subsequent users of the transcription to avoid treating IU boundaries as though they were all of equal phonological status. In reality, if the distinction made in this paper is valid, then it follows that some of the stretches of speech between marked boundaries are of no phonological status at all.

REFERENCES

[1] Svartvik, J. (ed) (1990) *The London-Lund Corpus of Spoken English: Description and Research*, Lund University Press, Lund.

[2] Knowles, G., Williams, B., Taylor, L. (eds) (to appear) *The Lancaster/IBM Spoken English Corpus: A Corpus of Formal British English Speech*, London: Longman.

[3] Chafe, W., Du Bois, J.W. & Thompson, S. A. (1991) "Towards a new corpus of spoken American English" in: Aijmer K. & Altenberg, B. (eds) *English Corpus Linguistics*. London: Longman.

[5] Couper-Kuhlen E. (1986) *An Introduction to English Prosody*, London: Edward Arnold.

[6] Crystal, D. (1969) *Prosodic Systems and Intonation in English*, Cambridge: Cambridge University Press.

[7] Pierrehumbert, J.B. (1980) "The phonology and phonetics of English intonation". PhD thesis, MIT.

[8] Ladd, D.R. (1986) "Intonational phrasing: the case for recursive prosodic

structure". *Phonology Yearbook* 3: 311-340.

[9] 't Hart, J., Collier, R. & Cohen, A. (1990) *A perceptual study of intonation*. Cambridge: University Press.

[10] Stenström, A-B. (1990) "Pauses in Monologue and Dialogue", in: [1].

[11] Brown, G. (1977) *Listening to Spoken English*, London: Longman.

[12] Knowles, G. (1991) "Prosodic labelling: the problem of tone group boundaries", in: Johansson, S. & Stenström, A-B. (eds) *English Computer Corpora: selected papers and research guide*. Berlin: Mouton de Gruyter.

[13] Chafe, W. (1995) "Adequacy, User-Friendliness, and practicality in Transcribing", in: Leech, G., Myers, G., Thomas, J. (eds) *Spoken English on Computer: Transcription, mark-up and application* Longman: London.

MULTIPULSE LPC MODELING OF ARTICULATORY MOVEMENTS: DETERMINATION OF MINIMUM PULSE SEQUENCES

Soumya BOUABANA and Shinji MAEDA

ENST - CNRS URA-820 - 46 Rue Barrault 75634 Paris 13, FRANCE

e-mail : soumya@sig.enst.fr

ABSTRACT

The frame-by-frame variations of tongue profiles derived from X-ray film data are described in terms of the temporal patterns of four articulatory parameters. The temporal variation of each parameter, i.e., movement, is assumed to be the output of a time-invariant auto-regressive filter. These filters are excited by a sequence of pulses, representing articulatory commands. The curve of synthesis error for each movement shows a rapid decrease up to the number of pulses corresponding to that of the syllables in the sentence and then the decreasing rate becomes distinctively slower. In this paper, the minimum number of pulses is determined by using acoustic criterion. It depends on the number of the phonetics features, in the sentence, of which their realization is related to particular parameters.

1. INTRODUCTION

Digitized lateral X-ray film data were used to monitor the temporal variation of tongue profiles in the mid-sagittal plane. The profiles are obtained by manually tracing, frame-by-frame, radio films shot at a rate of 50 frames per second during the production of 10 French sentences uttered by two female speakers. An articulatory model is derived as the result of a factor analysis on the measured tongue contours. In this model, the tongue profile is specified by one extrinsic parameter, jaw position **jw** (open/close), and three intrinsic parameters, tongue-body position **tp** (back/front), tongue-body shape **ts** (arched/flat) and tongue tip position **tt** (up/down). These four parameters suffice to specify the entire mid-sagittal tongue shape with reasonable accuracy, since they explain more than 90% of the variance of observed tongue profiles. The parameter values are calculated from each frame of the X-ray data.

The phonetics features of vowels F-patterns can be specified by the values of one or two dominant parameters. It appears, in a preliminary analysis of movements, that not the whole four parameters but only some selected parameters at time are involved in the production of a given phoneme. Our objective is to introduce a constraint for controlling individual articulator movements by using a simple source-model filter, in order to effectively describe the coordinated orchestration of the individual articulatory movements during sentences.

2. MODELING MOVEMENTS

Each articulatory movement is assumed to be the output of a time-invariant auto-regressive (LPC) filter. This hypotheses means that motor programs controlling muscular forces might be such as to produce a nearly constant stiffness condition because of the physical benefits of achieving movements which are both optimally smooth and energy efficient, as speculated by Nelson [1]. Actually, the comparison between the optimum movements with respect to various physical performance constraints, such as energy and rate of change of acceleration (jerk), shows the remarkable similarity of movements predicted by the linear-spring invariant model and by performance with minimum-energy-cost constraint.

2.1. LPC analysis

The LPC analysis filter consists of two parts, the pre-emphasis and the filter to be identified. The role of the pre-emphasis, which is the first-order, is to flatten spectral tilt, in order to correctly identify the poles at high frequencies. An information criterion indicates a second order filter cascaded with first-order filter as optimum for modeling movements of the tongue parameters. A standard LPC technique is used to identify the values of filter coefficients [6]. The corresponding impulse responses including the de-emphasis filter exhibit highly da-mped characteristics with an effective duration of 140 to 200ms [2]. Because the command is represented by the train of pulses, these impulse responses behave as an elementary gesture.

2.2. Biomechanical interpretation of the filters

The biomechanical interpretation leads us to consider the second order filter as neuro-muscular system (a force generation) and the pre-emphasis filter as a passive mechanical system. This result is supported by the natural frequency calculations. The pole frequency of the identified second order system is about $3Hz$ [2]. However, the natural frequency of the human tissues seems to be much higher than $3Hz$. For example, the oscillation frequency of the lips during bilabial stops release is approximately 33 Hz [3]. The mechanical resonance of cheek tissues, measured for tensed or relaxed condition, was found in the frequency range of 30 to $60Hz$ [4]. These values are much higher than the calculated natural frequencies of the second order filter. Because in frequencies below the natural frequency of the second order system the motion is determined by the stiffness and resistance, the identified mechanical system would behave as a viscoelastic system corresponding to the first-order filter. Our filter approch corresponds to a Hill type muscular contraction model [5].

2.3. Multi-pulse synthesis

Each filter is excited by a sequence of pulses representing articulatory commands. The position and the amplitude of the excitation pulses are determined from the measured movements using an MLPC method (Multi-pulse LPC) proposed by Atal and Remde [7]. The accuracy of the MLPC synthesized movements monotonously improves with an increase in the number of excitation pulses [2]. So, the question we raise is how many pulses are needed to synthesize the observed articulatory movements. In the articulatory domain, it is difficult to establish a criterion for this. We, therefore, resort to an acoustic criterion.

3. ACOUSTICAL EFFECTS OF MOVEMENTS

We attempt to determine which parameters have dominant influence on the phonetics features of vowels in the sentence. We calculated the first four formants frequencies along 10 sentences from the measured articulatory movements as references, $\mathcal{F}_{ref}$. These references formants patterns are then compared with those calculated from only one parameter synthesized with the M-LPC model, $\mathcal{F}_{syn,j}$, $j =$ **jw**, **tp**, **ts**, or **tt**. The values of the remaining six parameters are the measured ones. In these formant calculations, the measured lip movements are used.

The effect of each tongue parameter manifests when the number of pulses is

equal to 0. In this paper, we present only $\mathcal{F}_{syn,jw}$ and $\mathcal{F}_{syn,tp}$ with $m = 0$, (*see* figure 1). The greatest distance error between $\mathcal{F}_{ref}$ (solid line) and $\mathcal{F}_{syn,j}$ (dashed line) patterns occurs with **tp** synthesized movement. The back/front feature is highly damage. The **jw** effect is more important for the close vowels than the open vowels. Except when the open vowel is located at the end of the sentence. The two patterns $\mathcal{F}_{syn,ts}$ and $\mathcal{F}_{syn,tt}$ are almost identical with $\mathcal{F}_{ref}$. In spite of the important articulatory activity of these two parameters, their contribution to the F-pattern vowels is small. These observations imply that the tongue articulatory parameters aren't controlled in the same importance. In the next section, the value of the minimum pulses m_j^* is determined for each parameter j.

4. MINIMUM NUMBER OF PULSES

The error between the first four formants frequencies of $\mathcal{F}_{ref}$ and $\mathcal{F}_{syn,j}$, $F(n), n = 1, ..., 4$, in terms of energy is calculated by

$$\mathcal{E}_{i,m,j} = \sum_{k=1}^{N} \sum_{n=1}^{4} (F_{i,m,j}(n,k) - F_i(n,k))^2,$$

where i is the sentence number, j is the synthesized parameter which depends on the number of the pulse m, and k is the frame number. An error tolerance, $\mathcal{E}_{i,m_j^*,j}$, is estimated from the difference limen (DL) of formant frequencies [8]. The value of m_j^* represents the number of pulses involved to synthesized movement j with the minimum accuracy necessary in synthesizing the perceptually acceptable formants patterns. This number is always less or equal to the syllable number in the sentence. Figure 2, illustrates one example for the sentence "une réponse ambiguë". The determined number of pulses depends on the type of parameters; thus, $m_{jw}^* = 5$, $m_{tp}^* = 6$, $m_{ts}^* = 3$ and $m_{tt}^* = 2$. Note that the number of syllables in this sentence is equal to 6. The observed and synthesized movements show large differences, especially for **tt** and **ts**. $\mathcal{F}_{syn}$ calculated with a small number of pulses exhibits the error tolerance less than DL. So, a great accuracy in synthesized movements isn't necessary to produce formants frequencies of vowels with reasonable accuracy. Otherwise, error in the articulation can be very high when the parameter isn't directly involving in the production. For example **ts** for the vowels $[aN,]$, **tt** for $[aN, i, y]$. The number m_j^* seems to be in relation with the number of the phonetics features inherent to the parameter.

5. CONCLUSION

The results show that the number of pulses necessary for each movement to synthesized formants frequencies of vowels is always less or equal to the number of syllables in the sentence. This is explained by the fact that the effective duration of the filter's impulse response is comparable to that of the average syllable duration in the 10 sentences, $180ms$. Moreover, it suggests that articulatory controls involve a unit of production whose length is about the size of syllable. Having the method that allows us to describe the relatively complex articulatory movements in terms of a small number of pulses, the problem now is to find out something about the nature of spatio-temporal organization, specifically the inter-articulator coordination during speech.

This work is supported by european project *Esprit/BRA, n° 6975, SPEECH MAPS.*

6. REFERENCES

[1] W.L. Nelson, Physical principles for economies of skilled movements. *Biol. Cyber.*, 46 (1983), pp. 135-147.

[2] S. Bouabana and S. Maeda, Effets acoustiques de la modlisation articulatoire. *Springer-Verlag*, 2nd Edition, (1994).

[3] O. Fujimura, Bilabial stop and nasal consonants : a motion picture study and its acoustical implications. *Journal of Speech and Hearing Research*, 4(3) (1961), pp. 233-247.

[4] K. Ishizaka, J. C. French and J. L. Flanagan, Direct determination of vocal tract wall impedance. *JIEEE. Trans. on ASSP*, 30(11) (1983), pp.828-832.

[5] A. V. Hill, The heat of shortening and the dynamic constants of muscle. *Proc. Roy. Soc. Brit.*, 19(6) (1974), pp. 136-136.

[6] F. Makhoul, A new look at the statistical model identification. *IEEE Trans. on Automatic Control*, (1986) pp. 716-723.

[7] B.S. Atal and R. Remde, High-quality speech at low bit rates : Multi-pulse and stochastically excited linear predictive coders. *Proc. ICASSP*, (1986) pp. 614-617.

[8] J. Flanagan, Speech analysis synthesis and perception. *Springer-Verlag*, 2nd Edition, (1972).

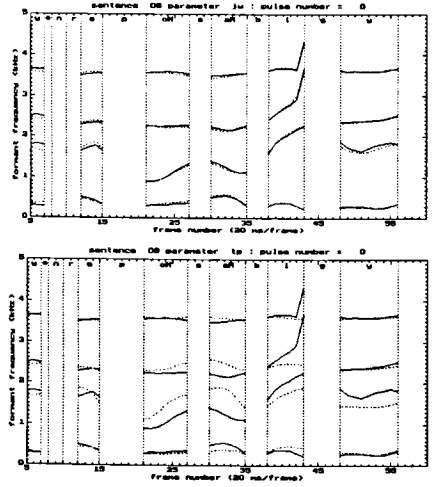


Figure 1: *Solid line: $\mathcal{F}_{ref}$. Dashed line: top, $\mathcal{F}_{syn,jw}$; bottom, $\mathcal{F}_{syn,tp}$, with $m = 0$.*

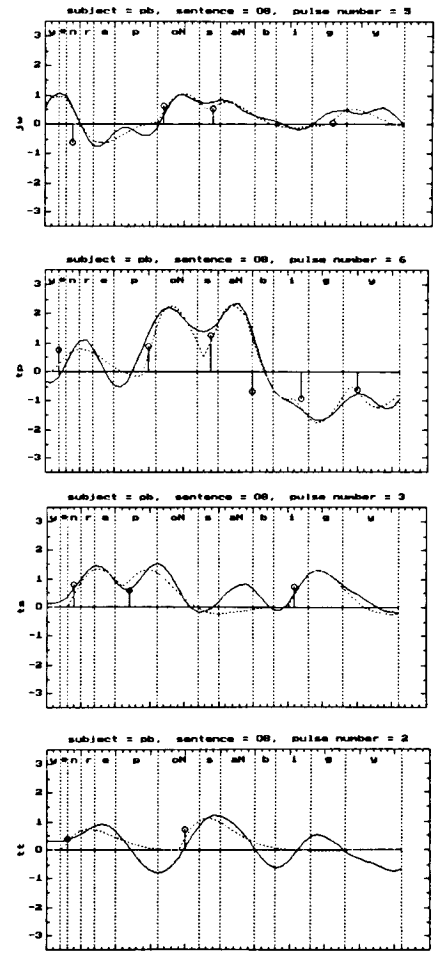


Figure 2: *observed (solid line) and synthesized (dashed line) movements with the minimum pulses for jw, tp, ts and tt parameters.*

Recovering place of articulation for occlusives in VCVs

Gérard BAILLY
Institut de la Communication Parlée U.R.A. - CNRS N° 368
I.N.P.G./E.N.S.E.R.G. - Université STENDHAL
46, Avenue Félix Viallet, 38031 GRENOBLE Cedex 1 France

ABSTRACT

We present a method for specifying sensory-motor templates for articulatory synthesis. Control of articulation plant is done by two modules: (a) a module which converts the phonetic string into a "sensory-motor" score, i.e. spatio-temporal templates of the desired properties of the sounds to be produced. (b) a trajectory formation module which computes a smooth articulatory trajectory satisfying all these requirements. We examine here the possibility of specifying simple VV and CV transitions from an acoustic score.

INTRODUCTION

We proposed in [1] a general framework for articulatory speech synthesis where gestures made audible by the extensive use of physical models are controlled by motor control principles. The advantages of such an approach are : (a) the control part of the synthesiser is only dedicated to high-level semiotic constraints on the gestures: physical modelling guaranties the production of ecological sounds since laws governing movements, aerodynamics and acoustics are part of the sound production device - the articulatory plant. (b) the distinction between proximal (actual control parameters of the plant) and distal space (heterogeneous space where tasks are defined) enables a flexible and optimal definition and control of speech tasks: if vocalic systems could be easily predicted in the acoustic space [9], geometric description of vocal tract targets in terms of constrictions seem well-suited for consonants. Nevertheless, models using tasks dynamics [10, 7] which unifies speech control in the geometric space, lack acoustic [12] and articulatory supervision [8] including the important role played by the jaw in language acquisition and control [5]. Complementarity of sensory-motor descriptions is thus essential to the definition of speech goals. The actual proximal trajectories must be then computed by a trajectory formation module [7] which find an optimal solution filling all constraints with minimal effort.

This article demonstrates that complementarity is effective in defining articulatory prototypes and that both geometric or acoustic definitions of CV syllables lead to stable articulations.

THE ARTICULATORY PLANT

We use an improved version of the

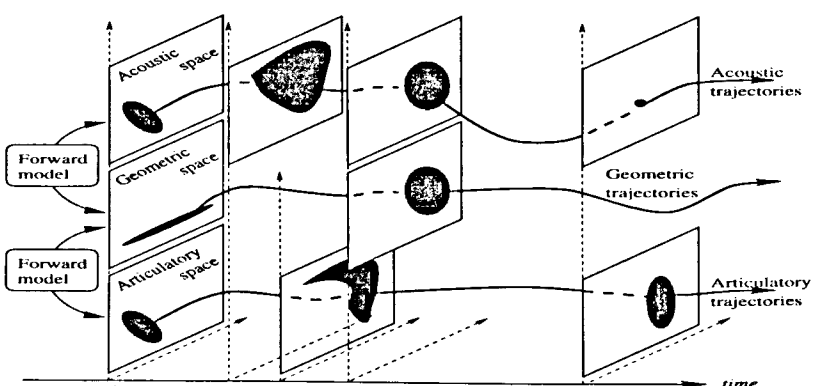


Figure 1: The sensory-motor score: all spaces may collaborate for defining templates. They are linked by Forward models capturing the different proximal-to-distal transforms.

Maeda's model [4] based on a re-analysis of the 519 tracings of mid-sagittal pictures used by the original statistical analysis (see Fig.2). A more detailed description of the model is given in [4]. The acoustic space (defined as babbling procedure around the neutral articulatory configuration in the limit of ±3 the standard deviation for each articulator) is presented in Fig.3. It illustrates how clearly the vocalic triangle is defined and evidences the overlap between F1-F2 and F1-F3 planes.

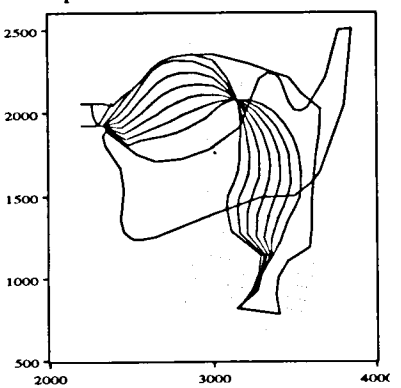


Fig.2. The articulatory plant.

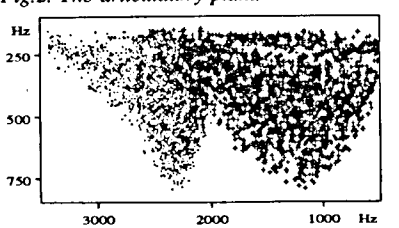


Fig.3. The maximal space

The different Forward models linking the articulatory space to the formant (computed using [2]) and the geometric spaces (command parameters of the Fant's model) are implemented as polynomial interpolators.

TRAJECTORY FORMATION

The module uses a constrained back-propagation of sensory-motor errors through the different Forward models augmented with a smoothness term minimising the articulatory jerk.

VOCALIC PROTOTYPES

Vocalic prototypes are computed by defining acoustic targets for the ten French oral vowels /a,ɛ,e,i,œ,ø,y,ɔ,o,u/. These targets are defined as dispersion ellipsis in the formant space. Starting from a neutral position (similarly to [9]), the gradient descent converges towards articulatory configurations with are given in Fig.4. Known articulatory hierarchy is clearly respected especially for the jaw: during the gradient descent, articulators move in synergy to fulfil the acoustic target. Although macro-sensitivities for the protrusion parameter around closed vowels /u/ and /y/ are almost null, this is not the case during the gradient descent: protrusion participate actively in the formation of the Helmholtz resonator neck.

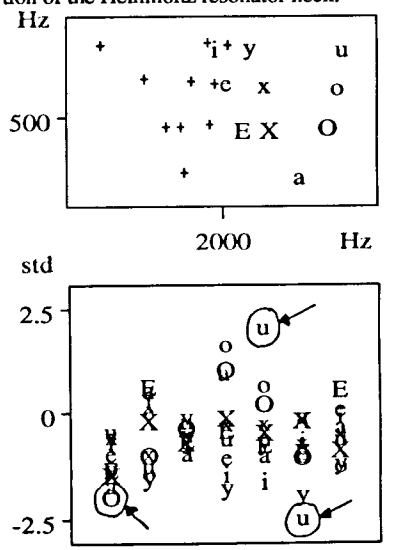


Fig.4. Top: acoustic targets. Bottom: superposition of articulatory targets. From left to right: jaw, lip height and protrusion, tongue body, dorsum and tip and larynx height. Please note /u/ and /y/ are the most closed and protruded, /u/ being realised with the higher dorsum and the lower tongue tip (bunched tongue) vowels whereas /a/ and /ɔ/ have the lowest jaw.

VOCALIC TRANSITIONS

What is the best control space for generating movements? We compared the acoustic results of the most simple strategy for generating inter-vocalic transitions: proximal movements are supposed to be zero-phased. Although this strategy is surely far too simple, large phasing relations are not expected to occur in these simple movements. Acoustic tran-

sitions between maximal vowels are presented in Fig.6. They can be compared to natural ones in Fig.5. Please not the lack of convergence of F2-F3 for the /i/-/a/ gesture which is however obtained by the Fant's model.

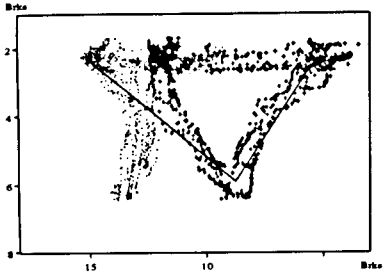


Fig.5. Natural VV transitions.

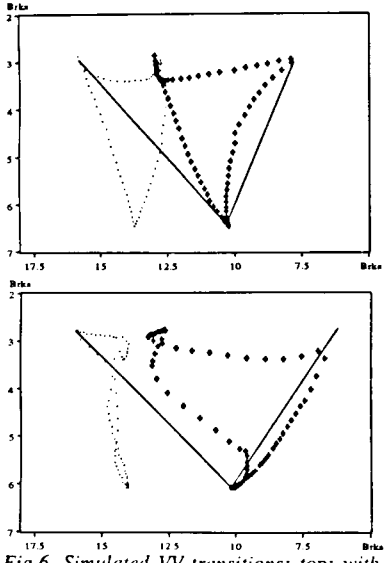


Fig.6. Simulated VV transitions: top: with prototypes proposed above; bottom: with prototypes proposed by Fant [6].

PROTOTYPES FOR STOPS

Articulatory prototypes for stops in a symmetric vocalic context were determined assuming that closure is obtained from the vocalic positions above via a simple efficient gradient descent towards geometric targets. These targets were defined as Al=0 & Ac>0 for /p/, Al>0 & Ac=0 for /t/ and /k/ with 0<Xc<1 for /t/ and 2<Xc<5 for /k/. These constraints were sometimes completed by some additional constraints avoiding for example double constrictions. The /ata/ gesture obtained by this procedure is given Fig.7. If bilabials are simply produce by synergetic actions on jaw and lip closure, dentals and palatals result in more complex gestures since both jaw and tongue body contribute at carrying the final tongue articulator, respectively. tongue tip and dorsum to the right place of articulation.

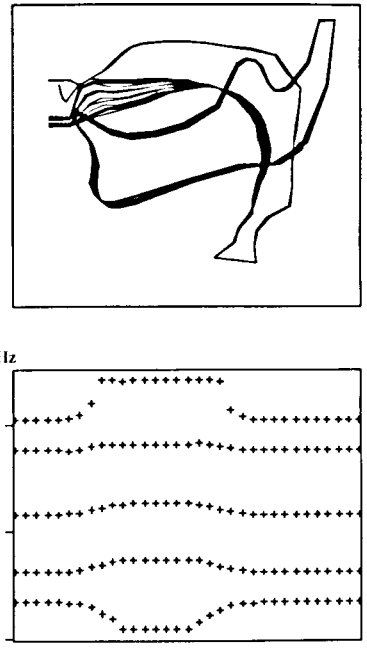


Fig.7. Articulatory prototype for /ata/.

We verified that regression lines linking the F2 at release and the F2 target of the vowel were in accordance with the so-called locus equations intensively measured in the literature [11]. The respective correlations with b-, d-, g-locus equations are .964, .884 and .891. When two loci are used for /g/ then the correlation is .923 and .79[1] for back and front articulations. The low correlation for /d/ is explained by the lack of sublingual volume predicted by our present plant. We hope to present new data at the conference using a new plant [3].

[1]This constant locus is due to affiliation changes and mainly depends on the length of the pharynx.

COHERENCE OF INVERSION

The acoustic VCV trajectories produced by these prototypes were used as acoustic templates in our trajectory formation module in order to show that place of articulation may be recovered from an acoustic specification using a reference model. Fig. 8 summarises the inversion results for /b/ and /g/. The results for /d/ are poor since no additional sublingual volume may explain low F2 values. /d/ is thus often confused with /b/.

CONCLUSION

I proposed here a heterogeneous control of speech articulation using sensory-motor templates. We will apply this scheme to articulatory synthesis using an articulatory model and intensive collection of acoustic, aerodynamic and articulatory data gathered on the same subject to enable a real assessment of inversion results. We will present more intensive simulations at the conference.

Acknowledgements

This theoretical proposal is part of a collective work which inspired the ARTIST project submitted to LTR call for proposal founded by the EEC.

REFERENCES

[1] Bailly, Laboissière & Schwartz (91) Formant trajectories as audible gestures: An alternative for speech synthesis, *J. of Phonetics*, 19(1), 9-23.

[2] Badin & Fant (84) Notes on vocal tract computations, *STL-QPSR*, 2/3, 53-108.

[3] Badin, Gabioud, Beautemps,, Lallouache, Bailly, Maeda, Zerling & Brock (95) Cineradiography of VCV sequences: articulatory-acoustic data for a speech production model, *ICA, (to appear)*.

[4] Beautemps & Gabioud (94) Adaptation d'un modèle articulatoire à un locuteur, dans le but de contraindre l'inversion articulatori-acoustique, *XXe JEP*, Lannion, 119-124.

[5] Davis & MacNeilage (94) Organization of babbling: a case study, *Language & Speech*, 37(4), 341-355.

[6] Fant (1992) Vocal tract area functions of swedish vowels and a new three-parameter model, *ICSLP*, 1, 807-810.

[7] Honda & Kaburagi (94) A dynamical articulatory model using potential task representation, *ICSLP*, 179-184.

[8] Lee & Beckman (94) Jaw targets for strident fricatives, *ICSLP*, 37-40.

[9] Lindblom, MacNeilage & Studdert-Kennedy (84) Self-organizing processes and the explanation of phonological universals.in *Explanation of Languages* Mouton, 181-203.

[10] Saltzman and Munhall (89) A dynamical approach to gestural patterning in speech production, *Ecological Psychology*, 1(4),1615-1623.

[11] Sussman, McCaffrey & Matthews (91) An investigation of locus equations as a source of relational invariance for stop place categorization. *JASA*, 90(3), 1309-1325.

[12] Tatham (95) The Supervision of Speech Production, in *Levels in Speech Communication*, Elsevier, 115-126.

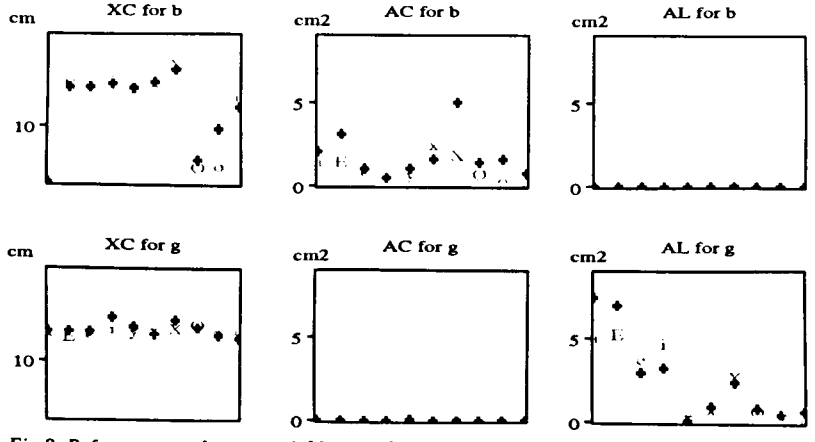


Fig.8. Reference vocal tract variables for /b/ and /g/ vs results from articulatory-to-acoustic inversion. Note that inverted constriction areas are smaller since vowels are also reduced.

FROM SENSITIVITY FUNCTIONS... ...TO MACRO-VARIATIONS

L.-J. Boë, P. Badin and P. Perrier

Institut de la Communication Parlée, Grenoble, France

ABSTRACT

This paper presents an approach to study and to predict the effects of articulatory gestures on formants for vowels. The classical sensitivity functions have been extended to macro-variations: large variations have been applied to the seven control parameters of an articulatory model for ten French vocalic prototypes. Non linearities have been analysed and the importance of the different articulators has been expressed in terms of audibility.

FROM SENSITIVITY FUNCTIONS...

In the tradition of dividing the vocal tract into a set of n-tubes of varying lengths and sections, sensitivity functions have been considered as a fairly efficient tools to study articulatory-to-acoustic relationships. In fact the expression *sensitivity function* was first coined in 1974 by Fant & Pauli [1], who then developed this notion in the frame of speech production theory. Sensitivity functions enable an evaluation of the consequences of small variations in area function ($\Delta A_i / A_i$ or $\Delta l_i / l_i$) of an undamped vocal tract on the corresponding resonance frequencies: transversal and longitudinal sensitivities.

Recently a variational formulation of the resonance modes in the vocal tract has been presented by Jospa [2]. This formulation provides and explicits (non linear) link between the formant frequencies and the area function: analytical expressions for the sensitivity functions that take into account wall admittance and lip radiation effects.

If $[l_A]$ and $[l_x]$ are units of area and length, the sensitivity function defined by Jospa, for a resonance mode n is:

$$S_n(x) = [l_A]^{-1} \delta f_n / f_n$$

It expresses the ratio of variation due to a small variation of the area function at the position x and corresponding to a Dirac distribution:

$$\delta A(x) = [l_A] \delta (x' - x)$$

Sensitivity functions can be useful to study the effects of small changes in area function, particularly at the constriction. However, it is not possible to extrapolate sensitivity functions to large variations. This is due to the fact that the relationship between area function and the corresponding formants can be highly non-linear (sometimes non-monotonous). This important property is at the origin of Stevens' Quantal Theory [3].

...TO MACRO-VARIATIONS

In fact, the approach which considers the vocal tract as a set of n-tubes whose dimensions (area and length) can be manipulated independently, without referring to an underlying articulatory model, even with few articulatory constraints, seems intrinsically very limited. The area function may not be the right control parameter for the vocal tract. This explains why the acoustic properties attributed to a set of n-tubes are of little use for the study of speech production if they do not correspond to realistic operations at the articulatory level. We thus propose to deal with the relations between vocal tract and acoustic by means of an articulatory model. Moreover, we have introduced the notion of *macro-variation* [4] corresponding to large variations (and not differential) of an *articulatory command* for a given articulation.

Non linearities

Using Maeda's model [5] implemented in an adapted environment [6], we have generated macro-variations for the prototypes of French vowels [i e ɛ a y ø œ u o ɔ] elaborated by Boë et al. [7, 8] for variations of the seven command parameters (Lip Height and Protrusion; Jaw Height; Tongue Dorsum, Body and Apex; Larynx Height), within a range of $\pm 1\sigma$, and a step of 0.25σ (except for the larynx: range of $\pm 3\sigma$, and step of 0.1σ). Configurations with constrictions under a threshold of 0.3 cm^2 were not considered as vowels and thus discarded. A dictionary of 1,421 configurations has thus been generated.

To evaluate the non linearity of the articulatory-acoustic relationship, the first four formants have been expressed as first degree polynomial combinations of the seven articulatory parameters:

$$F(P) = w_o + \sum_{i=1}^{7} w_i P_i$$

where the w_i coefficients have been determined by optimisation [9]. The relative mean quadratic errors (root mean square error / standard-deviation) are respectively, 43%, 18%, 46% and 54% for the first four formants. For 40% of the items in the dictionary, the explained percentage of the variance is less than 90% with a linear model. Two forms of non linearity can be observed: saturation and non-monotony (Figure 1).

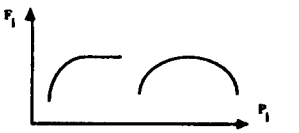


Figure 1: Forms of non linearity between articulatory parameters P_i and formants F_i: saturation and non-monotony.

Vowels presenting non linear relationships are listed in Table 1. The relation between LipH and F_1 for [i, y] is a typical example of the saturation mentioned by Stevens [3]. Dorsum presents non-monotonous relationships with F_1 and F_2.

	F_1	F_2	F_3	F_4
LipH	i y	i y	i y	i e y ɔ
Jaw		ɔ		
Body	y		u y ɔ	i e ɛ a ø œ u o
Dorsum	i y ø œ o	e ɛ a y œ ɔ		
Apex	œ		a	

Table 1: Vowels presenting non-linear relationships between aticulatory parameters and formants (explained variance < 50% with a linear relation).

From the point of view of the articulatory-acoustic relationship, independence can also be considered as a non linearity. This is the case of labial protrusion for all vowels and all formants. Independence and saturation result in *non-audible gestures*, i.e. gestures without acoustic effects, as defined by Abry and Schwartz [10].

Multinomial fits

Each formant frequency F_i in the dictionary was modelled by a separate multinomial function (degree up to 3) of the seven articulatory parameters [9].

$$F(P) = w_o + \sum_i w_i P_i + \sum_i \sum_{j \geq i} w_{ij} P_i P_j + \sum_i \sum_{j \geq i} \sum_{k \geq j} w_{ijk} P_i P_j P_k$$

Table 2 shows the RMSE (Root Mean Square Error / Standard Deviation) as a function of the polynom degree. A fit of degree 3 leads, for the main cardinal vowels [i a u], to relative errors less than 6% for the first four formants, except for a 10% error on F_1 of [u].

	deg. 1	deg. 2	deg. 3
F_1	43.1	13.8	4.5
F_2	18.1	8.8	3.8
F_3	45.7	19.1	8.4
F_4	53.8	26.6	12.6

Table 2: RMSE in % and polynom degree

Influence of gestures on formants

The influence of an articulatory parameter on a formant can be expressed by the relative variation of the formant for a given variation of the parameter. Thus, the resulting variations of the formants for variations of $\pm 1\sigma$ of each of the seven parameters, and for each vowel prototypes, have been compared to perceptual thresholds, in order to determine their audibility. The thresholds were 5% for F_1 and 10% for F_2 F_3, approximately following Flanagan [11].

Non audible gestures

Tables 3 and 4 indicate the cases where the variation of a given parameter (within the range of $\pm 1\sigma$, whenever

possible, but avoiding constrictions under 0.3 cm^2) is not audible. Note that the larynx has no effect at all. As well, protrusion is an articulatory parameter with low acoustic influence.

	LipH	LipP	Jaw	Body	Drsm	Apex
i	•	•	•		•	•
e	•	•				
ɛ		•				•
a		•	•		•	•
y		•				
ø		•				•
œ		•				•
u		•	•			
o				•	•	
ɔ		•	•			

Table 3: Non audible gestures for an increase of lip height and protrusion; a raising of jaw, dorsum and apex; a backward displacement of tongue.

	LipH	LipP	Jaw	Body	Drsm	Apex
i	•	•		•	•	
e		•			•	•
ɛ		•				•
a		•	•		•	•
y	•	•		•		•
ø		•			•	•
œ		•			•	•
u	•					
o	•		•		•	
ɔ		•	•			•

Table 4: Non audible gestures for a decrease of lip height and protrusion; a lowering of jaw, dorsum and apex; a forward displacement of tongue.

Audible gestures

Table 5 presents the relative variations of F_1 and F_2 that are larger than 10% for F_1, and 20% for F_2. Note that variations for F_3 did never reach 20%. The body is the most influencing for all vowels, mainly for F_2 [a ø œ u o ɔ], but also for F_1 [i e ɛ y ø œ]. LipH is mainly influent on F_1 of all vowels except [i]; Jaw has influence on F_1 only, for [i e ɛ y ø]. At last, dorsum influences only F_1 of [ɔ].

An example of combination of audible and non-audible gestures

Taking into account the sensivity functions of [i], it is impossible to infer the gesture of the [iy] transition. At the acoustic level: stability of F_1, displacement of the focalisation F_3-F_4 towards a focalisation F_2-F_3 with lowering of F_4, F_3 and to a lesser extent of F_2 [12]. At the articulatory level, this is achieved by an important decrease of lip height, an important increase of protrusion [13], a slight backward displacement of tongue body [14] and a slight larynx lowering [15]. The observation of macro-variations corresponding to these gestures allow to predict these acoustic effects. Protrusion, which has no acoustic effect, could be interpreted as a gesture facilitating the accurate realisation of a small area at the lips.

	F_1	F_2
i	+Body +41 -Jaw +24 -Apex +15	
e	-Body -25 -LipH -24 +Body +19 +Jaw -15 -Jaw +14	
ɛ	-LipH -22 -Body -20 +Body +13 +Jaw -10	
a	-LipH -24	-Body +22 +Body -21
y	+LipH +17 -Jaw +15 +Jaw -13 +Body +13	
ø	-LipH -32 +LipH +19 -Body -17 +Jaw -11	-Body +21 +Body -20
œ	-LipH -29 -Body -17	-Body +24 +Body -21
u	+LipH +73	-Body +26
o	+LipH +40	-Body +37 +LipH +26
ɔ	-LipH -27 +Drsm -12	-Body +30 +LipH +29

Table 5: Percentages of variation of the audible gestures corresponding to ± 1 σ variations of articulatory parameters.

CONCLUSION

With an articulatory model which integrates morphological constraints, it becomes possible to relate geometric variations and associated formant changes to real behaviours in speech production systems.

We have presented an approach that exploits the notion of macro-variations, instead of using sensitivity functions. Thus, the macro-sensitivity functions can be interpreted in terms of non-audible and audible gestures of speech production and motor control commands.

Non-linearities have been quantified by means of a polynomial fit, and it has been show that a multinomial fit of at least the third degree is needed.

The influence of the different articulators on formants for the French vowels has been analysed. Particularly, non audible gestures have been enlighted: larynx height and lip protrusion variations, observed during speech production, seem to have small effects on formants. Protrusion may be interpreted as a gesture facilitating the control of a parameter such as lip area. Oppositely, it has been verified and quantified that the forward/backward displacement of tongue body is one of the most imortant control parameters for F_1 and F_2, with lip height and jaw opening for F_1.

ACKNOWLEDGEMENT

This work has been partially funded by the EC ESPRIT/BR project *Speech Maps*. We are indebted to Andrew Morris of the ICP who has developped and processed the fitting of the articulatori-acoustic relationship.

REFERENCES

[1] Fant G.& Pauli S. (1974) Spatial characteristics of vocal tract resonance modes. *Speech Communication Seminar* 2, 121-126.

[2] Jospa P. (1994) Formulation variationnelle du lien acoustico-articulatoire. *20e Journées d'Étude sur la Parole, GFCP SFA*, 113-118.

[3] Stevens K.N. (1972) *The Quantal nature of speech: Evidence from the articulatory-acoustic data.* In *Human Communication: A Unified View*, 51-66, E.E. David Jr. & P.B. Denes (eds.), Mac-Graw Hill, New-York.

[4] Majid R., Abry C., Boë L.-J. & Perrier P. (1987) Contribution à la classification articulatori-acoustique des voyelles. Étude des macrosensibilités à l'aide d'un modèle articulatoire. *XIth Int. Congr. of Phonetic Sciences*, 2, 348-351.

[5] Maeda S. (1989) *Compensatory articulation during speech: evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model.* In *Speech production and modelling*, 131-149. W.J. Hardcastle & A. Marchal (Eds.), Academic Pub., Kluwer.

[6] Boë L.-J. (1993) *Speech Maps Interactive Plant "SMIP"*.WP2 From speech signal to vocal tract geometry. Sound-to-Gesture Inversion in Speech. ESPRIT/BR n° 6975.

[7] Boë L.-J., Perrier P. & Morris A. (1992) Une prédiction de l'"audibilité" des gestes de la parole à partir d'une modélisation articulatoire. *19e Journées d'Étude sur la Parole, GFCP SFA*, 151-157.

[8] Boë L.-J., Vallée N., Perrier P., Payan Y. & Savariaux C. (1994) *Codebook and sound prototypes with the final state plant.* WP2.2, RP2, Deliverable 17, Sound-to-Gesture Inversion in Speech, SPEECH MAPS (ESPRIT/BR n° 6975).

[9] Morris A. (1992) Least-Squares Fit to Maeda model dictionary. Summary of procedures used and results to date. *Technical Report, Institut de la Communication Parlée, Grenoble.* 8p.

[10] Abry C. & Schwartz J.L. (1988/89) in "La perception visuelle de la parole" Cathiard M.A., *Bull. Institut de Phonétique de Grenoble* 17-18, 110-114.

[11] Flanagan J.L. (1955) A difference limen for vowel frequency. *J. Acoust. Soc. Am.* 27, 613-617.

[12] Mantakas M. (1989) *Application du second formant effectif F'2 à l'étude de l'opposition d'arrondissement des voyelles antérieures du français.* Doctorat INP, Grenoble.

[13] Abry C. & Boë L.-J. (1986) "Laws" for lips. *Speech Communication* 5, 97-104.

[14] Wood S. (1989) Vowel gestures and spectra: from raw data to simulation and application. *12th Int. Congr. of Phonetic Sci.*, 1, 215-219.

[15] Riordan A. (1977) Control of vocal tract length in speech. *J. Acoust. Soc. Am.* 62, 998-1002.

A NOVEL SELF-ORGANISING SPEECH PRODUCTION SYSTEM USING PSEUDO-ARTICULATORS

C. S. Blackburn and S. J. Young
Cambridge University Engineering Department (CUED), England
email: csb@eng.cam.ac.uk

ABSTRACT

A novel articulatory speech production system which is stochastically trained from a pre-specified initialisation state is presented. The target positions for a set of pseudo-articulators and the mapping from these to output speech spectral vectors are jointly optimised using linearised Kalman filtering and an assembly of neural networks. The techniques used to initialise and train the system are described, and preliminary results when synthesising speech are demonstrated.

INTRODUCTION

Articulatory speech synthesisers model human speech dynamics and hence theoretically can produce very high quality speech waveforms with explicit time-domain modelling of co-articulation [8, 12, 15]. Two major problems confronting such systems are:

- Specification of the sequence of articulator positions or vocal tract area functions corresponding to a given text.
- Provision of an accurate model of the human vocal tract.

The former is frequently achieved using an "inverse" model to map parametrised speech, usually in the form of spectral vectors, into articulator positions or vocal tract areas and hence determine target positions for the phonemes to be synthesised. We use a Kelly-Lochbaum synthesiser [6, 12] to generate a codebook of (articulator vector, spectral vector) pairs [13] which is inverted using dynamic programming (DP) incorporating geometrical constraints on the articulator trajectories, as shown in figure 1.

The inverse mapping is non-unique, so dissimilar articulator positions may result in similar acoustic outputs [2, 7], hence attempts to model the inverse transformation using acoustic error alone [1, 10] are likely to produce discontinuous articulatory output. A continuity constraint should therefore be applied to such trajectories, which may be implicit as in inverse filtering techniques [16], or explicitly imposed via a restriction to critically damped second order transitions [14] or the minimisation of geometrical distances [13, 17].

In addition, the non-linearity of the inverse mapping combined with its non-uniqueness can result in non-convex target regions in articulator space [4], so gradient-based algorithms which average over a number of training vectors, whether a single neural network [1, 10, 17], Jacobian computation [5] or unconstrained optimisation [7], may converge to an average which does not lie within the target class, resulting in an incorrect inverse model. This problem can be avoided either by subdividing the input space into regions in which the non-linear mapping is unique [11], or by jointly optimising an (inverse, forward) model pair to restrict the inverse model to a particular solution [3].

In our system the use of codebook look-up guarantees that a particular inverse solution is chosen at each point in time, and the DP search incorporates both acoustic and geometric constraints to ensure continuity.

The second problem, that of determining an accurate vocal tract model, is approached in our system by relaxing the constraint that the system exactly mimic human physiology. Instead, we use "pseudo-articulators" which fulfil roles similar to those of human articulators but whose positions are stochastically estimated from the training data. The initial articulator trajectory estimates obtained from the DP algorithm are iteratively re-estimated using linearised Kalman filtering and an assembly of neural networks which map from articulator positions to output speech.

SYSTEM INITIALISATION

System initialisation is shown in figure 1. Vocal tract area functions are determined from a set of five pseudo-articulators as in [9]. Four of these, roughly specifying tongue position, are sampled at regular intervals to give 6321 basic vocal tract shapes. A logarithmic quantisation is then applied to eliminate very similar shapes; since our aim in initialisation is to determine a set of articulator trajectories, time domain quantisation

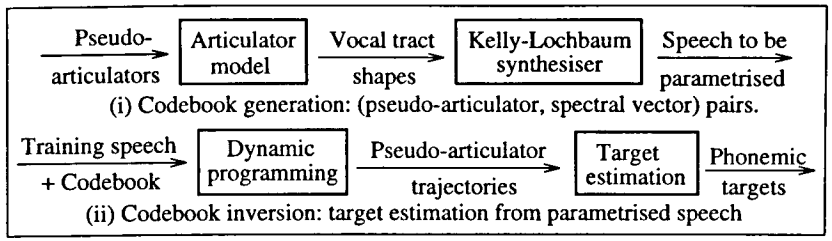


Figure 1: System initialisation.

is preferable to that in the frequency domain as used elsewhere [13].

Quantised lip opening is then added as a fifth parameter giving 27651 pseudo-articulator vectors which are used to generate a corresponding set of 10-section vocal tract area functions. These are interpolated in the logarithmic domain and re-sampled to yield an appropriate number of area sections for use in the Kelly-Lochbaum synthesiser, which treats the vocal tract as a variable number of fixed cross-sectional area tubes and incorporates separate oral and nasal tracts, as well as modelling transmission loss. A sampling frequency of 16kHz corresponding to area sections of length ≈1.1cm was chosen, and both 15 and 16-section resampled area functions were used, giving a total of 55302 basic shapes.

Fricative waveforms are created from shapes with a constriction of less than 0.3cm^2 using a random noise source at the constricted point which is correlated with the voiced excitation, if any. Nasals are generated from the parallel combination of a variable oral tract and a fixed nasal tract, for three values of velum opening. In all, 31848 voiced and unvoiced fricatives and 15126 nasals were included, in addition to 55302 purely voiced waveforms. In each case the speech waveforms were parametrised by the CUED HTK recogniser to give one 12-dimensional liftered cepstral vector per 10msec of speech. Finally, 212 cepstral vectors representing "silence" or background noise in the training speech were added to give a total codebook size of 102488 vectors.

Codebook inversion

The training speech comprised 600 sentences of one adult male from the speaker-dependent training portion of the Defence Advanced Research Projects Agency (DARPA) Resource Management corpus. This speech was also coded into 12-dimensional cepstral vectors, and dynamic programming was used to find a path through the lattice of possible articulator trajectories.

At each step of the DP algorithm, both the acoustical mismatch between the parametrised training speech vector and the codebook acoustic vectors *and* the geometrical mismatch between successive articulatory vectors are combined into a weighted score when evaluating paths. To reduce the computational load, a sub-optimal search was used in which only the 500 codebook vectors with the best acoustic match were considered at each step.

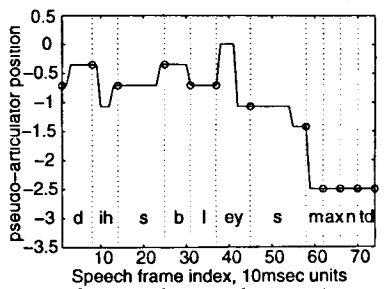


Figure 2: Pseudo-articulator trajectory for "displacement".

Pseudo-articulator trajectories such as that in figure 2 were generated in this way for all 600 sentences. This figure shows the trajectory of one pseudo-articulator during the word "displacement", where phoneme boundaries taken from the HTK-produced label file are marked as vertical lines. The phone labels are taken from the DARPA transcription of the speech, and in most cases the pseudo-articulator is steady between boundaries.

Target statistics are thus determined from the values of the articulators at the midpoint of each occurrence of each phoneme to give initial target means and covariance matrices for each of the five basic articulators for the 47 phonemes in the data set.

TRAINING

A separate neural network is used to learn the mapping from the pseudo-articulator trajectories of each phoneme to output speech. The trajectories are piecewise linear interpolations of the phoneme target means, constrained to pass through the average of two adjacent target means at the phonemic boundary. The training set output vectors were 24-dimensional mel-scaled log spectral coefficients; while this is a less efficient representation than the cepstral coefficients used previously, their use results in a more easily learned non-linear function.

The purpose of the neural networks is to approximate this mapping from articulatory to acoustic space, so that the linearised Jacobian matrix can be used to re-estimate the phonemic targets; hence their performance and architecture are not crucial to the training process. We trained feed-forward multi-layer perceptrons with 12 inputs, 30 hidden units, 24 outputs and sigmoid non-linearities at the hidden units using resilient back-propagation (rprop) for 1000 batch update epochs, giving mean errors in estimated spectral coefficients of around 10%.

The global error covariance matrix for each network mapping is estimated from its performance on an unseen test set, and the Jacobian matrix is found by extending the usual error back-propagation formulae to evaluate the derivative of each output with respect to each input:

$$\frac{\partial y_k}{\partial y_i} = \sum_j (w_{ij} w_{jk} y_j (1 - y_j))$$

where y_i, y_j, y_k are the outputs of nodes in the input, hidden and output layers respectively and w_{ij}, w_{jk} are the input-hidden and hidden-output weights respectively.

If the initial estimate of a phoneme's articulatory target mean vector is denoted $\hat{\mathbf{x}}$, with associated initial covariance matrix $\hat{P}$, and if the neural mapping is denoted $h()$ with Jacobian matrix H at the target estimate, output $\mathbf{z}$ and output error covariance matrix R, the target vector can be re-estimated using linearised Kalman filtering as:

$$\mathbf{x} = \hat{\mathbf{x}} + \hat{P} H^T (H \hat{P} H^T + R)^{-1} (\mathbf{z} - h(\hat{\mathbf{x}}))$$

This gives a re-estimated target vector for each occurrence of each phoneme, from which new target mean and covariance statistics are computed. Updated pseudo-articulator trajectories are then derived and the networks re-trained. This process is iterated until the optimum set of phoneme targets is obtained, from which speech is synthesised.

RESULTS

Figure 3 shows original and synthetic smoothed 24-dimensional mel-scaled filter bank vectors for the phrase "clear windows". The phoneme alignment produced by HTK has resulted in small timing errors at phoneme boundary positions, however the gross spectral characteristics of the two plots correlate well.

Formant transitions are generally well defined, although the co-articulation from the stop /d/ to the following vowel /ow/ in "windows" has been missed by the synthesiser. The use of a separate neural network for each phoneme results in some discontinuities at phoneme boundaries, for example immediately preceding the final fricated /z/ in "windows", however the formants themselves are well-defined across boundaries, and high-frequency frication is successfully modelled.

Future work

The system is still under development, and many features have yet to be implemented. In particular, improved co-articulation modelling could be provided via the explicit modification of the target means according to their context. Since we have statistics for target means *and* variances for each phoneme, this should permit statistically-based co-articulation effects to be modelled.

In addition, the use of pseudo-articulators which are not constrained to human physiology provides the possibility of adding additional articulators during the training phase, thus potentially increasing the amount of information available to the neural mappings.

Finally, a method for smoothly combining the outputs of the neural networks across phoneme boundaries should reduce errors due to discontinuities.

CONCLUSIONS

This paper has presented a novel pseudo-articulatory speech production model, which is initialised by generating a codebook of (acoustic vector, spectral vector) pairs using a conventional Kelly-Lochbaum articulatory synthesiser which is inverted using sub-optimal dynamic programming search combining acoustic and geometric cost functions. The means and covariance matrices of the articulator targets for each phoneme are then estimated over 600 sentences of one speaker, and articulator trajectories corresponding to the training speech are constructed using constrained piecewise linear interpolation between the means.

An individual neural network is then trained to learn the mapping from articulators to parametrised speech vectors for each of 47 phonemes, and the target means are re-estimated using these mappings and a linearised Kalman filter. This process is iterated to find the optimum set of target means from which output speech is synthesised.

While articulatory synthesisers still do not produce speech comparable to that of the best rule-based synthesisers, we have attempted to show that the inability to exactly model the human speech production mechanism need not limit their viability, and have demonstrated a preliminary stochastically-trained system which yields promising results.

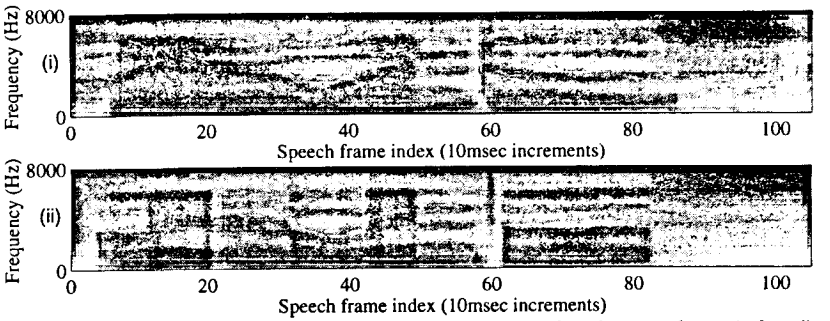


Figure 3: (i) original and (ii) synthesised filter bank output for phrase "clear windows".

REFERENCES

[1] B. S. Atal. "Neural networks for estimating articulatory positions from speech". *J. Acoust. Soc. Am.*, 86:S67, 1989.

[2] B. S. Atal et al. "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique". *J. Acoust. Soc. Am.*, 63(5):1535–1555, May 1978.

[3] M. George, P. Jospa, and A. Soquet. "Articulatory trajectories generated by the control of the vocal tract by a neural network". In *Proc. Int. Conf. Sp. Lang. Proc.*, volume 2, pages 583–586, 1994.

[4] M. I. Jordan and D. E. Rumelhart. "Forward models: Supervised learning with a distal teacher". *Cog. Sc.*, 16:307–354, 1992.

[5] P. Jospa and A. Soquet. "The acoustic-articulatory mapping and the variational method". In *Proc. Int. Conf. Sp. Lang. Proc.*, volume 2, pages 595–598, 1994.

[6] J. L. Kelly Jr. and C. Lochbaum. "Speech synthesis". In *Sp. Comm. Sem.*, Stockholm, 1962.

[7] S. E. Levinson and C. E. Schmidt. "Adaptive computation of articulatory parameters from the speech signal". *J. Acoust. Soc. Am.*, 74(4):1145–1154, 1983.

[8] P. Mermelstein. "Articulatory model for the study of speech production". *J. Acoust. Soc. Am.*, 53(4):1070–1082, 1973.

[9] P. Meyer, R. Wilhelms, and H. W. Strube. "A quasiarticulatory speech synthesizer for German language running in real time". *J. Acoust. Soc. Am.*, 86(2):523–539, 1989.

[10] G. Papcun et al. "Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data". *J. Acoust. Soc. Am.*, 92(2 Part 1):688–700, Aug. 1992.

[11] M. G. Rahim et al. "On the use of neural networks in articulatory speech synthesis". *J. Acoust. Soc. Am.*, 93(2):1109–1121, Feb. 1993.

[12] P. Rubin, T. Baer, and P. Mermelstein. "An articulatory synthesizer for perceptual research". *J. Acoust. Soc. Am.*, 70(2):321–328, Aug. 1981.

[13] J. Schroeter and M. M. Sondhi. "Techniques for Estimating Vocal-Tract Shapes from the Speech Signal". *IEEE Trans. Sp. Aud. Proc.*, 2(1):133–150, Jan. 1994.

[14] K. Shirai and T. Kobayashi. "Estimating articulatory motion from speech wave". *Sp. Comm.*, 5(2):159–170, June 1986.

[15] M. M. Sondhi and J. Schroeter. "A hybrid time-frequency domain articulatory speech synthesizer". *IEEE Trans. Acoust. Sp. Sig. Proc.*, ASSP-35(7):955–967, July 1987.

[16] H. Wakita. "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms". *IEEE Trans. Aud. Electroacoust.*, AU-21(5):417–427, Oct. 1973.

[17] J. Zacks and T. R. Thomas. "A new neural network for articulatory speech recognition and its application to vowel identification". *Comp. Sp. Lang.*, 8:189–209, 1994.

IRREGULARITIES IN THE VOICE: SOME PERCEPTUAL EXPERIMENTS USING SYNTHETIC VOICES.

Jan Gauffin and Svante Granqvist, Dept. of Speech Communication and Music Acoustics, Royal Institute of Technology (KTH), Stockholm, Sweden
Britta Hammarberg, Stellan Hertegård and Alf Håkansson, Dept. of Logopedics and Phoniatrics, Karolinska Institute, Huddinge University Hospital, Huddinge, Sweden

ABSTRACT

Sustained vowels with different kinds and magnitudes of jitter and shimmer sequences were synthesised and evaluated in a listening test. Though having the same jitter and shimmer magnitude, the different sequences were rated differently.

INTRODUCTION

For normal voices the acoustic effect of irregularities in the voice source is, in most cases, subtle, but pathological changes in the larynx may cause it to be a prominent feature. As we can hear rather small perturbations in the voice, we should also be able to measure and make objective classification of such perceptual voice qualities. Unfortunately, most attempts to do so have been rather disappointing.

Methods of chaos physics can be used to explain the mechanism behind observed irregularities in the voice source such as period doubling and chaotic vibrations of the vocal folds [1]. Simple simulation models of the vocal folds, like the two-mass model, can also be used to illustrate these mechanisms. However, understanding the mechanisms behind the irregularities in the voice source does not automatically give us the methods to analyse the corresponding acoustic features.

One kind of irregularity in the voice is the occurrence of more or less regular patterns of period-to-period variability. Such voices are often referred to as rough or creaky [2]. In the frequency domain, this corresponds to subharmonics in the spectrum. In the present investigation we have been studying the perception of synthetic voices with subharmonics by using synthetic vowels with repetitive patterns of perturbed periods.

METHOD

The synthesis was produced using the LF-model by Fant, Liljencrants & Lin [3], followed by a vocal tract filter tuned for the vowel /a/ [4]. Average fundamental frequency was 100 Hz and sampling frequency 16000 Hz. Different sequences of fundamental periods were generated, following the sequences AB, AOBO and AABB (see Tables 1 and 2). In the jitter case, A stands for a prolonged period and B for a shortened one, whereas O stands for an unmodified one. In the shimmer case, A stands for a period with higher amplitude and B stands for a period with lower amplitude.

This results in six different types of stimuli. Each of these was generated with 10 different magnitudes giving a total of 60 stimuli.

The sequences AOBO and AABB should give the same average jitter or shimmer measure since the AOBO sequence varies 1 unit between every period and the AABB sequence varies 2 units every other period. The AB sequence varies 2 units between every period. Therefore the magnitude of the peak period-to-period variability was divided by 2 in the AB sequence. The different sequences are illustrated in Figure 1 and the spectra of the six types of stimuli in Figure 2. All jitter and shimmer values refer to the voice source.

Table 1. The jitter sequences, period time deviations, n=0,1,..,9

Per. #	AB	AOBO	AABB
1	+n·0.6%	+n·1.2%	+n·1.2%
2	-n·0.6%	0%	+n·1.2%
3	+n·0.6%	-n·1.2%	-n·1.2%
4	-n·0.6%	0%	-n·1.2%

Table 2. The shimmer sequences, period amplitude deviations, n=0,1,..,9

Per. #	AB	AOBO	AABB
1	+n·1%	+n·2%	+n·2%
2	-n·1%	0%	+n·2%
3	+n·1%	-n·2%	-n·2%
4	-n·1%	0%	-n·2%

The equal average period-to-period variability should make it possible to examine and compare the perceptual sensitivity between the different sequences. However, the jitter and shimmer sequences with the same n are not comparable, since the step magnitudes of 0.6% and 1% are chosen arbitrarily.

Figure 1. The three different shimmer sequences. n=9 (Table 2).

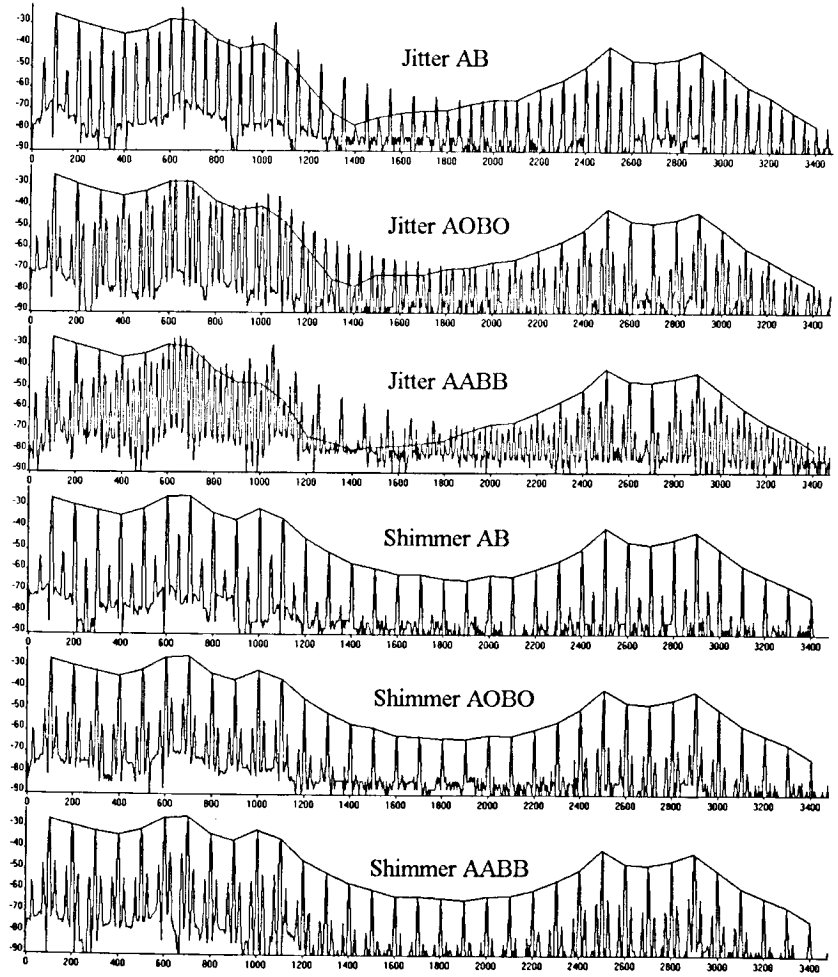


Figure 2. Narrow-band spectra of the six different types of stimuli and the envelope of harmonics. n=6 (Tables 1&2).

In Figure 2, one may note the difference between the jitter and shimmer stimuli by the envelope of the subharmonics. The jitter stimuli contain much weaker low-frequency subharmonics than higher-frequency subharmonics whereas the shimmer subharmonic envelope follows the envelope of the harmonics. It is also obvious that the stimuli with AABB or AOBO sequences have 25 Hz between spectral peaks, while the AB sequences have 50 Hz between the peaks. (Strictly speaking, the fundamental frequency has dropped from 100 Hz to 50 or 25 Hz, but if the subharmonics are weak enough the perceived pitch will still be 100 Hz). The spectra of AOBO and AABB stimuli appear similar but not identical.

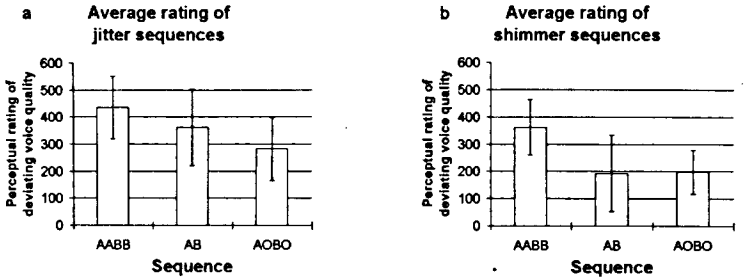


Figure 3a, b. Average and standard deviation of ratings of jitter (a) and shimmer (b) sequences. Each bar is an average of ratings of n=0,1,..,9 and all listeners.

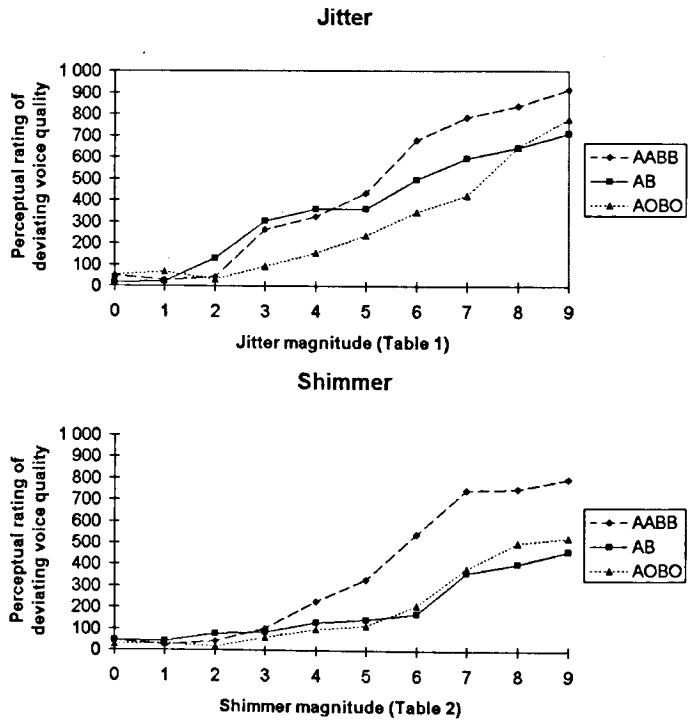


Figure 4. Average of all listeners' ratings of the different sequences.

The 60 stimuli were presented in random order to two experienced voice clinicians, two voice researchers and one musician. Each listener listened to the stimuli between one and three times, giving a total of 12 listening series. The listeners were asked to rate the degree of deviating voice quality on a visual analogue scale [5]. The position on the analogue scale was translated to a number between 0 (no deviation) and 1000 (maximum deviation).

RESULTS AND DISCUSSION

The results indicated that the sequences were rated quite differently regarding deviating voice quality. In the jitter case, the AABB sequence was rated 20% higher than the AB sequence and the AOBO was rated 22% lower than the AB sequence on average (see Figure 3a). In the shimmer case, the AABB sequence was rated 88% higher than the AB sequence and the AOBO was rated 3% higher than the AB sequence on average (see Figure 3b).

Figure 4 shows the perceptual ratings as a function of the magnitudes of jitter and shimmer for the different sequences. As expected, an increased degree of jitter or shimmer results in an increased perceptual rating of deviating voice quality.

A closer look at the spectra for the (Figure 2) might explain why the different types of stimuli were rated differently. Comparing the AABB and AOBO spectra, the latter have slightly weaker subharmonics and, furthermore, the jitter AOBO spectrum lacks the subharmonics 0.5·F0, 1.5·F0 etc.

This indicates that the levels of subharmonics could be more perceptually relevant than the jitter and shimmer measures.

All the listeners rated the AABB and AOBO stimuli as above. However they disagreed on how the AB sequence should be rated compared to AABB and AOBO sequences. This can be seen in the standard deviation bars in Figure 3. The AABB and AOBO sequences had a similar sound quality, while the AB stimuli differed from them. The different sound qualities can be explained by differences in the spectra. The AB sequence had 50 Hz between spectral peaks and the AABB and AOBO had 25 Hz. This might explain why the listeners disagreed on how to compare the AB stimulus with the AABB or AOBO stimuli.

CONCLUSIONS

We have shown that different kinds of jitter and shimmer sequences with the same average period-to-period variability are rated differently. This indicates that methods using period-to-period variability as a way to rate voice qualities might fail. We have also seen that there are differences in narrow-band spectra of the stimuli that might explain the differences in perceptual rating of the stimuli. For voices with repetitive patterns in period-to-period variability, this suggests that a method analysing spectral characteristics might yield better results than jitter or shimmer methods.

ACKNOWLEDGEMENT

This work was supported by research grants from the Bank of Sweden Tercentenary Foundation and the Karolinska institute.

REFERENCES

[1] Lauterborn, W. & Parlitz, U. (1988): Methods of chaos physics and their application to acoustics. *J. Acoust. Soc. Am.*, vol. 84, pp. 1975-1993.

[2] Imaizumi, S. (1986): Acoustic measures of roughness in pathological voice. *J. Phonetics*, vol. 14, pp. 457-462.

[3] Fant, G., Liljencrants, J. & Lin, Q. (1985): A four-parameter model of the glottal flow. *STL-QPSR*, vol. 4, pp. 1-13 (Speech Communication and Music Acoustics, Royal Institute of Technology, Stockholm).

[4] Håkansson A. (1995) *LF-Edit. Windowsprogram för LF-modellen.* Manual for custom-made program. Dept of Logopedics and Phoniatrics, Karolinska institute, Huddinge University Hospital

[5] Wewers, M.E., Lowe, N.K. (1990): A critical review of visual analogue scales in the measurement of clinical phenomena. *Research in Nursing & Health*, vol. 13, pp. 227-236.

VALIDATION OF TEMPORAL & SPECTRAL NOISE PARAMETERS USING (RE)SYNTHESIS.

Peter Pabon° & Guus de Krom
Research Institute for Language and Speech, University of Utrecht, the Netherlands.
°Institute for Sonology, Royal Conservatory, the Hague, the Netherlands.

ABSTRACT

Our main objective in this study was to develop a numerical algorithm or processing scheme that, starting from normal speech samples, produces a natural rough or breathy sound quality. Three methods are presented that illustrate the limited validity of spectral models for period-to-period variation and noise.

INTRODUCTION

Apart from being a playful effect processor, a "roughner" or a "breathinizer" has a serious use in the calibration of algorithms for voice-quality measurement. Although the auditory effects produced by such a processor may seem obvious or evident, the underlying models are often not. To produce a natural rough or breathy quality, the corresponding acoustic characteristics must also be modelled very accurately. Knowing the credentials of the acoustic effect means knowing how to develop the ultimate quality measurement device. Apart from being fun, the option of evaluating the quality by listening to the results makes the research less abstract and more effective. Often, experimenting with test signals quickly helps to define improvements of a model.

Usually, the only criterion for the validity of an acoustic voice-quality parameter is the comparison to perceptual or clinical ratings. The correlations that are found generally indicate only global relations. As a result, no clues, only educated guesses can be given how improvements should be made.

In this paper we try to take a by-way around this approach by using analysis/resynthesis based on the Overlap-Add (OLA) method [1]. With this method, much can be said on forehand about the quality of a model or spectral rating. Modern DSP technology allows a real-time implementation of the OLA method. OLA is a challenging research instrument which allows complex models that use elaborate processing schemes to be judged interactively.

The OLA method

The OLA method is an extension of the general method of spectrum analysis in which a time-varying spectral representation is derived by processing successive time frames. In the OLA method, the inverse process is also formalised. A frequency-to-time-domain transformation followed by an addition of subsequent frames restores the continuous time signal. OLA enables us to perform specific operations on the time-signal by manipulating the related spectral features.

A major limitation of the OLA method is that every spectral modification should represent a valid manipulation of the related time-domain signal. For instance, too rigorous cutting in the amplitude spectrum could disrupt the window-structure in the time domain and thereby produce discontinuities in the resynthesized output signal. Each spectral modification should prevent the complex liaison of amplitude and phase information from being disunited. To guarantee this, each spectral (re)organisation should always represent a realistic time-domain principle like filtering, correlation, shifting, integration, and so on. It is questionable if this restraint is a draw-back. Bizarre spectral or cepstral models could be used to rate voice-quality, but when their corresponding time-domain representation is intangible their relevance will probably be difficult to validate.

Our main approach in the simulation of a rough or breathy quality is a perturbation of the phase-relations of a periodic signal, while keeping the amplitude spectrum unchanged. Although the processing is done in the frequency domain, the main effect will be in the time domain.

Method I, Jittering

The first method uses OLA to randomly vary the linear phase component. The time domain result is a time shift of the entire frame. The frequency domain implementation allows for window shifts in fractions of the sample period. This variable window displacement produces a jittered version of the original (see Fig.1).

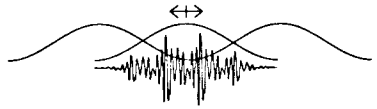


Figure 1, Jittering by random positioning of the frame.

The shape of the window/taper will prevent the emergence of hard discontinuities. The slight phase mismatch is smoothed by the fade-in-fade-out curve of the connected frames. This method is generally used in granular synthesis of single sound samples. If the window is periodically matched to the pitch period, this technique corresponds to a PSOLA method with a slightly perturbed periodicity.

The phase perturbation is proportional for all frequency components and is typically located around the frame boundaries. Seen over the frame, the amplitude spectrum is unaltered. Seen over an analysis period that includes more frames, the harmonic structure is slightly shattered, but the overall spectral envelope remains the same.

Method II, phase perturbation

In the first method, the entire frame was shifted as a block. In the second method, the frame positioning is randomised per frequency band. This effect is realised by convolving the signal with white noise. For band-limited random modulation of the phase, this is the best approach as it will guarantee continuity in both domains. Another way to look at this process, is to consider it as an all-pass filtering technique, where the phase delay of the all-pass filter is updated dynamically with each frame. Basically, the process is a cross-correlation or phase-vocoding technique [2] using white noise as a modulator (see Figure 2).

As can be seen in Figure 2, the amplitude spectrum is wobbly, but still remains maximally flat given the chosen phase curve. The overall spectral envelope is preserved and the speech output is still intelligible, although the quality is severely rough. The random time shift per frequency band breaks up all time-synchronisation that used to lead to sharply defined periodic excitation moments. If the low frequency part of the spectrum is processed in this way, our perception of a clear pitch is largely gone.

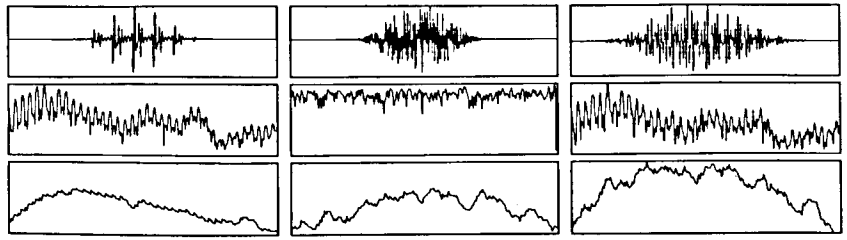


Figure 2, Perturbation by multiplying the signal spectrum with a white noise spectrum. (top) time signals, (middle) log spectra, range 80 dB, (bottom) phase spectra, range 16π, (left) speech signal, (middle) white noise and (right) noise modulated speech signal.

Phase versus amplitude information

Both methods were based on the principle that the periodicity information held by the phase spectrum was altered while the periodicity information in the amplitude spectrum was largely preserved. In Figure 2, the shape of the harmonics in the amplitude spectrum is deformed, but their average spacing remains the same. The auto-correlation-function, the cepstrum, even the complex cepstrum that also includes phase information, could still show a peak that indicates a clear fundamental periodicity, while the remains of this periodicity in the time-domain are completely lost by the randomisation of the phase information. Of course, the situation that the phase information is largely scrambled while the amplitude information is not, is not likely to appear to that extent in natural signals.

Still, the fact that large perturbations can occur without a corresponding large modification of the harmonic structure in the amplitude spectrum shows that every perturbation or noise model that is solely based on this amplitude representation is incomplete.

Additional information can of course be found in the regularity of the phase spectrum. However, it is questionable if the FFT phase spectrum is a good candidate for a separate model of (a)periodicity. Even if an elegant phase-unwrapping method is available, any correlational measure based on FFT phase information will suffer from the fact that a large part of the phase spectrum is non-deterministic. Overall, the curve may seem smooth, but, when inspected on an enlarged scale, the curve is jagged. For the FFT phase spectrum, a strange duality exist. If there are spectral components that have a clearly defined phase, they are more likely to be overshadowed by neighbouring spectral components that have not. The more concentrated the information due to periodicity, the less defined (and thus the more jumpy) the phase values of the other undefined components in between. A comparable principle is found with the cepstrum; the sharper the harmonics, the sharper and more jumpy the dips in the log-spectrum and the more noisy the base line of the cepstrum.

Phase/frequency stability is the base for the concentration of spectral energy in harmonics. The Fourier transform translates stability to a distinct spectral amplitude, thereby leaving only a superficial footprint on the mostly irregular terrain of the phase spectrum. On this terrain, the harmonics form stepping stones at which FFT phase information makes sense.

Apart from the models shown above, the amplitude spectrum can misrepresent what we consider noise or periodic on more occasions. For instance, the amplitude spectrum of a periodic noise burst can show a nice harmonic structure, while the noise within the bursts is not correlated, only the envelope (see Fig. 3).

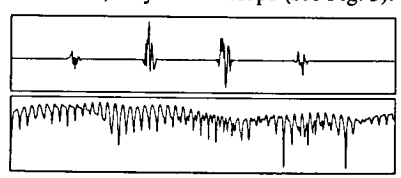


Figure 3, Windowed periodic bursts of white noise (top) and the corresponding logarithmic amplitude spectrum (bottom), vertical scale range 40 dB.

For a signal as in Figure 3, a repetition pitch can be perceived, and a harmonic series can be expected. However, it does illuminate an important question in the design of spectral models for voice (a)periodicity and noise: *In what way is the information on period-to-period correlation represented in the harmonic series?* This question is of relevance for the modelling of a breathy voice quality for which a period-linked noise burst seems an important attribute. It even

seems that the time synchronisation is not the only prerequisite, more complex correlations within the ensemble [3] are likely to play a role. A related idea is that any source (a)periodicity is linked to the period while any (a)periodicity linked to the tract/resonance could be judged on a different time scale. The above question is also of importance for the definition of harmonic-to-noise measures. Any reference to levels in between harmonics is a reference to a noise that is likely to be uncorrelated to the periodicity, e.g. to be randomly distributed over the analysis frame. Depending on the number of periods in the frame, e.g. the harmonic density, this noise can originate from many sources, which makes it an unreliable reference. This does not mean that such a parameter is insensitive, to the contrary, but it is questionable what it discriminates. Again, the overlap-add method allows a check of the above questions by resynthesizing both groups of information.

Method III, killing periodicity

The Fourier analysis principle is based on a phase stability criterion. If a frequency component is stable, it will match to a center-frequency of a band-filter and thus lead to a cumulative result over a given amount of time yielding a spectral peak. To see how effective the tuning/adding is, we have two options: (A) compare amplitudes between two adjoining spectral frames [4], or (B) check the stability of the phase curve as a function of band-filter frequency. Phase stability is weighted in the group delay phase (GDP) function, the differentiated phase curve. The GDP function is a good candidate to mark and thus to remove harmonics from a spectrum. The spectral amplitude is high only due to band-filter phase matching/stability, but the absolute level has no influence on the stabilising principle. A comb-filter design need not be based on a strictly regular pattern in the harmonic series (the pitch).

In our implementation, each harmonic is attenuated using a non-linear mapping function. The resulting output signal shows different degrees of periodicity killing, depending on the averaging time and stability threshold used. In general, the non-harmonic residue illustrates the inaptness of the spectral model to separate period-to-period aperiodicity from noise.

CONCLUSIONS

The first two methods demonstrate the additional role of phase information in the description of voice aperiodicity. The fact that the produced "perturbation" is often perceived as being synthetic, makes us question the completeness and even the validity of common spectral models for voice aperiodicity. The way phase stability is linked to spectral amplitude information, and the way this complex is condensed in the harmonics is vital in our description of period-to-period variability, and the definition of noise as being a non-harmonic residue. Spectral parameters that rate voice (a)periodicity should therefore also include phase (stability) information.

REFERENCES

[1] Allen, J.B., and Rabiner, L.R. (1977). A Unified approach to short-time Fourier analysis and synthesis., *Proc. of the IEEE,* (65), No. 11, pp 1558-1564 .

[2] Moorer, J.A. (1987). The use of phase vocoder in computer music applications, *J. of the Audio Engineering Society,* (26), No. 1/2, pp 42-45.

[3] Pabon, P (1994). A real-time singing voice synthesizer (Alto). *SMAC Proceedings.*. Royal Swedish Academy of Music, Issue No. 79, pp 288-293.

[4] Serra, X. and Smith, J. (1990) Spectral Modeling Synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition, *Computer Music Journal,* (14,) No. 4, pp 12-24.

Time and Frequency Synthesis Parameters of Severely Pathological Voice Qualities

Abeer Alwan, Philbert Bangayan, Jody Kreiman, and Christopher Long***
Department of Electrical Engineering, UCLA, Los Angeles, CA
**Head and Neck Surgery, UCLA, **HST-MIT, Cambridge, MA*

ABSTRACT

This paper describes a pilot study into the mechanics of synthesizing severely pathological voices. Successful synthesis of such voices may ultimately provide a quantitative method for evaluating and documenting voice qualities. An analysis-by-synthesis approach using the formant synthesizer KLSYN was used to model the voices of 24 patients suffering from voice disorders. Results suggest a number of modifications to KLSYN that would facilitate synthesis of these voices.

INTRODUCTION:

No standard system of description exists for pathological voice qualities. Qualities are labeled based on the perceptual judgments of individual clinicians, a procedure plagued by inter- and intra-rater inconsistencies and terminological confusions. Synthetic pathological voices could be useful in creating a standard protocol for quality assessment.

A serious limitation of past studies on synthesizing pathological voices is the focus on single aspects of quality and/or of the acoustic signal. Accordingly, previous studies provide little insight into the techniques necessary to generate reasonable copies of natural pathological voices.

The present study used Sensyn 1.1, the Sensimetrics version of the Klatt formant synthesizer KLSYN [1]. The Klatt synthesizer was chosen because it is commercially available, widely used, and often referenced. In addition, the synthesizer includes a turbulent noise component, pole and zero pairs that can be used to model tracheal or nasal coupling, a provision for time-varying parameters to model unsteady quality, and a "diplophonia" parameter to model bicyclic (period doubled, bifurcated) phonation. However, KLSYN was designed for synthesizing normal voices, and questions remain about its suitability for producing acceptable pathological stimuli.

METHODOLOGY:

Twenty-four voice samples of the vowel /a/ were selected from a library of recordings. Signals were digitized at 20 kHz and then downsampled to 10 kHz, the maximum sampling rate at which all synthesizer parameters could be manipulated. One second segments were excerpted from the middle portion of each natural sample. Stimuli were informally grouped (based on the perceptual judgment of author JK) into the following categories: rough and rough-breathy (11 tokens), bicyclic (8 tokens), rough-bicyclic (1 token), strained-breathy (2 tokens), and strained-rough (2 tokens).

Time and frequency domain analyses of each voice sample were undertaken to guide synthesis efforts. In the time-domain, we tracked long-term amplitude and frequency modulations. In the frequency-domain, the fundamental frequency (F0), formant frequencies, strengths of the first three harmonics, and any additional resonances were measured.

SYNTHESIS PROCEDURES

Synthetic waveforms were modeled after each of the natural tokens using the cascade branch of the synthesizer. Synthesis proceeded as follows.

Step 1: Match Formant Frequencies and Mean F0: As a first step, the formants' frequencies and bandwidths were matched. In addition, the mean value of F0 was used.

Step 2: Adjust Amplitude of Voicing (AV) and Amplitude of Aspiration Noise (AH): When synthesizing pathological voices, matching AH is as important as matching AV because increased breathiness often enhances the perception of rough and bicyclic qualities.

Step 3: Adjust Open Quotient (OQ): The degree of the strained quality in a voice, if present, was matched by altering OQ, which defines the percentage of the pitch period in which the glottis is open.

Step 4: Boost Low Frequency Components: It was often difficult to match the amplitudes of harmonics below F1 in the synthetic voice to those of the natural waveforms. This harmonic mismatch resulted in synthetic voices which did not sound as "rich" as the natural voices. In these cases, additional pole/zero pairs (the nasal and/or tracheal) were placed around the frequencies of the first formant. Typically, one pole/zero pair was placed at the first harmonic, and the other pair, at the second harmonic.

Step 5: Alter F0: F0 was varied to model the natural utterances. For voices with high jitter, F0 values were generated such that they followed a Gaussian distribution, given a mean and variance calculated from the natural sample. In other cases (particularly with bicyclic voices), values were measured manually from the natural waveform and imported into the synthesizer. For bicyclic voices where F0 values were not entered manually, the diplophonia (DI) and flutter (FL) parameters were used. DI was useful for synthesizing some bicyclic voices, but failed to capture the pattern of F0 and amplitude alterations for others. Flutter creates slowly varying and regularly repeating F0 values, as described by:

$$\Delta F_0 = \frac{FL}{50}\frac{F0}{100}\left[\sin(2\pi 12.7t) + \sin(2\pi 7.1t) + \sin(2\pi 4.7t)\right]$$

Step 6: Alter AV as a Time-Varying Parameter for Amplitude-Modulated Voices: The parameter AV was altered in a time-varying fashion to model shimmer.

Step 7: Add Additional Pole/Zero Pairs if Necessary: Some voices required additional pole/zero pairs to model nasal and/or tracheal coupling.

In some cases, the speed quotient of the glottal waveform (parameter SQ) was adjusted to match the overall spectral shape.

These steps were repeated as necessary to fine-tune the synthesis, until the synthesized voice was judged to be a reasonable match to the original waveform.

SYNTHESIS RESULTS

Rough and Rough-Breathy Voices

Eleven rough and rough-breathy voices (4 female, 7 male) were analyzed. Seven samples had fairly steady qualities, but four voices varied considerably. Capturing the variation in F0 proved critical for successful synthesis of these voices. Seven of the ten voices required time-varying AV, and the rough voices were generally accompanied by turbulent noise. Eight out of the ten voices had OQ > 50%.

Bicyclic Voices: Bicyclic voices (also referred to as diplophonia or bifurcated phonation) present a pattern of cycles that alternate in frequency, amplitude, or both, in a large-small-large-small pattern. Eight bicyclic voices (4 female, 4 male) were analyzed. None showed a perfect pattern of periods alternating in an ABAB fashion. Instead, three patterns emerged: (1) three voices had fundamental frequencies alternating among a small number of values (typically 5 to 9); (2) F0 was bimodally distributed for 4 voices; and (3) one voice had increasing bicyclicity with time.

The male voices tended to sound more strained and also had weaker first harmonics than their female counterparts. Hence, the OQ was less than 50% for all the male voices, but only for one female voice.

Strained-Breathy and Strained-Rough Voices: These were the most difficult voices to synthesize. Attempts were made to capture the strained yet breathy quality by changing the speed quotient, changing bandwidths, sequentially altering OQ, time-varying AH to modulate breathiness, utilizing FL and

and varying F0. None of these techniques proved entirely successful.

The strained-rough voices were unsteady during the periods when the voices become strained. Techniques used to model these voices included time-varying SQ, OQ, and AV. The gargly nature and unsteadiness of the voices were not captured well in the synthesized versions.

PERCEPTUAL EVALUATION

As suggested above, some attempts at synthesizing pathological voices were subjectively more successful than others. The following experiment was undertaken to evaluate the overall quality of the synthesis, and to determine which voices listeners considered good matches to the original samples.

Methods

Ten expert listeners participated in this experiment. The 24 voices described above were used as stimuli. Stimuli were normalized for peak voltage, and onsets and offsets were multiplied by 25 ms ramps to eliminate click artifacts.

Listeners heard each natural sample paired with its synthetic copy, and were asked to judge how well the copy matched the original on a 1-7 scale (1: perfect match). Stimuli were presented in free field at a comfortable listening level.

Results

Listeners unanimously reported being pleased by the overall quality of the synthesis. Mean ratings ranged from 1.30 to 6.30. As Figure 1 shows, listeners agreed well in their ratings when they thought that a copy was nearly identical to the original sample. For less successful copies, both the mean rating and the variability in ratings increased.

On the whole, copies of male voices (filled circles in Fig. 1) were more successful than were copies of female voices (asterisks in Fig. 1). Most "unsuccessful" ratings reflect failures to model unsteady or gargled qualities or failures in modeling voices with strong low frequency components and a muffled quality. The spectrogram of an unsteady voice which was difficult to synthesize is shown in Fig. 2.

SUMMARY AND DISCUSSION

On the whole, our efforts at synthesizing severely pathological voices were fairly successful. Less severe pathological voices (in particular, male voices with steady F0 contours) were synthesized best. Our results suggest that several modifications to the Klatt synthesizer would improve the quality of synthesis, and would facilitate the production of a wide range of pathological qualities. (1) More than six formants are needed to synthesize voices at high sampling rates. KLSYN provides enough variable formants to support sampling rates of 10-12 kHz. Providing more formants, with variable frequencies and bandwidths, would alleviate this difficulty. (2) A parameter is needed to increase the spectral energy below F1. KLSYN-synthesized voices often lack energy at frequencies below the first formant. One solution is to increase the open quotient (OQ), which increases the first harmonic energy. Another is to add pole/zero pairs below F1. Both solutions were often inadequate. A new parameter that boosts the harmonics below F1 would provide more low-frequency energy. (3) More pole/zero pairs are needed to account for increased coupling to the nasal tract or to the trachea in pathological cases. (4) Jitter and shimmer parameters are needed to facilitate modeling of perturbations in natural voices. In this study, jitter and shimmer were modeled by manually entering the time-varying values of F0 and AV. This approach is cumbersome to use. KLSYN does offer a flutter parameter (FL) which slowly time-varies F0, but this does not model jitter appropriately. (5) The diplophonia parameter (DI) should be split into separate F0 and amplitude perturbation parameters. DI was designed to model bicyclic (bifurcated, period doubled) phonation, and as presently implemented it attenuates and delays every other glottal pulse. The resulting patterns of amplitude and frequency variation do not match measurements of natural bicyclic waveforms, for which there is no consistent correlation between amplitude and F0. Modeling would be improved by allowing amplitude to be changed independent of delay, and/or by allowing amplitude to be specified for each individual period. (6) The update interval (UI) should be implemented as a time-varying function that could be updated at any time instant and not necessarily at the beginning/end of a pitch period. Some parameters, such as F0, are updated at the end of each period while most other time-varying parameters (such as AV) are specified at multiples of UI, and linear interpolation is used to determine values between updates. This is problematic when one needs to change attributes of one glottal pulse without affecting other pulses.

Finally, more acoustic modeling of severe vocal pathology is necessary. As discussed above, most acoustic models are based on variations in normal speech, and do not easily accommodate pathologic cases. Effective synthesis of strained/breathy voices and gargled qualities in particular must await improvements in modeling as well as improvements in synthesizers.

ACKNOWLEDGEMENT:

This research was supported in part by NIDCD grant DC 01797.

REFERENCES:

[1] Klatt, D.H., and Klatt, L.C. (1990). "Analysis, Synthesis and Perception of Voice Quality Variation among Female and Male Talkers," JASA, 820-857.

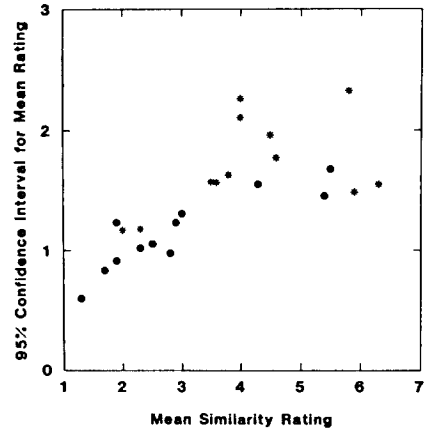


Fig. 1: Variability in perceptual ratings versus mean similarity of the synthetic tokens to the natural voices, as judged by ten expert listeners.

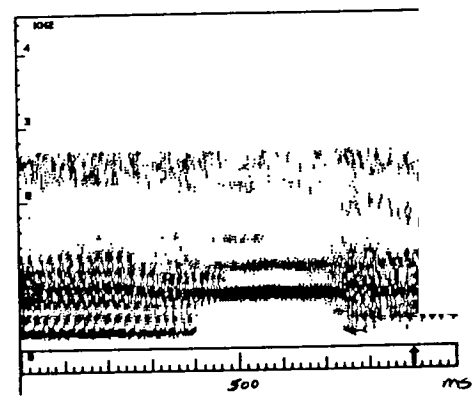


Fig. 2: Spectrogram of an unsteady female voice which was the most difficult voice to synthesize.

An Analysis-by-Synthesis Approach to the Estimation of Vocal Cord Nodule Features

Takuya Koizumi, Shuji Taniguchi, Mikio Mori, and Akemi Imazawa
Dept. of Information Science, Fukui University
3-9-1 Bunkyo, Fukui 910, Japan

ABSTRACT

This paper deals with a new non-invasive method of estimating vocal cord nodule features through hoarse voice analysis. A noteworthy feature of this procedure is that it enables us to estimate vocal cord nodule features such as the mass and dimensions of nodules through the use of a novel model of pathological vocal cords which has been devised to simulate the subtle movement of the vocal cords with nodules.

INTRODUCTION

A number of studies on acoustic analysis of hoarse voice which have been reported in the last decade may be divided into two groups by their objectives. Some studies tried to establish techniques of discriminating voices caused by pathological vocal cords from normal voices, while others aimed at developing a procedure for the classification of laryngeal disease only through the analysis of hoarse voice. However, few of them have ever attempted to develop a non-invasive procedure for estimating the pathological condition of the larynx [1].

An attempt toward this end has recently been made, resulting in the development of a new noninvasive procedure for the estimation of vocal cord nodule features through hoarse voice analysis. A noteworthy feature of this procedure is that it enables us to estimate vocal cord nodule features such as the mass and dimensions of nodules through the use of a novel model of pathological vocal cords which has been devised to simulate the subtle movement of the vocal cords with nodules. This method, which might well be called an analysis-by-synthesis approach, is characterized by a fact that a hoarse-voice synthesizer comprising the model of pathological vocal cords and a vocal tract model is used to estimate the nodule features along with an acvoustic distance measure for hoarse voices defined as a function of glottal volume flow waveform and power spectral density of glottal turbulent noise.

A synthetic hoarse voice produced with this hoarse-voice synthesizer is compared with a natural hoarse voice produced by pathological vocal cords with nodules in terms of the distance measure, and values of the nodule features that minimize the distance measure are chosen as reasonable estimates of them. The effectiveness of this method can be examined by comparing the estimates of dimensions of nodules with actual dimensions of those nodules. Some estimates of nodule dimensions that have been obtained by applying the procedure to hoarse voices of patients who have laryngeal nodules are found to compare favorably with actual nodule dimensions, demonstrating that the procedutre is effective.

THE HOARSE-VOICE SYNTHESIZER

An Eight-mass Model of Pathological Vocal Cords

The hoarse-voice synthesizer cosists of a vocal tract model and a model of pathological vocal cords. The pathological vocal cords can be modeled as a mechanical vibration system made up of four independent masses coupled by nonlinear springs and dampers. To properly model the vocal cords with a couple of nodules, however, it is necessary to add a couple of split masses representing the nodules to the four mass model. This results in an eight-mass model of the vocal cords with nodules as shown in Figure1. A brief description of this model will be given below. The upper and lower edges of the vocal cords are represented by the masses m_1, m_2 and M_1,M_2, respectively.

The split masses m_{iN} and M_{iN} (i=1,2, $m_{1N}=m_{2N}$, $M_{1N}=M_{2N}$) representing the nodules are attached to the upper and lower masses, respectively, and therefore move synchronously with the upper and lower masses as their integal parts. Most of laryngeal nodules seem to have a paraboloid-like shape, however, for the

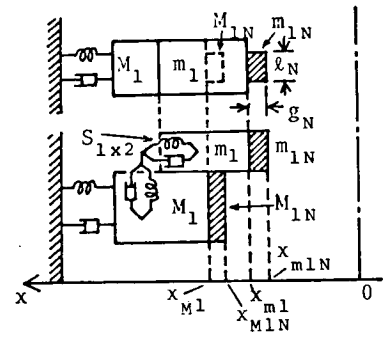


Figure 1. The structure of the eight-mass model of vocal cords with nodules.

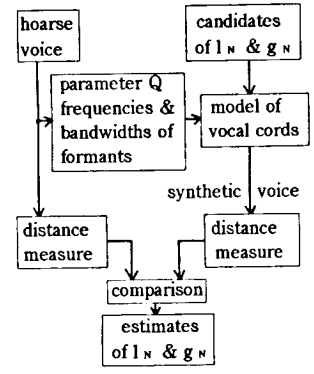


Figure 2. The method of estimating the nodule features.

simplicity of analysis it is assumed that the masses m_{iN} and M_{iN} are in the form of a rectangular parallelpiped and that those masses and their counterparts on the opposing vocal cord are placed in a proper position to collide with each other, when the model is in motion. The nodules are considered to cause a hindrance in one way or another to the motion of vocal cords known as the mucosal surface wave. This effect can be expressed as an increase in the stiffness K_{x2} of the spring S_{ix2} coupling the upper and lower masses, which is denoted as ΔK_{x2} in the following. ΔK_{x2} is thought of as a function of the width l_N of the nodules.

The Hoarse-voice Synthesizer

The hoarse-voice synthesizer consists of a glottal volume flow source and a vocal tract model. The glottal volume flow source comprises a subglottal system, a glottal impedance controlled by the eight-mass model of the vocal cords, an aspiration noise source, and a first formant load. An incomplete closure of vocal cords caused by the nodules gives rise to an aspiration noise due to turbulent glottal flow during the closed-glottis condition. To take this phenomenon into account the glottal volume flow source is provided with a noise source controlled by the vocal cord model, which supplies the glottal volume flow source with the aspiration noise when the Reynold's number for the slit formed by the opposing masses of the model exceeds a critical value. The vocal tract model that is made up of five formant filters and a simple radiation load produces the output P_0 which is a synthetic hoarse voice, when excited by the output U_G of the glottal volume flow source.

THE METHOD OF ESTIMATING THE NODULE FEATURES

The procedure for estimating vocal cord nodule features is shown in Figure 2. For an input voice sample of, say, vowel /a/, which has been uttered by a patient who has pathological vocal cords with nodules, the nodule features, i.e., dimensions of the nodules are estimated through the use of an optimization technique to be described below.

A Feature Vector for Hoarse Voice

A specific feature vector is used to describe ascoustic properties of hoarse voice. A part of this vector represents an estimated noise power spectrum of hoarse voice, which is calculated by subtracting fundamental and higher harmonic frequency components of vocal cord oscillation from power spectrum of hoarse voice. This noise power spectrum is represented as a 32-dimensional vector. This vector is augmented by a 10-dimensional glottal volume flow waveform vector to yield a 42-dimensional feature vector for hoarse voice. The glottal volume flow waveform vector has as its components real and imaginary parts of the short-time discrete Fourier transform (DFT) of hoarse voice which has been filtered using a low-pass filter with cutoff frequency of 1.5 kHz. Those real and imaginary parts of the DFT of the

hoarse voice have been calculated using amplitudes and phases relative to the phase of the fundamental frequency component of the first five harmonic frequency components of the hoarse voice.

Estimation of the nodule features

It is possible to estimate the nodule features from a given hoarse voice uttered by a patient who has pathological vocal cords with nodules by maximizing a similarity in terms of the distance measure between the hoarse voice and a synthetic hoarse voice produced with the hoarse-voice synthesizer. The method of estimating the nodule features in this work is based on this analysis-by-synthesis approach. It is illustrated in Figure 2. For a stationary portion of the input voice a frequency spectral analysis is performed by means of the FFT and comb filtering to obtain the aforementioned augmented vector. This spectral analysis is necessary to produce a synthetic hoarse voice which is similar to the input voice in the sense described above. In addition to the frequency spectra several other important acoustic parameters have to be extracted from the input voice. First the input voice is subjected to the linear predictive analysis, and frequencies and bandwidths of the first five formants are estimated. Then an average fundamental period is calculated from a sequence of the fundamental period to determine a pitch control parameter Q. Using these parameters and a set of m_{iN}, M_{iN}, K_{c2}, and the width of the nodules l_N specified somehow in the hoarse-voice synthesizer will result in a synthetic hoarse voice which may or may not be similar to the input voice.

If the distance measure which will be defined in the next section is considered as a function of the mass $m_{iN}+M_{iN}=M_N$ and width l_N of nodule, then the problem of estimating the nodule features will be equivalent to that of finding a minimum of a surface representing the distance measure over the l_N-g_N plane. Here g_N represent the thickness of nodule. To find the minimum requires the evaluation of the distance measure at a number of points in the l_N-g_N plane. Finding the position of the minimum provides us with the desired nodule dimensions.

The Distance Measure

The distance measure used for the aforementioned purpose comprises the components of the feature vector defined previously, i.e., the estimated noise power spectrum and glottal volume flow waveform spectrum. It is given by

$$C=W_1 \sum_{i=1}^{32} (P_{i1}-P_{i2}) + W_2 \sum_{j=1}^{5} \{(R_{j1}-R_{j2})^2 + (I_{j1}-I_{j2})^2\}, \quad (1)$$

where P_{i1} and P_{i2} denote estimated noise power spectra of input and synthetic voices, respectively, and R_j and I_j are real and imaginary parts of the jth harmonic frequency component of the input or synthetic voice, respectively. R_j and I_j are calculated using amplitude and phase relative to that of the fundamental frequency component of the jth harmonic frequency component. The subscript 1 denotes the input voice, and the subscript 2 the synthetic voice. The weights W_1 and W_2 have been chosen as follows: $W_1=3$, $W_2=40$.

RESULTS OF EXPERIMENT

The method of estimation was applied to hoarse voices /a/ uttered by six patients who have vocal cord nodules. Since video pictures of vocal cords of those patients were available, it was possible to somewhat precisely measure dimensions of the laryngeal nodules from the pictures, taking into account a fact that the average length of the vocal cords is 14 mm for adult men and 10.5 mm for adult women. Estimates of dimensions, i.e., the width and thickness, of the nodules obtained by the method of estimation are shown in Table 1 along with measured dimensions of them.

In the model of vocal cords with nodules the incomplete glottal closure necessarily occurs, however, in actual vocal systems of patients who have vocal cord nodules it does not necessarily occur, because opposing vocal cords which are made up of a soft mucosal tissue called cover and flexible body can collide with each other even in the presence of nodules.

Some of the estimates, spesifically for voice samples 4, 5, and 6 corresponding to small nodules, are in large error because of a mismatch mentioned above between the model of vocal cords and actual vocal cords. Estimates of nodule dimensions for voice samples 1, 2, and 3

Table 1. Estimates of dimensions of the nodules obtained by the estimation scheme and corresponding measured dimensions of the nodules.

Voice Sample	Measured l_N (mm)	Estimate of l_N (mm)	Estimation Error(%)	Measured g_N (mm)	Estimate of g_N (mm)	Estimation Error(%)
1	2.96	3.15	+6.4	0.78	0.95	+21.8
2	2.11	3.15	+49.3	0.91	0.95	+4.4
3	2.96	3.15	+6.4	0.71	0.14	−80.3
4	2.87	3.15	+9.8	0.48	0.95	+97.9
5	2.16	3.15	+45.8	0.54	0.63	+16.7
6	1.63	3.15	+93.3	0.66	0.79	+19.7

corresponding to larger nodules are in agreement with actual dimensions of nodules within estimation errors less than 20 %.

Figure 3 shows waveforms of synthetic hoarse voice (sound pressure) and glottal volume flow, glottal areas, and displacements of masses for the voice sample 1. The output sound pressure and glottal volume flow waveforms are found to be rather irregular because of glottal turblent noise. The displacements of masses clearly show the occurrence of incomplete closure of the opposing upper and lower masses.

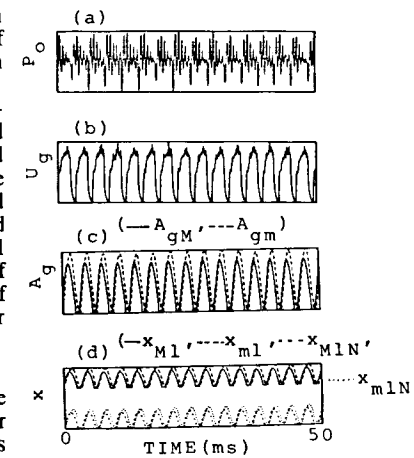


Figure 3. Waveforms derived from the estimation scheme: (a) synthetic hoarse voice(sound pressure), (b) glottal volume flow, (c) glottal areas, and (d) displacements of masses for voice sample 1.

CONCLUSIONS

In this study a new noninvasive procedure have been developed for estimating vocal cord nodule features through the use of a novel model of pathological vocal cords with a couple of nodules. This model is able to simulate the subtle movement of vocal cords with nodules in the presence of aspiration noise in the glottis. By this newly developed procedure it is possible to estimate dimensions of laryngeal nodules only through hoarse voice analysis.

This procedure has been applied to several hoarse voice samples and has been shown to be capable of estimating the state of vibration of vocal cords and dimensions of vocal cord nodules with satisfactory accuracy for large nodules. For smaller nodules, however, the estimates of dimensions of nodules were found to be in error because of the mismatch between the model and actual vocal cords.

REFERENCE

[1] Koizumi, T., Taniguchi, S., and Itakura, F. (1993), "An Analysis-by-Synthesis Approach to the Estimation of Vocal Cord Polyp Features", Laryngoscope, vol.103, pp.1035-1042.

THE ROLE OF TRANSITION VELOCITY IN THE PERCEPTION OF V_1V_2 COMPLEXES

Pierre Divenyi
Speech and Hearing Research, V.A. Medical Center, Martinez, CA, USA
Björn Lindblom
Department of Linguistics, Stockholm University, Stockholm, Sweden
René Carré
C.N.R.S. and Ecole Nationale de Télécommunications, Paris, France

ABSTRACT

V_1V_2 tokens were digitally synthesized, with the transition plus the V_2 segments cut back to various degrees. Listeners were asked to either identify the vowel at the end of the stimulus (Exp. 1) or to judge its proximity to a designated target (Exp. 2). Results suggest that the overshoot effect in the dynamic perception of vowels may be at least partially attributed to cochlear processes. Thus, while results are consistent with a gesture-bound theory of speech perception, they also support alternative accounts.

INTRODUCTION

It has been known for some time that, for a vowel embedded in a C_1VC_1 or $V_1V_2V_1$ context to be recognized, its formant frequencies do not need to reach the values characteristic to those of the same vowel in isolation [1]. This phenomenon has been termed "vowel reduction" or "perceptual overshoot." Recent results in this area overwhelmingly suggest that vowel reduction is an integral part of the production and perception of connected discourse [2][3], and it represents an accessible point of entry into the study of the dynamic aspects of speech. One aspect of vowel reduction and the associated perceptual overshoot that has not received sufficient attention is its dependence on the velocity of vocalic transitions [4]. The experiments presented in this paper investigated the effect of transition velocity on vowel perception and addressed two specific questions: (1) Can formant transitions leading to a certain vowel target define the target as a function of the transition velocity alone? (2) Is the trajectory leading from vowel V_1 to vowel V_2 perceived identically to the trajectory leading from V_2 to V_1?

EXPERIMENT 1

V_1V_2 samples were generated digitally using a PC computer. Each sample had a falling f_0-contour corresponding to that of a male voice. In Experiment 1, the two vowels were selected at either endpoint or midway between the linear trajectory between the two French vowels /a/ and /i/, represented in the F1-F2 plane. The vowel exactly bisecting this trajectory was one which French-speaking listeners identified as an acceptable token of /ɛ/. The trajectory in the F1-F2 vowel space is illustrated in Fig. 1a. The duration of the V_1 segment was held constant at 100 ms. Two velocities of frequency change were synthesized, one covering the distance between /a/ and /i/ in 100-ms, and the other in 200 ms. The trajectory thus reached the vowel /ɛ/ in 50 or 100 ms, respectively. A 100-ms terminal steady-state portion of V_2 was appended to the transition. From each V_1-V_2 complex, a series of tokens were generated by cutting back an increasingly longer segment from the end of the complex. The duration difference between any two adjacent tokens was 5 ms. Four V_1V_2 pairs, /ai/, /ia/, /aɛ/, and /iɛ/, were investigated. A schematic of a

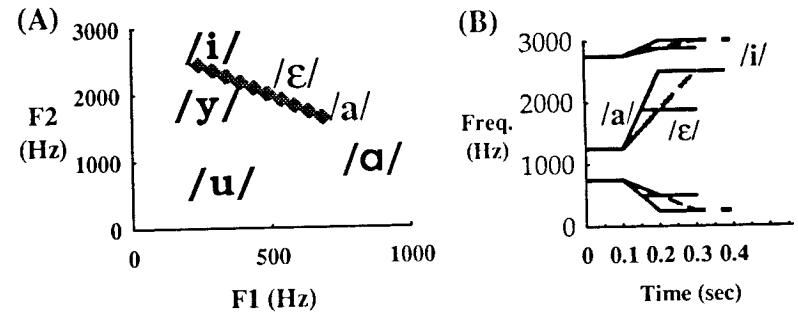


Figure 1. Stimuli used in Experiment 1. A: F1-F2 plane representation of the /il-/al transition trajectory. B: Spectrographic representation of /ail and /aɛ / at the two transition velocities (shown as different line styles).

sample stimulus is shown in Figure 1b. Three experienced subjects served as listeners. They had to indicate, by key press, which of the four vowels /a/, /e/, /ɛ/, or /i/ sounded most similar to the final vowel of the stimulus they just heard. Blocks of stimuli contained 10 repetitions of 10 to 16 different tokens. Each token was introduced by the monosyllabic word "dis" (="say").

Figures 2a-d illustrate combined results of three subjects. Each panel repre-

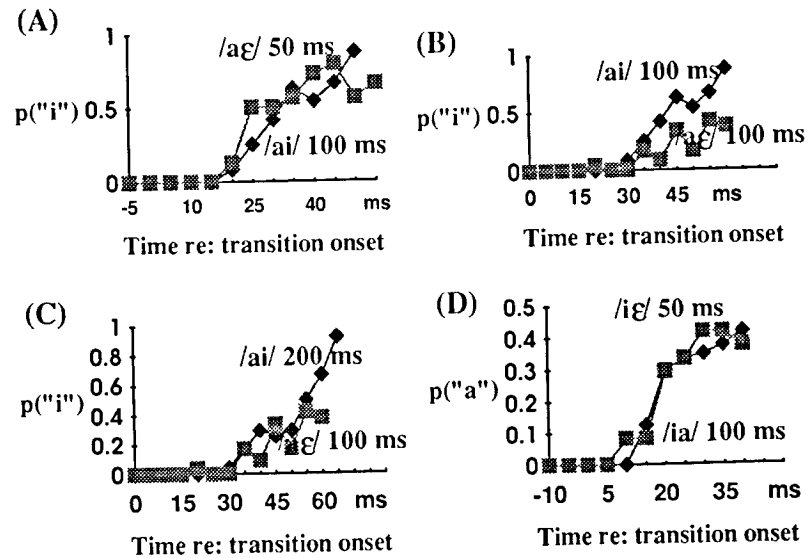


Figure 2. Results of Experiment 1. A: Percentages of a final "i" reported for /ai/ with a 100-ms transition and for /aɛ/ with a 50-ms transition (solid lines in Fig. 1b). B: Same as "A", but with 100-ms /aɛ/ transition. C: Same as "B", but with 200-ms /ai/ transition. D: Percentage of "a" reported for /ia/ with 100-ms and /iɛ/ with 50-ms transitions. Note the compressed ordinate scale. Averaged data for three listeners.

sents the proportion of the responses in which the terminal vowel segment was heard as the *remote* anchor of the trajectory, i.e., /i/ for the /ai/ and /aɛ/ transitions, and /a/ for the /ia/ and /iɛ/ transitions. The three main results are: (1) For identical Hz/ms transition rates, responses to the /ai/-/aɛ/ and /ia/-/iɛ/ pairs overlap even beyond the onset of the steady-state /ɛ/. (2) For the low

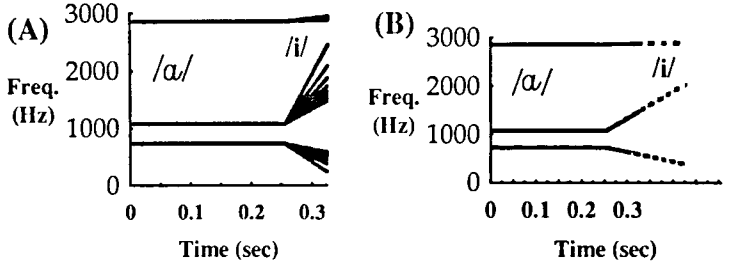


Figure 3. Spectrographic representations of stimuli used in Experiment 2. A: F1-F2-F3 of the eight /ia/ tokens used in the constant transition duration series. B: F1-F2-F3 of two of the eight /ai/ tokens used in the constant transition velocity series.

transition rate, there is little or no overshoot. (3) The /ia/-/iɛ/ and /ai/-/aɛ/ transitions are judged asymmetrically.

EXPERIMENT 2

Experiment 2 was designed to test for overshoot along trajectories outlined by the three corners of the vowel triangle with series of tokens differing *either* in transition velocity *or* in transition duration. Two series of eight tokens were generated from the six V_1V_2 complexes */ai/-/ia/*, */au/-/ua/*, and */ui/-/iu/* with the final formant frequencies of the tokens covering the V_1-V_2 formant space, with transition *duration* (Fig. 3a) or transition *velocity* (Fig. 3b) fixed. Four trained subjects served as listeners and *rated*, on a four-point scale, the *proximity* between the V_2 target and the terminal vowel segment of the token.

From the results, we estimated F1 and F2 values of the final vowel which was judged with a 50 percent confidence to be the target. Figure 4a shows these frequencies for the constant-duration and Fig. 4b for the constant-velocity series and demonstrate consistent overshoot. The largest overshoots are found in the constant duration conditions, indicating that, as the transition velocity increases, the final vowel is increasingly heard as the V_2 target. Although the overshoot for the constant-velocity vowel pairs is more modest, the variability is also less. The small size of these overshoots may be due to the averaging process that computes the pitch of sounds with changing frequencies [5]. Large directional asymmetry was observed only for one vowel pair and only in one series.

DISCUSSION AND CONCLUSIONS

The above two experiments demonstrate that, although contrast and linguistic context may influence vowel recognition, they do not represent *sine qua non* conditions for the phenomenon of overshoot to occur. The present results suggests that peripheral auditory processing may play a substantial role in the dynamic perception of vowels. In Fig. 4, it seems that the overshoot is associated with either a large F1 difference (*/ua/-/au/*) or an F2 trajectory that traverses low-frequency (<1000 Hz) regions. Since adjacent harmonics are individually resolved at low frequencies, both inter-

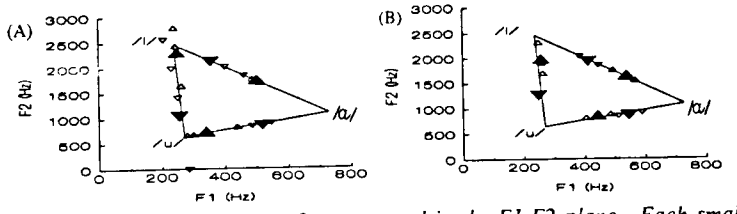


Figure 4. Results of Experiment 2 represented in the F1-F2 plane. Each small open triangle represents data from one of the four subjects and the large filled triangles, the average data for any condition. Data represent estimated formant frequencies for terminal vowels judged to reach the target V_2 with a 50-percent confidence rating. Each triangle points toward the target V_2 frequency, e.g., on the trajectory between /u/ and /i/ triangles pointing downward are data for /iu/, and those pointing upward are data for /ui/ conditions. Panel A: results of the constant transition duration series. Panel B: results for the constant transition velocity series.

harmonic and broad (especially low-by-high-frequency) *suppression* may shift the *effective* formant peak (i. e., the peak in the peripheral excitation pattern) away from another, relatively high-energy region [6].

Fig. 2 also suggests that the presence and extent of the overshoot depend heavily on the *slope* of the transition. In fact, the intended target may not even matter: An /aɛ/ doublet will generate the percept of a terminal /i/ as long as the transition velocity is identical to that of an /ai/ doublet. This illusion is strongest at the very beginning of the transition and remains quite compelling as long as a final steady-state /ɛ/ is absent. Since in the natural production of such vowel pairs the first part of the transition coincides with a period of maximum acceleration of the articulators (i.e., a period of maximum force), our data are consistent with a speech-gesture interpretation but also with certain alternative accounts.

ACKNOWLEDGEMENT

The authors would like to thank Drs. Steven Greenberg, Shinji Maeda, and John Ohala for their patiently expressed opinions on the ideas discussed in this paper. The research was supported by NIH and VA Medical Research in the U.S. and by the E.U. Science Project.

REFERENCES

[1] Lindblom, B. and M. Studdert-Kennedy (1967), "On the role of formant transitions in vowel recognition", *Journal of the Acoustical Society of America*, 42: pp. 830-843.

[2] van Son, R.J.J.H. (1993), "Vowel perception: A closer look at the literature", *Proceedings of the Institute of Phonetic Sciences, University of Amsterdam*, 17: pp. 33-64.

[3] van Wieringen, A. (1995), *Perceiving dynamic speechlike sounds*. University of Amsterdam (the Netherlands):

[4] Pols, L.C.W. and R.J.J.H. van Son (1993), "Acoustics and perception of dynamic vowel segments", *Speech Communication*, 13: pp. 135-147.

[5] Nabelek, I.V., A.K. Nabelek, and I.J. Hirsh (1973), "Pitch of sound bursts with continuous or discontinuous change of frequency", *Journal of the Acoustical Society of America*, 53: pp. 1305-1315.

[6] Weber, D.L. and D.M. Green (1978), "Temporal factors and suppression effects in backward and forward masking," *Journal of the Acoustical Society of America*, 64: pp. 1392-1399.

EFFECT OF PRIOR KNOWLEDGE OF THE VOWEL ON THE PERCEPTION OF FRENCH STOP BURSTS

Linda Djezzar
CRIN-CNRS & INRIA Lorraine
B.P. 239 54506 Vandœuvre-lès-Nancy France

ABSTRACT

This paper presents a perceptual experiment on the effect of prior knowledge of the vowel on the identification of French stop bursts in natural speech. In order to evaluate the discrimination power of only spectral characteristics of the burst, stimuli consisted of fixed-length bursts of approximately 25 ms (neither VOT nor transitions are present). Results showed that knowing the identity of the following vowel caused a slight but statistically significant improvement of stop identification.

1 INTRODUCTION

Several studies have indicated that the release burst provides reliable information to correctly identify the stop place of articulation [1] [5] [6]. Nevertheless, in current recognition systems, the recognition rates of the palatovelars in front contexts and the dentals in rounded contexts are still far from human performance. Moreover, questions still remain about the role of the knowledge of the identity of the adjacent vowel on the identification of stops. In order to better understand the discrimination power of the burst for identifying the stop articulation place, three perceptual experiments were carried out: the first tested listeners' ability to identify the burst plus a segment of the following vowel, the second investigated the identification of fixed-length bursts independent of the knowledge of the following vowel, the third dealt with the effect of knowing the vowel on this identification. The first and the second experiments are described in [2] [3]. In the following section, we present them briefly in order to introduce the third experiment which is the central issue of this paper.

2 PERCEPTION OF STOP BURSTS WITHOUT KNOWING THE VOWEL

2.1 Preliminary experiment

We checked the listeners' ability to identify i) bursts plus a segment of the following vowel, and ii) bursts only. The results indicated that the presence of a vocalic segment, even a very short one, allowed an almost perfect stop identification. The high identification rates obtained for the burst-only stimuli encouraged us to propose the following experiment in which we tried to clarify the contribution of just the spectral burst characteristics.

2.2 Identification of fixed-length bursts

2.2.1 Corpus and stimuli

The stimuli were extracted from a corpus made of CVC and CV syllables, in which each of /p, t, k/ appeared in combination with each of the 11 vowels /i, e, y, ø, ɛ, a, œ, ɔ, u, o, ɔ̃/. The stimuli consisted of burst portions of the same duration (25 ms) with no remaining vocalic segment. Figure 1 presents the spectrogram of a few stimuli.

2.2.2 Major results

The average identification was 87%. Table 1 shows that, independent of context, all the identification rates were not only well above chance but also very high. A three way repeated analysis of variance, ANOVA, was performed in order to estimate the effects of the audition (2 auditions, one per session), of the vowel (8 vowels) and the stop (3 stops). Moreover, we used the Scheffé test for all post-hoc comparisons. These tests indicated significant main effects for all the parameters but only one significant interaction, the stop-vowel one.

To summarize, the burst onset provides reliable spectral information about the place of articulation of the stops, independent of explicit knowledge of the identity of the vowel. Neverthless, listeners' performance varied significantly depending on the following vowel and on the syllable. Does the knowledge about the identity of this vowel improve stop recognition, at least in the worst contexts? This question is the object of the following experiment.

Table 1. Consonant confusion matrice for fixed-length stimuli in eleven unknown vocalic contexts.

	p(%)	t(%)	k(%)
/pi/	89	6	5
/ti/	0	90	10
/ki/	0	28	72
/pe/	89	9	2
/te/	2	94	4
/pɛ/	87	11	2
/tɛ/	1	85	14
/kɛ/	1	28	71
/pa/	87	10	3
/ta/	15	79	6
/ka/	0	28	72
/py/	83	7	10
/ty/	0	72	28
/ky/	3	13	84
/pø/	97	2	1
/tø /	0	83	17
/kø /	2	6	92
/tœ/	0	98	2
/kœ/	1	12	87
/pu/	85	6	9
/tu/	0	97	3
/ku/	0	2	98
/po/	91	8	1
/to/	1	86	13
/ko/	0	1	99
/pɔ/	93	4	3
/tɔ/	8	83	9
/kɔ/	1	1	98
/pɔ̃/	92	6	2
/tɔ̃/	8	77	15
/kɔ̃/	2	0	298

3 EFFECT OF KNOWING THE IDENTITY OF THE VOWEL

Winitz *et al.* [8] have shown that, when listening to the entire burst of /p, t, k/, subjects could identify the adjacent vowel better than chance. According to Repp and Lin [6], the explicit knowledge of the following vowel slightly improved (3%) the consonant recognition. In order to verify whether Repp and Lin's results would be the same for natural (non whispered) bursts, we conducted an experiment to test the influence of the explicit knowledge of the vowel on the identification of the stops.

3.1 Identification of fixed-length bursts with knowing the vowel

3.1.1 Corpus and stimuli

The corpus was made up of 90 tokens: /p, t, k/ uttered twice by 5 male speakers, in each of the 3 vocalic contexts /i, a, u/. We used all the burst stimuli of the "vowel unknown" experiment appearing in /i, a, u/ contexts.

3.1.2 Subjects and procedure

The subjects were submitted to 3 sessions on 3 days, each day a session. The first session was a familiarization task, which consisted in hearing the stimuli of the preliminary experiment. The second session included 3 stages, one per vowel. At these stages, the identity of the vowel was revealed to the listener. Each stage was composed of a training phase, a test then a rest. The training phase was divided in 3 tasks.

- First, the subjects listened to the training corpus (/p, t, k/ appearing before each of /i, a, u/, and uttered twice by 3 male speakers) and simultaneously read the corresponding answers.
- Second, they underwent a test on the same corpus, but randomized, and they were asked to choose their response from among /p, t, k, ?/, the symbol /?/ means they were unable to supply an answer.
- Third, they compared their false responses by looking at the correct answers, and by simultaneously listening to the sounds.

The test task consisted in hearing the test corpus devised for one of /i, a, u/ (30 stimuli). The third session is a replication of the second one but the corpora were randomized differently. In this way, we obtained 2 auditions for each stimulus.

3.1.3 Acoustic analysis of the stimuli

In order to interpret the perceptual results, we analyzed the release burst of voiceless stops in all vocalic contexts.

- **Palatovelars.** As might be expected, the frequency of the most prominent peak varied as a function of both the place of articulation and the degree of rounding of the following vowel. More precisely, the mean values were 1020 Hz in the back context /u, o, ɔ, ɔ̃, ɑ̃/, 2500 Hz in the central context /a, ɛ̃, œ/ and 2800 Hz in the front context /y, ø, i, e, ɛ/.
- **Dentals.** In the rounded context, they had a prominent peak at a relatively low frequency, at approximately 2500 Hz. In the central context, the spectra of our dental stimuli were either flat or prominent at 1800 Hz (*Locus*).

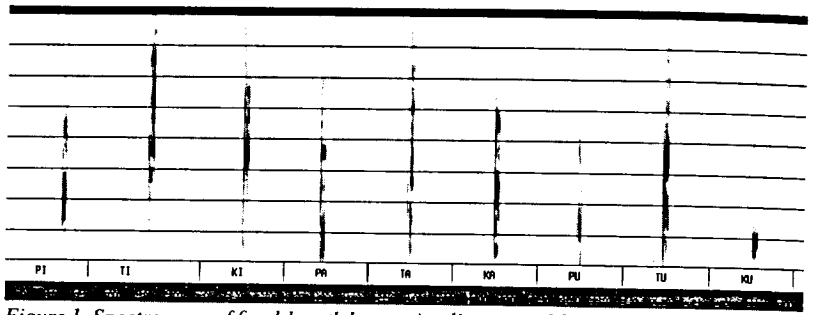


Figure 1. Spectrogram of fixed-length burst stimuli extracted from the syllables /pi, ti, ki, pa, ta, ka, pu, tu, ku/.

• **Labials.** The maximum was not situated in a well defined region, even for the same vocalic context. Nevertheless, the energy was stronger in low frequencies in the back context, and in mid frequencies in the central and front contexts.

Moreover, the global form of the spectrum did not always fit the templates proposed by Blumstein and Stevens [1]. In fact, the spectra of our dental stimuli followed by rounded vowels were not diffuse and those of our palatovelar stimuli followed by centrals and unrounded front vowels were not always compact [4].

3.1.4 Results

The overall identification rate was 89%, 4% higher than the identification rate obtained in the "vowel unknown" experiment for the same vocalic contexts /i, a, u/. A three way repeated analysis of variance was conducted in order to estimate the effects of the audition (2 auditions, one per session), of the vowel (3 vowels) and the stop (3 stops).

There was only one significant main effect, the vowel one [$F(2,166)=13.28$, p.c.<0.001], and one significant interaction [$F(4,166) = 7.67$, p.c.< 0.001], which occurred between the consonant and the vocalic context. According to the Scheffé test, /u/ was considered as the best vocalic context, and the interaction between /k/ and /i/ reduced the intelligibility of the consonant. Let us interpret briefly these results.

• **Context /u/.** In this context, the very high rate of identification of /k/ is most probably due to its particular spectral form (2 prominent peaks separated by a region without energy: generally the most prominent was situated at the F2 of the following vowel and the second one at 4000 Hz). Similar identification rates have been reported in [1] and [5]. Although the spectral form of /t/ was not diffuse, it was correctly identified; the spectral maximum frequency of /t/ in this context was clearly different of that of /p/ and /k/. /p/ was mostly confused with /k/ because its relatively prominent peak situated in the low frequencies.

Table 2. Consonant confusion matrice for fixed-length stimuli in three known vocalic contexts.

	p(%)	t(%)	k(%)
/pi/	92	1	7
/ti/	0	86	14
/ki/	0	22	78
/pa/	87	8	5
/ta/	8	79	13
/ka/	0	9	90
/pu/	93	3	4
/tu/	0	99	1
/ku/	1	1	98

• **Context /a/.** In most cases /k/ was confused with /t/ because: i) its non compact spectral form, and ii) the frequencies of its maxima were situated in the same region as those of /t/. /t/ was confused with /p/; the missclassified stimuli had a weak energy and their maxima were situated bellow 2500 Hz. /p/ was very well identified; it might be due to its weak energy.

• **Context /i/.** Most of the confusions were between /t/ and /k/. The latter was not always compact, and the frequencies of the maxima of the two stops were situated in the same region (3500 Hz-4500 Hz).

4 COMPARISON BETWEEN THE "VOWEL UNKNOWN" AND "VOWEL KNOWN" EXPERIMENTS

A four way repeated analysis of variance, was conducted in order to estimate the effects of the audition (2 auditions, one per session), of the vowel (3 vowels), of the stop (3 stops) and of the experiment (2 experiments).

The significant main effect of the experiment parameter [$F(1, 336) = 7.05$, p.c. < 0.01] proved that knowing the vowel improved the identification of stop bursts. There were significant effects of the vowel [$F(2,336)=28.83$, p.c. < 0.001] and a significant interaction between the consonant and the vocalic context [$F(4,336) = 14.40$, p.c. < 0.001]. The Scheffé test revealed that /u/ was the best vocalic context, and /a/ the worst one. The interaction between /k/ and /i/, /p/ and /u/, reduced the intelligibility of the consonant, while the interaction between /k/ and /u/ favored it. There was a marginal significant interaction between the experiment and the place of articulation [$F(2,336)=1.90$; p.c.< 0.25]. The Scheffé test indicated an improvement in the identification of /k/ in the "vowel known" experiment: there was no change for /k/ followed by /u/, already very well identified in the "vowel unknown" experiment, whereas the improvement for /k/ followed by /i/ and by /a/ was 6% and 18% respectively. The identification of /p/ was slightly improved (4%), while that of /t/ was reduced by 2%. Another more significant interaction was the triple interaction between the experiment, the stop, and the vowel [$F(4,336) = 2.42$, p.c. < 0.05]. The Scheffé test indicated that the improvement of the identification of /k/ followed by /a/ in the "vowel known" experiment was very significant while the improvement of /p/ followed by /u/ was only slightly significant. Consequently, vowel knowledge was especially beneficial to the identification of /p/ followed by /u/ and /k/ followed by /a/.

5 CONCLUDING REMARKS

The explicit knowledge of the vowel caused a slight but statistically significant improvement of stop identification. More precisely, knowing the vowel was of benefit to /k/ followed by /a/ and to /p/ followed by /u/. The increase appeared without an accompanying decrease of other stops in the same vocalic context, therefore the knowledge of the identity of the vowels /u/ and /a/ had only positive effects. On the other hand, the listeners' performance remained constant in the /i/ context. These deductions allowed us to appreciate how the vocalic information may help stop recognition. For this purpose, we are developing a system that uses vowel features to recognize the stop place of articulation.

ACKNOWLEDGMENT

These experiments were first reported in a Ph. D. thesis [3] supervised by J.-P. Haton, A. Bonneau and Y. Laprie. M. Depaix and F. Lonchamp provided precious advice for the succes of these experiments. The assistance of all these people is gratefully acknowledged.

REFERENCES

[1] Blumstein, S. and Stevens, K. (1979). Acoustic invariance in speech production : Evidence from measurements of the spectral characteristics of stop consonants. *J.A.S.A.*, 66(4):1001–1017.

[2] Bonneau, A., Djezzar, L. and Laprie, Y. (1993). Perception of French stop bursts, implications for stop identification. *Eurospeech*, volume 1, 693–696, Berlin.

[3] Djezzar, L. (1995). Contribution à l'étude acoustique et perceptive des occlusives du français. Ph. D. thesis, U. Henri Poincaré Nancy, France.

[4] Djezzar, L. (1995). Some new considerations about the spectral form of the stop bursts. To appear in *Eurospeech'95*, Madrid.

[5] Kewley-Port, D., Pisoni, D. and Studdert-Kennedy, M. (1983). Perception of static and dynamic acoustic cues to place of articulation. *J.A.S.A.*, 73(5):1779–1793.

[6] Repp, B. and Lin, H.-B. (1980). Acoustic properties ansd perception of stop consonant release transients. J.A.S.A., 85(1):379-396.

[7] Cullinan, W. and Tekieli, M. (1979). Perception of vowel features in temporally segmented noise portions of stop consonant CV syllables. *J. S. H. R.*, 22:103–12.

[8] Winitz, H., Scheib, M. E. and Reeds, J. A. (1972). Identification of stops and vowels for the burst portion of /p,t,k/ isolated from conversational speech. *J.A.S.A.*, 51(4):1309–1317.

THE ROLE OF CONSONANTAL DURATION AND TENSENESS IN THE PERCEPTION OF VOICING DISTINCTIONS OF PORTUGUESE STOPS

João Veloso
Universidade do Porto - Faculdade de Letras (Portugal)

ABSTRACT

Though often described as the main correlate of the distinction voiced *vs* voiceless stops of Portuguese, glottal vibrations (thus, [voiced] feature) seem to be less important than consonantal duration. This study provides data that suggest that the differences of consonantal duration between Portuguese voiced and voiceless stops are highly significant and that the manipulation of this variable significantly changes voicing processing among Portuguese native listeners.

GENERAL PRESENTATION

This paper aims to present and to discuss some results of research into the role of consonantal duration (CDR) and [tense] feature in the voicing oppositions of Portuguese stops.

Six oral stops exist in European Portuguese (henceforth: Portuguese): the voiced /b d g/ and the voiceless /p t k/.

Following Chomsky and Halle [1], the acoustic and phonetic correlate of the voiced/voiceless opposition is the presence/absence of glottal vibration.

Portuguese, however, presents allophonic realizations of /b d g/ without glottal vibration: [b̥ d̥ g̥]. These allophones are processed by native listeners of Portuguese as voiced [2].

There are some studies that show that [b̥ d̥ g̥] are found also in Spanish and that there, too, these allophones are processed as voiced consonants [3, 4].

To a certain point, CDR may explain these perceptual data. Several studies of Portuguese [5, 6, 2] and Spanish [7] have shown that mean CDR is higher in voiceless consonants than in voiced ones.

Since CDR is one of the main acoustic correlates of tenseness [1], some authors [2, 4, 8] have suggested that in Portuguese and Spanish, at the level of distinctive features, the [+tense]/[-tense] opposition may be the fundamental opposition in the separation between voiceless and voiced stops. Thus, in their proposals, the presence/absence of glottal vibrations (i. e., the opposition [+voiced]/[-voiced]) is a redundant, secondary opposition in the organization of the consonantal systems of these languages (these assumptions are clearer among the studies related to Spanish).

It is our aim, in this paper, to go deeper into the importance of these questions in Portuguese.

EXPERIMENTAL PROCEDURE

Corpus

The material for acoustic analysis and the stimuli of the perception tests were extracted from a corpus of spoken Portuguese. This corpus was recorded in an anechoic chamber and was produced by five male adult native speakers of Portuguese, whose dialects were very similar to the "pattern-dialect" of Portuguese; they read one set of sentences with different syntactic structures three times at least.

In all the sentences, sequences with the phonetic structure [aCá] (C=[p t k b d g]) could be isolated and submitted to acoustic analysis.

Acoustic Analysis

In the acoustic study, the CDR of the intervocalic consonant of the above mentioned [aCá] sequences was measured. This measurement corroborated the results of previous studies [5, 6, 2]: Portuguese voiceless stops show mean CDR values that are higher than mean CDR values of voiced ones (voiceless mean CDR > 120 ms; voiced mean CDR < 100 ms - see Table 1). In this study, these differences were evaluated by an Analysis of Variance which showed that they are highly significant (p=0.000).

Table 1. CDR (minimum, maximum and mean values and standard deviations) of each Portuguese stop. Unit: ms

	[p]	[t]	[k]	[b]	[d]	[g]
min	116	105	110	54	43	46
max	147	153	131	108	99	105
mea	132	133	123	80	70	75
SD	13	23	10	17	15	14

Perception Tests

Rationale

Acoustic data from previous studies [5, 6, 2] and our own acoustic study lead us to formulate the following hypothesis: if CDR is an important acoustic cue for the distinction between voiced and voiceless stops, then the manipulation of this variable will interfere in the processing of that distinction.

More precisely, if it is possible to build stimuli from natural Portuguese speech with the phonetic structure [aCá] (C=stop), in which C has the invariant duration of 100 ms, the identification of voicing will be more affected with voiceless stops than with voiced ones. Although mean CDR of voiced stops is below 100 ms, several realizations of [b d g] with CDR values very close to 100 ms were found. In the case of voiceless stops, one single realization (=[t]) was found in this study with a CDR value near 100 ms (=105 ms).

Stimuli

The stimuli of our perception tests consisted of 6 of the [aCá] (C=[p t k b d g]) sequences studied in our acoustic analysis, which form non-words in Portuguese.

All the stimuli were produced by the same speaker. In all of them, C was replaced by a portion of white noise (WN). The spectra of the adjacent vowels and the VC-CV transitions were entirely preserved in this manipulation.

The WN portions did not have the same duration in all the stimuli, which were divided into two sets (A and B):

- *set A:* C was replaced by a portion of WN with the same duration as the replaced consonant;
- *set B:* C was replaced by a portion of WN with the invariant duration of 100 ms.

Subjects

The subjects of the perception tests were 9 non-paid naïve (phonetically untrained) subjects. None of them reported suffering or having suffered from auditory disease. They were divided into two groups:

- *Group I:* 6 native listeners of Portuguese;
- *Group II:* 3 native listeners of German (n=2) and Italian (n=1).

Method

Subjects listened to the stimuli through binaural stereophonic headphones in individual sessions of testing which were divided into two distinct parts. These sessions took place in a quiet room.

Firstly, subjects listened to the stimuli of *set A* (WN=CDR); afterwards, they listened to the stimuli of *set B* (WN=100 ms).

Each stimulus was presented 3 times (6 consonants X 3 presentations = 18 stimuli per session), in a random order. Stimuli were presented spaced by a pause of 3 s.

Subjects were asked to transcribe the intervocalic consonant orthographically on special forms. They were all told that a noise could be heard and that this would not affect the identification of consonants; they were encouraged not to leave blank spaces, i. e., they were told that they should identify all the stimuli.

At the end of each session, the orthographic transcriptions were immediately converted into phonetic transcriptions by the experimenter, who asked the subjects for explanations whenever he had any doubts about their transcriptions.

RESULTS

Table 2 displays the results of the perceptual tests. The analysis of answers considered only voicing (i. e., if a subject identified the place or the manner of articulation of a consonant wrongly, but voicing was correctly identified, his/her answer was taken as correct).

Table 2. Percentage of correct identifications of voicing with voiceless and voiced stops by both groups of subjects and in both sets of stimuli

	Voice-less		Voic-ed	
	Native	Non Native	Native	Non Native
WN= CDR	62.9 %	70.4 %	94.4 %	100 %
WN= 100 ms	22.7 %	18.5 %	83.5 %	92.6 %

The differences of voicing processing between the two sets of stimuli are significant ($p<0.05$) only in the native listeners' group with voiceless stops. On the other hand, the manipulation of CDR did not significantly ($p\geq0.05$) alter the voicing processing in either the non-native listeners' group with voiceless and voiced stops, or the native listeners' with voiced stops (the values of p here stated were obtained from the *t* statistics).

GENERAL DISCUSSION AND CONCLUSIONS

The differences that we found and their significancy levels lead us to accept our initial hypothesis: CDR is an important acoustic cue for the processing of the distinction between voiced/voiceless stops, at least for native listeners of Portuguese.

If we consider only the native listeners' group, the differences of voicing processing were significant only with voiceless stops because of the manipulated values of CDR. In the set B of stimuli, the value of WN (=100 ms) is clearly below the minimum and mean values found in the set of voiceless stops. In voiced stops, this invariant CDR of 100 ms is higher than their mean CDR, although several realizations of /b d g/ with CDR values not very far from 100 ms were found.

Our results support the proposals of previous studies of Portuguese [2] and also of Spanish [3, 4, 8] which claim that in these languages [tense] feature is a very steady correlate of the opposition voiced/voiceless among stops: the present study shows that, in Portuguese, voiced and voiceless stops have significantly different mean CDR values - which are among the main acoustic correlates of tenseness - and that these acoustic differences are perceptually important.

This importance of CDR for the voicing processing seems to be more important in some languages than in others, as is shown by the different results of perceptual tests with listeners from different languages.

ACKNOWLEDGEMENTS

I am grateful to Prof. Maria da Graça Pinto (University of Porto) for all her advice and for her priceless comments on earlier drafts of this paper.

Part of the experimental work of this study was carried out at Stockholm University (Dept. of Linguistics, Phonetics Laboratory), thanks to a scholarship I was granted by the Swedish Institute in 1992. I thank Dr Francisco Lacerda (Stockholm University) for all his support during my stay at his institution.

I am also grateful to Dr Belinda Maia (University of Porto) and to Eleanor Underwood (University of Trás-os-Montes e Alto Douro), for having helped me with the English version of this paper.

REFERENCES

[1] Chomsky, N., & Halle, M. (1968), *The Sound Pattern of English,* New York: Harper & Row.

[2] Viana, M. do C. (1984), *Etude de Deux Aspects du Consonantisme du Portugais: Fricatisation et Dévoisement.* Ph. D. diss., Université des Sciences Humaines de Strasbourg.

[3] Alarcos Llorach, E. (1950), *Fonología Española,* Madrid, Gredos.

[4] Veiga, A. (1985), "Consideraciones relativas a la actuación y límites de las oposiciones fonológicas interrupto/continuo y tenso/flojo en español", *Verba - Anuario Galego de Filoloxía,* vol. 12, pp. 253-285.

[5] Delgado Martins, M. R. (1975), "Vogais e Consoantes do Português: Estatística de Ocorrência, Duração e Intensidade", *Boletim de Filologia,* tomo XXIV (Fascs. 1-4), pp. 1-11.

[6] Viana, M. do C. (1979), "O índice duração e a análise acústica das oclusivas orais em português", *Boletim de Filologia,* tomo XXV, pp. 1-20.

[7] Martínez Celdrán, E. (1984a), "Cantidad e intensidad en los sonidos obstruyentes del castellano: hacia una caracterización acústica de los sonidos aproximantes", *Estudios de Fonética Experimental,* vol. I, pp. 73-129.

[8] Martínez Celdrán, E. (1984b), "¿Hasta qué punto es importante la sonoridad en la discriminación auditiva de las obstruyentes mates del castellano?", *Estudios de Fonética Experimental,* vol. I, pp. 245-291.

DISCRIMINATION OF COARTICULATED GERMAN VOWELS IN THE SILENT-CENTER PARADIGM: "TARGET" SPECTRAL INFORMATION NOT NEEDED

Ocke-Schwen Bohn and Winifred Strange***
**English Department, Kiel University, Germany*
*** Dept. Communication Sciences and Disorders, U. South Florida, Tampa, USA*

ABSTRACT

German listeners were tested on their ability to discriminate naturally produced German /dVt/-syllables which were modified to manipulate the availability of formant target information (traditionally considered the primary information for vowel identity) and of dynamic spectral information. The results support Strange's [1] Dynamic Specification Theory, which states that vowels are specified by dynamic information defined over syllable onsets and offsets.

INTRODUCTION

In the traditional view of vowel perception, the target frequencies of the first two formants constitute the primary acoustic information for the perceptual identity of vowels. This Simple Target Model of vowel perception is inadequate because it fails to account for how listeners perceive speakers' intended messages in the face of various sources of variation in the acoustic signal. One kind of variation in vowel targets comes from coarticulation of vowels with consonants in consonant-vowel-consonant (CVC) syllables. Research by Strange and her collaborators has shown that vowels produced in CVC-syllables are identified with far greater accuracy than vowels produced in isolation, even though targets are often not reached in coarticulated vowels [1]. This led Strange to hypothesize that important information for vowel identity must be contained in the dynamic contour of the formants within the syllable.

Strange and her collaborators developed the Silent Center paradigm to test their hypotheses on the role of dynamic sources of information in vowel perception. Their methodology involves the systematic modification of CVC-syllables to explore the perceptual relevance of various sources of potential information contained in CVC-syllables. The methodologically most important modification in this paradigm is the generation of silent-center (SC) syllables, which are created by attenuating to silence the entire syllable nucleus, leaving only the initial and final transitions in their appropriate temporal relationship. The converse of SCs are vowel-centers (VCs), which are created by deleting the initial and final transitions, so that the syllable nuclei with target information are retained. Experiments employing SCs and VCs as stimuli allow one to test the perceptual importance of acoustic information associated either with the opening and closing gestures of the vocal tract in the production of CVC-syllables (i.e., SCs), or with the (approximate) target configuration of the vocal tract in vowel production (i.e., VCs). Experiments employing the SC paradigm typically also test the perception of initials (INIs) and finals (FINs), which are, respectively, the initial and final transitions alone. This is done to test whether SCs in their entirety, or their initial or final part alone, contribute to perceived vowel identity.

Several studies have examined the perception of AE vowels by AE listeners in the SC paradigm and found high levels of identifiability for SCs. Strange [1] concluded a review of these studies by stating that no single spectral cross-section adequately captures the perceptually relevant information; rather, the acoustic information for vowel identity resides in the changing spectral structure. Because most previous SC studies examined the perception of AE vowels, some ambiguity as to the nature of the dynamic information remained. According to Nearey's Compound Target Theory (CTT), the more or less diphthongized English vowels can be differentiated by contrasting patterns of vowel-inherent spectral change [2, 3]). Strange's Dynamic Specification Theory (DST), on the other hand, states that vowels are specified by dynamic information defined over syllable onsets and offsets [1]. The dynamic information

reflects each vowels' characteristic opening and closing phases in their appropriate temporal relationship and style of movement of the vocal tract.

In order to examine the generality of previous findings on the identification of AE vowels in the SC paradigm, and to assess alternative hypotheses about the nature of dynamic information, Strange & Bohn [4] examined the identification of coarticulated German vowels in /dVt/-syllables by native German listeners in the SC paradigm. Like English, German has a large vowel inventory whose monophthongs differ in tenseness and/or length, but which have little or no diphthongization. Strange's DST predicts that dynamic spectral information plays an important role in the perception of vowels which have little or no diphthongization. Nearey's CTT, on the other hand, predicts that the importance of dynamic spectral information is restricted to the perception of diphthongized vowels.

Strange & Bohn [4] reported that the availability of dynamic spectral vs. target spectral information affected German listeners' identification of coarticulated German vowels in much the same way as AE listeners' identification of coarticulated AE vowels. Vowel identity was maintained very well in German SC-syllables even though the vocalic nucleus with information on formant targets was not presented in that condition. Strange & Bohn also reported extremely high error rates in both INI and FIN conditions, indicating that neither onsets alone nor offstes alone were sufficient to maintain vowel identity for German listeners.

The present study was designed to test the robustness of the findings of Strange & Bohn by using a new set of stimuli and employing a different experimental paradigm. Four vowel contrasts which German listeners confused most frequently in Strange & Bohn's [4] identification experiments were examined for their discriminability in experimental conditions which selectively presented different types of acoustic information contained in CVC-syllables. Compared to previous identification studies, the discrimination paradigm used in the present study presents listeners with a very simple task in terms of memory load and cognitive demands. One might expect, therefore, that performance levels in INI and FIN conditions will be much higher in a task involving simple discrimination rather than identification with, e.g., 14 response alternatives as in the Strange & Bohn study. If this were so, an important assumption of the DST would be questioned, namely, that vowel identity is specified not by onsets alone or offsets alone, but by trajectory information defined over syllables onsets **and** offsets in their appropriate temporal relationship. On the other hand, the DST would be strengthened if the pattern of results obtained from a simple discrimination experiment were much the same as the patterns obtained previously from identification experiments (i.e., low performance levels for INIs and FINs as opposed to the SC condition).

METHODS

Stimuli

Six tokens each of the vowels /i/, /I/, /e/, /ɛ/, /U/, /o/ were produced in isolated /dVt/-syllables by a male native speaker and recorded onto DAT. The vowels were selected for discrimination because the pairs /i/-/e/, /e/-/I/, /I/-/ɛ/ and /o/-/U/ were most frequently confused in the identification experiments of Strange & Bohn [4]. From the digitized waveforms, measurements of target syllable duration, voice onset time and fundamental frequency were used to make the final selection of four instances each of the six vowels.

SCs were generated by attenuating to silence the center portion of each of the target syllables, leaving onset and offset portions in their original temporal position. The onset and offset portions included the major part of the transitions associated with opening and closing gestures for all vowels. - VCs were the converse of SCs. They were generated by attenuating to silence the onset and offset portions. INIs were generated by silencing both center and offset portions in each syllable. FINs were generated by silencing both onset and center portions of each syllable.

Subjects

50 subjects who met the selection criteria (no history of hearing loss according to self-report, native speaker of North German, limited exposure to

languages other than German) were recruited mainly from linguistics courses at Kiel University and participated as unpaid volunteers. The mean age of the 33 female and 17 male subjects was 26 years (SD = 5.5).

Procedure

The stimulus syllables (unmodified syllables, SCs, VCs, INIs, and FINs) were redigitized and stored on the hard disk of a 486-PC. Groups of 10 subjects each were assigned to one of the five listening conditions (defined by stimulus type) and tested individually in a sound treated chamber, where stimuli were presented from a loudspeaker. Each subject was tested for discrimination of the contrasts /i/-/e/, /e/-/I/, /I/-/ɛ/, /o/-/U/ in a pseudo-randomized order. Because the results of the discrimination experiment were to provide baseline data for a study of infant vowel perception that employed the change/no change procedure [5], the adult subjects were tested in an age-appropriate version of that procedure. The subjects listened to presentations of the background stimuli (e.g., the four randomly presented tokens of /dit/ for the /i/-/e/ contrast) while being engaged in a simple distractor game, and were instructed to raise their hand when they detected a change to the foreground (e.g., /det/ for the /i/-/e/ contrast). The ISI was 1.5 sec; a change trial consisted of the presentation of three foreground stimuli. The change from background to foreground stimuli was initiated by an assistant, who could not hear whether she initiated a change or a control trial, and who observed the subject through a one-way mirror. The assistant pressed a button when she observed a hand signal by the subject during trials. Custom software controlled feedback for correct signals (illumination of a display above the loudspeaker.) The software also kept track of correct responses, false alarms, correct rejections, and misses throughout the 25 trials (15 changes and 10 controls) for each contrast.

RESULTS

Figure 1 gives the overall results for the five stimulus conditions for the vowel contrasts /i/-/e/, /e/-/I/, /I/-/ɛ/, /o/-/U/, expressed as percentages of correct responses averaged across subjects within each group. Discrimination levels for unmodified syllables (mean % correct: 97.5), SCs (mean % correct: 96.1), and VCs (mean % correct: 99.1) did not differ significantly. Vowel identity was well maintained in the SC condition, even though the vocalic nucleus with information on formant targets was not presented in that condition. All four vowel contrasts were discriminated highly accurately in the SC condition (/i/-/e/: 96.5%, /e/-/I/: 95.6 %, /I/-/ɛ/: 95.6, /o/-/U/: 96.8 %).

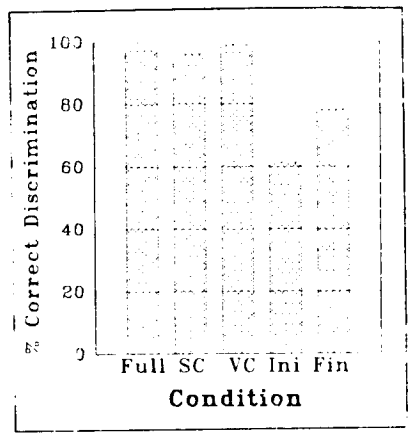


Figure 1: Overall per cent correct responses in vowel discrimination in unmodified syllables (Full), silent-center syllables (SC), vowel center (VC), initials (INI), and finals (FIN) conditions.

Both FINs and INIs on the one hand were discriminated siginificantly less accurately than unmodified syllables, SCs, and VCs on the other. This shows that neither onsets alone nor offsets alone were sufficient to maintain vowel identity. Overall discrimination of FINs (78.3 % correct) was significantly better than of INIs (61.2 % correct), but more detailed analyses revealed that this was true only for contrasts differing in tenseness (/e/-/I/, /o/-/U/). Performance on FINs and INIs was not significantly different for the tense /i/-/e/ or the lax /I/-/ɛ/ contrast.

CONCLUSIONS

The most important finding was that German listeners discriminated German vowels highly accurately when only dynamic spectral information specified over onsets and offsets together was presented. The pattern of results for the discrimination of confusable German vowel contrasts was very similar to that reported previously for the identification of German and AE vowels by native listeners. For an adequate perceptual representation of vowels in these two languages with large vowel inventories, target spectral information is not necessary; rather, trajectory information specified over syllable onsets and offsets is a very good source of information for vowel identity in both German in AE.

The finding that INIs and FINs were discriminated less accurately than SCs provides important support for Strange's DST [1], which states that vowel identity is specified by dynamic information defined over syllable onsets **and** offsets. Even when listeners were confronted with the very simple task of detecting a change from foreground to background stimuli, onsets alone (INIs) or offsets alone (FINs) did not specify vowel identity as well as trajectory information defined over syllable onsets **and** offsets in their appropriate temporal relationships (SCs).

The finding that German vowels presented as SCs were identified and discriminated highly accurately is incompatible with Nearey's ([2, 3]) Compound Target Theory (CTT). The CTT predicts that distinct patterns of vowel-inherent spectral change (VISC) contribute importantly to perceived vowel identity, at least when the more or less diphthongized vowels of AE are presented as SCs. However, acoustic analyses of the German vowels presented in the experiments reported here revealed very little or no VISC. Formant movement, particularly for F2, was observed in many of the vowels, but this movement was associated with coarticulatory influences from the preceding and following alveolar consonants (see also [6]). This means that the dynamic sources of information which German listeners used so successfully in the SC conditions were associated with the opening and closing gestures at the margins of CVC-syllables, as predicted by Strange's [1] DST.

Further studies are underway to establish the generality of these perceptual results with German vowels and adult German listeners. One series of studies in progress examines whether German vowels produced in multiple consonant contexts by multiple speakers are identified accurately when listeners are presented with SCs. Another series examines how the three types of acoustic information present in coarticulated German vowels (dynamic spectral, target spectral, and temporal) contribute to perceived vowel identity in prelingual German infants [5]. Preliminary results from these experiments suggest that almost all infants who can discriminate vowel contrasts with unmodified syllables can also discriminate the same contrast when vowels are presented as SCs.

ACKNOWLEDGEMENTS

Research supported by a grant from the *Deutsche Forschungsgemeinschaft* (DFG grant Bo-1055/3-1) to O. Bohn, and by a grant from the National Institutes of Health (NIDCD 00323) to W. Strange. The authors wish to thank Desiderio Saludes and Sonja Trent for their assistance in generating stimulus materials, and Kirsten Schriever and Anja Steinlen for assistance in testing subjects. Very special thanks go to Linda Polka for her help in setting up the speech perception lab at Kiel University.

REFERENCES

[1] Strange, W. (1989), "Evolving theories of vowel perception." *J. Acoustical Soc. America* 89, 2081-2087.

[2] Nearey, T. (1989), "Static, dynamic and relational properties in vowel perception." *J. Acoustical Soc. America* 85, 2088-2113.

[3] Andruski, J. E. & Nearey, T. M. (1992), "On the sufficiency of compound target specification of isolated vowels and vowels in /bVb/ syllables." *J. Acoustical Soc. America* 91, 390-410.

[4] Strange, W. & Bohn, O.-S. (1995), "Dynamic specification of coarticulated German vowels: I. Perceptual studies." To appear in *J. Acoustical Soc. America*.

[5] Bohn, O.-S. & Polka, L. (1995), "What defines vowel identity in prelingual infants?" *Proceedings 13th International Congress of Phonetics*.

[6] Strange, W. & Bohn, O.-S. (in prep.), "Dynamic specification of coarticulated German vowels: II. Acoustical studies."

PROSODIC STRUCTURES AND DISCOURSE ORGANIZATION

Li-chiung Yang

Department of Linguistics
Georgetown University, Washington, D.C. U.S.A. 20057

ABSTRACT

In this paper we examine how topic organization and discourse interaction are manifested in the intonational structure of a discourse. Topics are hierarchically organized to indicate the topic relationships between phrases. Patterns of intonational convergence and divergence in intonation arise naturally from the cooperative working out of discourse processes.

Introduction

In spontaneous discourse, intonational patterns of phrases are a critical component in conveying the coherence of topic development and the salience of communicated information between participants. Topic structure and discourse interaction are integral and inseparable parts of discourse organization, as participants cooperate and interact to achieve a mutual development of the conversation. The pitch level structuring of each phrase relative to other phrases through systematic patterns of downstepping and upstepping signals the coherence relations among phrases.

Downstepping and topic development

Examination of the pitch movements in the dialogue shows that discourse is hierarchically organized by intonation. In my data, dialogue, episodes and topic initiations often begin with a high pitch level or expanded pitch range, while endings are marked by a low pitch level or narrowed pitch range. The direction of pitch level movement is frequently downward.

Downstepping between phrases usually occurs when there is a natural elaboration of topic ideas which move towards a resolution. This process can be seen as progressive movement away from uncertainty, with each subsequent phrase closer to a final resolution. The degree of step lowering represents the degree of completeness and finality of the phrase relation in the topic hierarchy. Figure 1 presents a plot of the peak pitch points for a continuous subsection of 90 utterances for two speakers.

(1)
21 B: Oh wasn't ours around four
fifty or so?
22 then you have to add tax,
23 then altogether it was close to five
hundred.
24 B: Umhum.

In example (1), speaker B's discourse goal is to explain the price of an object, and this goal is resolved by the successful computation of a logical sequence of calculations. This sequence of utterances in U21 to U24 is triggered by the high-pitched doubt and self-reflection expressed by the speaker in U21. In each successive phrase, the speaker explicitly works out the calculations, and the growing certainty brings her progressively closer to a confident conclusion at a low pitch point at U23, and a closing discourse particle *umhum* (U24) to self-confirm at the lowest pitch point. The high pitch level of U21 followed by a large downstep of about 80Hz reflects the emotion associated with the high level of uncertainty at U21 and the subsequent shift towards a resolution in U22.

Regular stepping in more uniform development

The elaboration of topic in discourse often takes on a more settled and structured character. This frequently occurs when a speaker has a more extended series of ideas to present as the discourse enters a more narrative phase. In these situations, the further

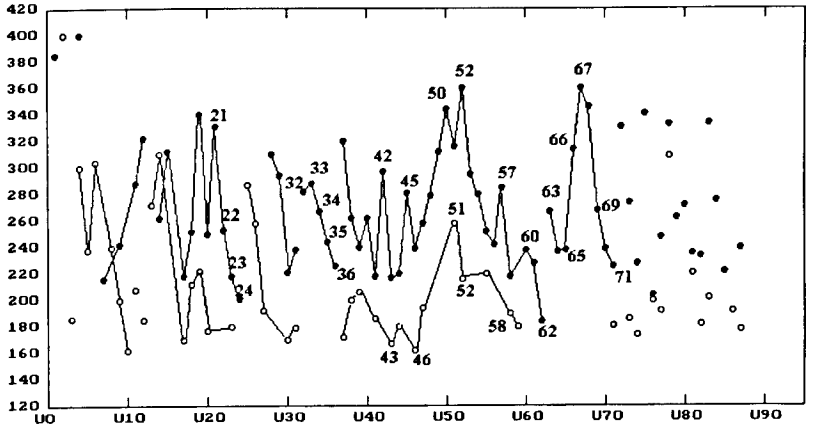


Figure 1. Plot of Pitch Peaks of 90 Utterances
** Speaker A: ○ Speaker B: •*

development and elaboration of topic is often associated with a systematic step structure between phrases which signals the topic organizational hierarchy and indicates that one phrase logically follows the other. In the case of downstepping, proportional or regular steps often occur as the speaker moves systematically towards the completion and resolution of a specific discourse goal. The size of the step is correlated with the degree of emotional intensity or degree of cognitive transformation.

(2)
32 B: Then - I once before
33 used the facilities of another
school -
34 (f) At that time I didn't know
how they did it,
35 (f) anyway everything there was
computerized,
36 (f) they did it for us.

In example (2), speaker B first introduces the topic, and then adds on successive qualifying expressions in U34 to U36; each qualifying expression drops by about 15 to 20hz. The smaller and uniform step sizes reflect the relative stability and constancy of the speaker's state during this short segment.

Upstepping as cognitive uncertainty

In narrative-like speech, new topics or subtopics are often introduced as more gradual and natural developments of previous topics, and topic initiation phrases often start at a more intermediate or low level and subsequent phrases develop in patterns of climax and resolution. This may occur because the unity of the narrative development may take precedence over the need to signal new information. Successive upstepping from such low topic initiation points often occurs as emotional elements unexpectedly enter the conversation.

(3)
63 B: Then we just
asked people to come in to record
64 Then after the recording
65 then we just took that tape -
66 Oh, let me think
what happened then?
67 In any case, in any case ---
68 I forget what happened,
anyway it's just
70 that sound, huh,
69 they just took that tape,
71 and put it into the computer.

Language often reveals the cognitive process. In U63 to U65, speaker B starts out confidently and develops the topic further in U64, but is already hesitant, and her hesitation continues in U65. In U66-U67, speaker B is in a state of cognitive uncertainty, and steps out of

the current topic line to attempt to recall information. The uncertainty of topic development is reflected in the upstepping of these phrases. The speaker finally reaches a cognitive turning point which is indicated by the phrase "in any case, in any case" at U67 to signal a return to the main topic after a digression. Having made this decision, the speaker then returns to a more certain and confident cognitive state, and continues to develop the next logical step. This results in downstepping in U68-U71.

Upstepping between phrases often occurs in situations of cognitive uncertainty, as in self-reflection and doubt, in contrast to downstepping, which is typically associated with definiteness, finality and completion. Upstepping often continues until a climax occurs at a high cognitive turning point, which in this example occurs when speaker B is unable to retrieve the missing information, and decides to go on. The cognitive uncertainty is then resolved and expressed in the subsequent phrases.

Topic, planning and cognitive states

The complexity of topic structure increases as discourse gets more involved and more complicated. In extended discourse, topics and subtopics flow into one another, therefore it is often difficult to pinpoint where the boundaries of a topic are. Most importantly, the progression of natural discourse differs from strict semantic-logical discourse progression. Instead of following exact semantic implications or inferences, as often seen in formal logical models, the reasoning used in natural discourse works according to the principle of *relevance*. In spontaneous conversation, ideas are often triggered in a chain reaction fashion, and topics develop by building on preceding phrases. In addition, speakers often backtrack or step out from the ongoing topic to monitor the communicative process and to provide appropriate background information. Topics are organized as psychologically logical and relevant to the speaker and hearer, and are initiated and motivated by this principle.

The complex hierarchical and systematic nature of topic intonation can be seen in the section from U43 to U62 (Due to space limitations, the discourse text is omitted). In this section, the speaker is presenting two accounts of a story. The hierarchical intonational structure appears as a sequence of generally upstepping phrases in the first account in U43 to U51, followed by a downstepping pattern in the retelling of the story in U53 to U62.

In the beginning part of this segment, the speaker is building up the topic, and at the same time is trying to recall the specific details and to accommodate and make the conversation relevant to the information status of the other participant. This results in frequent impromptu insertions, add-ons of new ideas, first attempts at making a point, and side developments. In this section, each phrase stems from some feeling of inadequacy about the incompleteness of the previous phrase, or some feeling of uncertainty arising from not having a definite idea of the topic direction. In the meantime, the speaker is working up to a climax point. Each phrase functions to add new information as a way to overcome the previous phrase, therefore each step is higher than the other, until the speaker finally comes to the climax - a high point - in the story. The upstepping pattern is just such an expression and representation of these elements in the speaker's mind. In the subsequent section, the speaker has already organized the essential points in her mind, has also established the appropriate common background, so is free of disruptions, and can concentrate more on the topic structure, i.e. the elements of the story, itself. Therefore her account here is smoother and more certain, as signalled by the gradually descending anti-climactic downslope. The essential point is that the *planning* in these two versions differs: the upstepping section, the 1st version, is less planned, whereas the downstepping section, the 2nd version, is more planned. This example is evidence that discourse is *not* always pre-planned, and the intonational structure and topic development reflect the degree of planning involved.

This evidence for the use of pitch level to signal topic structure is enhanced by B's three upward spurts within that gradual rise-fall slopes. These three spurts do not follow the general trend but have a temporary rise in pitch level above that trend. In each case, in U45, 52 and 57, the upward spurt occurs because the phrase is a summarization of the previous statements, and is a signal that the speaker has temporarily interrupted the logical stream.

Topic, convergence and divergence

My data also show that discourse is interactionally and cooperatively organized. In discourse, participants accommodate and interact frequently to mutually work out development and resolution of the conversation. The cooperative nature of discourse is manifested in a process of intonational convergence and divergence among participants. Intonational convergence occurs when participants match corresponding movements in pitch ranges of utterance to indicate support of topic hierarchy and development, and to signal enthusiasm and sympathetic agreement with the other speaker's point of view.

The intonational interactions between participants in the segment U41 to U62 show how patterns of convergence and divergence can vary according to the signals of topic organization which the main speaker provides. In the regularly structured narrative, speaker B starts at a low point and then accelerates until the climax is reached in U50. The hearer's pattern is remarkably similar to the main speaker's. When speaker B starts low at U41, speaker A also responds low at U43. As speaker B rises to reach the climax at U50, speaker A also responds with heightened pitch and rises to her highest point at U51. On the anti-climatic downslope, speaker B gradually returns to a state of equilibrium in a series of downsteps, and speaker A's feedback also follows a downstepping pattern and ends low at U59.

On the whole, speaker A's pitch movements are mirroring exactly the pitch movements of speaker B's, and the two speakers are moving together in an emotionally synchronized manner. At phrases in which speaker B continues the further elaboration of topic, the two speakers' step movements are parallel, i.e. they converge, and this convergence signals that both speakers are in agreement on the topic development.

This overall pattern of convergence is interspersed with isolated instances of divergence where both speakers move in opposite directions. In U42, U45, U52, and U57, when speaker B steps out of the current topic flow to summarize or make a point, as signalled by the abrupt pitch rise, speaker A actually drops her pitch, at U43, U46, U52 and U58, to indicate her acknowledgement and understanding that the phrase is outside the regular topic flow.

The above examples illustrate that discourse is both hierarchically and interactionally organized and that intonation signals systematically the organization of topic and discourse activity. Results of this study demonstrate that progress in the understanding of prosody is gained if adequate consideration is given to the varied functions of intonation in discourse. Topic, discourse context, and discourse interactions are fundamental to the extended meaning which intonation uniquely provides. A clear understanding of the determinants of hierarchical structuring and discourse intonational development is of crucial importance in the further study of natural discourse.

ACKNOWLEDGEMENT

This research was conducted at the Phonetics Lab of Stanford University. I am grateful to Bill Poser for his generous encouragement, to Tom Veatch for help with WAVES, and to Stanford Linguistics Department for supporting this research. Thanks also to Shaligram Shukla, Charles Kreidler, Peter Patrick, and Lisa Zsiga for their encouragement.

MODELLING INTONATION IN DIALOGUE

G. Ayers, G. Bruce**, B. Granström***, K. Gustafson***, M. Horne**, D. House**, and P. Touati ** (Names in alphabetic order)*
**Dept. of Linguistics , Ohio State University, 222 Oxley Hall, 1712 Neil Avenue , Columbus, OH 43210-1298, U.S.A.*
***Dept. of Linguistics and Phonetics, Helgonabacken 12, S-22362 Lund, Sweden*
****Dept. of Speech Communication and Music Acoustics, KTH, Box 70014, S-100 44, Stockholm, Sweden*

ABSTRACT

The analysis of spontaneous dialogue in Swedish is discussed. The methodology, speech material and analysis types are presented, as well as the intonational aspects under study. In particular, tonal downtrend is examined in relation to lexical semantic aspects of topic structure, and the environments in which downstepping occurs are outlined.

BACKGROUND

The framework for our present research on prosody is the project Prosodic Segmentation and Structuring of Dialogue (proZodiag) centered around the description of Swedish. The project is supported within the second phase of the Swedish Language Technology Programme (1993-1996) and represents cooperation between Phonetics at Lund and Speech Communication at KTH, Stockholm [1].

The object of study is the prosody of spontaneous speech and dialogue. The main goals of this research are to increase our understanding of prosody in its natural environment - dialogue and spontaneous speech - and eventually to create a more powerful prosody model.

The background for our present research effort is experience from two decades of study of the prosody of prepared speech in a laboratory setting. The intention of using such laboratory speech has not been to study the prosody of reading, but rather to simulate natural, spontaneous speech. Laboratory speech provides us with a high degree of experimental control but presents an artificial situation with often pragmatically anomalous speech material where the informant's skill in acting becomes important in the recording of the data. For a discussion see [2, 3].

A reasonable question to ask in this context is then why up until fairly recently have there been so few studies on the prosody of spontaneous speech. The apparent reason for this state of affairs is the well testified complexity of prosody; the control of variables is much easier in the phonetics laboratory.

METHODOLOGY

We are exploiting a number of different kinds of speech material: true spontaneous dialogue (recorded with no intention of being studied phonetically), spontaneous 'lab speech' dialogue (with partly predetermined conversational topics but otherwise spontaneous), acted dialogue from scripts (using the 'lab speech' dialogues), dialogue simulation in text-to-speech synthesis, and man – machine dialogue.

For our study of dialogue prosody and spontaneous speech we are using the following types of analysis: analysis of the dialogue structure itself (without taking prosodic information into account), and a two part prosodic analysis: an auditory analysis in the prosodic transcription stage and an acoustic-phonetic analysis which takes F0 and waveform information into consideration.

We conceive of the analysis of the dialogue structure as comprising at least the following aspects: textual aspects (both the structure of conversational topics and its relation to lexical-semantic aspects of focus), turn regulating aspects (keeping, yielding, taking the turn), initiative / response structure, feedback (both giving and seeking feedback), and rhetoric activity (which appears to occur in all kinds of speech, to varying degrees).

In our prosodic analysis of Swedish, focus is on intonational aspects. Our auditory analysis results in a prosodic transcription involving two levels of prominence (accented, focussed), tonal junctures and two levels of grouping (minor, major phrase). The choice of word accent (accent I / accent II) in Swedish is lexically determined. The prosodic categories used are the following:

Tonal Structure		
accented	accent I (HL*)	
	accent II (H*L)	
focussed	accent I ([H]L*H)	
	accent II (H*LH)	
	compound (H*L...L*H)	
juncture	initial (%L; %H)	
	terminal (L%; LH%)	
Grouping		
boundary	minor	|
	major	‖

The acoustic-phonetic analysis is based on F0 and waveform information, whereby both global features such as, for example, F0 level and F0 range and local features such as direction and timing of F0 events are taken into consideration and interpreted in our current prosody model.

The three types of analysis – analysis of dialogue structure, auditory analysis, acoustic-phonetic analysis – involving both symbol and signal information are combined and synchronized with each other in the same ESPS/Waves+ environment. The labelling used (symbol information) is similar to the ToBI transcription system for English [4]. It consists of an orthographic tier (marking the end of words), a tonal tier (symbols of tonal structure), a boundary tier (symbols of grouping), dialogue structure tier (hierarchy of conversational topics), and a miscellaneous tier (with extralinguistic and other information).

An important part of our research methodology is the use of speech synthesis. We are using two different approaches in our research. The first method is a synthesis matching technique to verify the prosodic transcription. Our model is implemented in the ESPS/Waves+ environment, and the input is the prosodic transcription with information about type and time location of tonal turning points. This information (with little segmentation indicated) together with phonetic rules from our prosody model are fed into a modified version of the ESPS/Waves+ synthesizer. The model contour is synthesized and compared to the original. Deviations are then studied, which leads to improvements in the model.

The other approach is to use the KTH text-to-speech synthesis system. Using an experimental version of this system which includes an extended set of prosodic markers, we have a flexible tool for manipulating prosodic parameters. It is particularly suited for testing our hypotheses about prosody on new speech material, specifically the simulation of dialogues.

TONAL DOWNTRENDS

One aspect of our modelling of dialogue intonation involves analysis of intonational downtrends. See for example [5, 6]. According to our earlier study of Standard Swedish in controlled laboratory experiments, the occurrence of a focal accent is a pivot for tonal downstepping [7]. Before a focal accent, non-focal accents do not appear downstepped, while after focus downstepping of successive accents within the same phrase / utterance is a characteristic feature. Thus an early focus in an utterance will typically trigger downstepping, while a late focus will tend to arrest this tonal downtrend. That is to say, downstepped accents occur on information that is given in the context as in (1b) which constitutes an answer to the question in (1a). The underlined words in (1b), which are contextually given in the preceding question, are characterized by downstepped accents. Words in bold print are focussed:

(1) a. Vem lämnar ungen nallar? 'Who gives the kid teddy bears?'
b. **Mamman** <u>lämnar ungen nallar</u> 'The **mamma** <u>gives the kid teddy bears</u>'

In our earlier study of spontaneous speech within the 'KIPROS' project, this regularity was found to appear in spontaneous dialogue as well [8, 9]. Several, typical examples of downstepped pitch patterns were observed which seemed to be triggered by the placement of focal accent in the same way as described above.

In our current analysis of spontaneous speech, downstep has also been seen to correlate with contextually given information. However, it has also been observed to correlate with information that cannot be classified as given. The following examples are taken from our analysis of a dialogue recorded from a Swedish radio program about jazz music called 'What's cooking?'.

In (2), the clause after the word *macka* 'sandwich' does not contain given information in the same sense as in example (1). Yet, it is characterized by downstep in the same way as the underlined words in (1).

(2) Jag har ett litet recept på en varm **macka** | som jag faktiskt har har utvecklats utvecklat en aning
'I've got a recipe for a hot **sandwich** | which I in fact have have improved improved a little'

In this example and in the ones that follow, the underlined words are characterized by downstepped accents, and words in bold print are focussed. Phrase boundaries are also indicated.

It should also be noted that in the examples there is no focal accent present in the underlined phrase itself as a direct trigger of the downstepping. Instead, each successive non-focal accent within the phrase is downstepped.

Thus, the generalization that it is given information that gets downstepped is perhaps too narrow to cover all cases of the phenomenon in spontaneous speech. It would, however, be insightful if one could relate examples like those in (1) and (2) to some more general discourse/semantic parameter(s).

One idea which we would like to pursue in this respect is to relate downstepped information to the development of discourse topics. In this regard, one could say that the downstepped information in (2) is similar to the nonfocal material in (1), in that it can be considered as information which is not central to the development of the topic. By 'central to the development of the topic', we then mean related to the specification of the lexically important/'generic' (see below) referents in the semantic field under discussion as well as the specification of the relationships among these referents. In the specific dialogue under consideration, the central topic involves the description of the ingredients in a recipe for a hot tuna fish sandwich. Just as the downstepped information in (1), being contextually given, does not provide anything new as regards the relationships between the referents *ungen* 'the kid' and *nallar* 'teddy bears', the downstepped information in (2) likewise does not lead to the development of the central topic in the discourse from which it is extracted, i.e. to the description of the discourse referent *macka* 'sandwich'. In other words, there is no information regarding the referents that are relevant to the description of the sandwich. The downstepped material constitutes parenthetical information which is unrelated to the specification of the ingredients in the sandwich, i.e. is non-central to the development of the topic.

Another example of the use of downstep is given in (3):

(3) ...man har vitt **bröd** förslagsvis | och på detta lägger man en röra av.....
'...you take white **bread** for example | and on that you put a mishmash of...'

In this case, noncentral can be related to the level of specificity of the discourse referents. The word *röra* 'mishmash' is semantically nonspecific or 'non-generic' as regards its status in terms of lexical hierarchies. That is to say, it is not "at the level of ordinary everyday names for things and creatures" [10] as regards its relation to the other referents which are mentioned in the dialogue, e.g. *bröd* and the ingredients that make up the 'mishmash': *tonfisk* 'tuna fish', *majonnäs* 'mayonnaise', etc. Thus, one can hypothesize that the downstepping in the utterance in (3) is related to the nonspecificity/nongenericness of the referent mentioned. In other words, one can hypothesize that referents that are central to the topic are ones that are relatively more 'generic' or more 'basic'.

A similar situation can also lead to prosodic downtoning, i.e. when the speaker 'comments on' /specifies an already introduced 'generic' discourse referent as in (4-5):

(4) ...man måste ha en burk **tonfisk**, en burk **crème fraiche** | tonfisk med vatten bara, eh, är bra...
'...you have to have a can of **tuna fish**, a package of **crème fraiche** | just tuna fish in water, uh, is good...'

(5)...en tredjedels burk **majonnäs** || gärna lätt majonnäs där också för att crème fraiche i...
'...a third of a jar of **mayonnaise** || preferably light mayonnaise there too since (there's) crème fraiche in (it)...'

Here the downstepping characterizing the second occurrences of *tuna fish* and *mayonnaise* and their respective specifications can be interpreted as reflecting the noncentrality of that information for the development of the topic, i.e. the central referents (generic terms), *tuna fish*, *mayonnaise* have already been mentioned; the comments concerning the fact that it is good if it is tuna fish in water, and that the mayonnaise should preferably be light, although new information, are relatively unimportant as regards the development of the central theme.

CONCLUSION

Since our material is restricted, we cannot be sure of the generality of the observations made here concerning downstepping. Nevertheless, by pinpointing the environments in terms of the topic structure and related lexical semantic correlates, we can test these hypotheses against more extensive data in future studies.

ACKNOWLEDGEMENTS

This work was carried out under a contract from the Swedish Language Technology Programme (HSFR-NUTEK). We would like to acknowledge assistance from Marcus Filipsson and Birgitta Lastow in developing the analysis by synthesis system.

REFERENCES

[1] Bruce, G., Granström, B., Gustafson, K., House, D. and Touati, P. (1994), "Modelling Swedish prosody in a dialogue framework", *Proc. ICSLP 94*, pp. 1099-1102, Yokohama, Japan.

[2] Beckman, M. (1995), "A typology of spontaneous speech", *Proc. ATR International Workshop on Computational Modeling of Prosody for Spontaneous Speech Processing*, pp. 2.23-2.34, Kyoto.

[3] Touati, P. (1995), "Pitch range and register in French political speech", *Proceedings of the ICPhS -95*, Stockholm.

[4] Pitrelli, J., Beckman, M., and Hirchberg, J. (1994), "Evaluation of prosodic transcription labeling reliability in the ToBI framework", *Proc. ICSLP 94*, pp. 123-126, Yokohama, Japan.

[5] Pierrehumbert, J. and Beckman, M. (1988), *Japanese tone structure*, Cambridge, MA: The MIT Press.

[6] Grønnum, N. (1992), *The groundworks of Danish intonation*, University of Copenhagen: Museum Tusculanum Press.

[7] Bruce, G. (1982), "Developing the Swedish intonation model", *Working Papers*, 22 (Dept. of Ling. Lund Univ.), pp. 51-116.

[8] Bruce, G., Touati, P., Botinis, A., and Willstedt, U. (1988), "Preliminary report from the KIPROS project", *Working Papers*, 33 (Dept. of Ling., Lund Univ.), pp. 23-50.

[9] Bruce, G. and Touati, P. (1992), "On the analysis of prosody in spontaneous speech with exemplification from Swedish and French", *Speech Communication* 11, pp. 453-458.

[10] Cruse, D. A. (1986), *Lexical semantics*. Cambridge: Cambridge U.P.

PROSODIC CUES FOR INFERRING AND REALIZING DISCOURSE RELATIONS

Ernst Buchberger
*Austrian Research Institute for Artificial Intelligence (ÖFAI), Vienna, Austria**

ABSTRACT

This paper reports on investigations into the relationship between prosody and discourse relations. We present the analysis of an example, situated in a formal framework composed of TSM for the representation of prosody and (an extension to) DRT for the representation of discourse, and we argue that the results can be exploited fruitfully for two main areas in the automatic processing of language and speech: analysis (parsing) and synthesis (generation).

INTRODUCTION

In computational linguistics and AI, much research has been performed concerning the structure of discourse, e.g. [1,2,3]. It has been noted that "there may even be a hope of using the various sorts of clues in a program to discover with a reasonable probability of success the underlying rhetorical relations" [3, p.310] and that "rhetorical relations may, though need not, be explicitly signaled by some expression in the text" [3, p.264]. Research specifying this relation between rhetorical relations and cue words includes among others [4,5]. Up to now, the search for cues has been mainly restricted to written text, though. We have investigated the possibility that clues may also be taken from prosody.

Our analyses are based on an adaptation of Fery's [6] application of Pierrehumbert's tone sequence model (TSM) [7] to German with regard to prosodic modelling, and on Asher's [3] extension of Disourse Representation Theory (DRT) [8], which integrates the notion of discourse relations into DRT.

AN EXAMPLE

The example below is one of a number of small discourses taken from German radio news which we have investigated . The text ("Nachrichten39" - see below) has been hand-labelled with ToBI Labelling Tools according to the conventions developed at IMS [9] (see figure 1).

Im Berliner Prozeß gegen zwei Polizeibeamte wegen der Mißhandlung eines Vietnamesen sind die beiden Angeklagten heute freigesprochen worden. Das Gericht wertete die Aussagen des geschädigten Vietnamesen als widersprüchlich. Offensichtlich hätten sich die Beamten zwar sehr barsch benommen, eine strafbare Tat sei aber nicht erkennbar.

The corresponding Structured Discourse Representation Structure SDRS (simplified) is depicted in figure 2. Be-

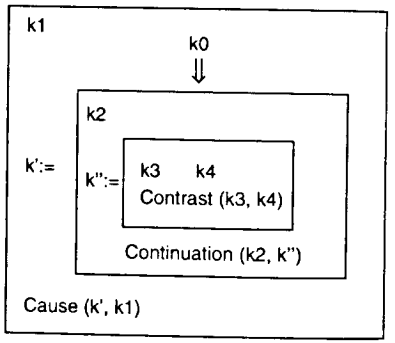


Figure 2. SDRS Representation of Text "Nachrichten39"

fore we start discussing how to arrive at this SDRS representation, we have to bear in mind that we should not expect to derive all elements of its structure purely by means of prosody. We will, however, show where prosodic cues may help us to arrive at a correct interpretation.

The basic constituents of our representation can be found by drawing regression lines through the F0 curve [10]. At a point where a noticeable discontinuity is found, a new line (and with it, a new DRS element) will start. This gives us the four DRSs k1 to k4. k1 corresponds to the first sentence, k2 to the second, k3 to the clause *Offensichtlich ... benommen*, and k4 to the rest of the third sentence (k0 does not correspond to an element in the text, but has to be constructed according

*Work reported here was performed while the author was a guest researcher at the Institute of Natural Language Processing (IMS), Stuttgart, Germany

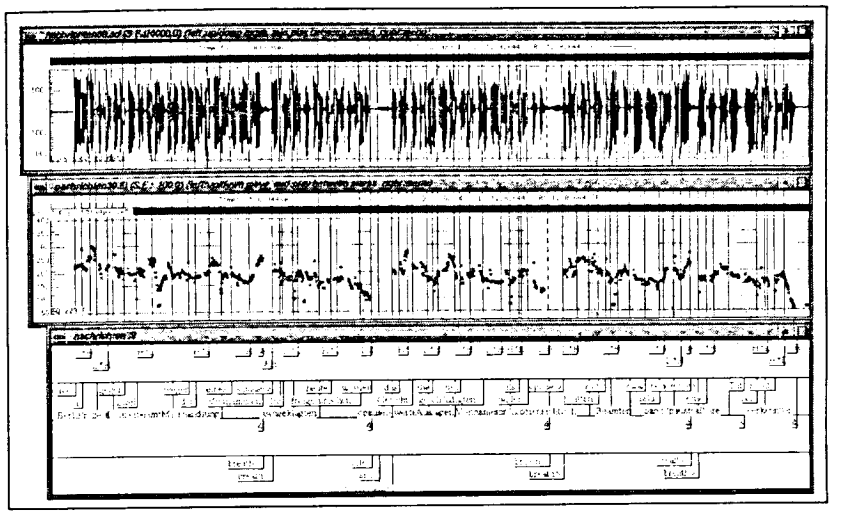


Figure 1. Signal, F0, and Labels for Text "Nachrichten39"

to Asher's topic construction rules). After having processed the first sentence and constructed k1, we find k2 with a compressed pitch range. This compression, and the relatively high ending of k2 suggests a Continuation relation to hold between k2 and what follows. k2 forms part of an SDRS k' which will be constructed only later on. We will, however, ask already at this point whether a clue can be found how this SDRS will be linked to k1. There is indeed one to be found, which is mainly inferential in nature along the lines of Dahlgren's naive semantics [11], but there is also a prosodic marking present, namely the conspicuous pitch accent on *widersprüchlich* ("contradictory"), marking the Cause for k1. In a news text, this prosodic marking is quite advantageous, especially for an inattentive hearer of radio news, (who might be listening to the news while being occupied with other tasks) whose attention is thus drawn to the relevant facts. The connection between k3 and k4 can be clearly seen from the progredient phrase tone % at the end of *benommen*. In the absence of further information we would take this as an indication for a Continuation relation between k3 and k4, but here we have the lexical clue of the word *aber* ("but") from which we infer a Contrast relation. The pitch accents help to find the contrastive elements *strafbar* ("punishable") and *nicht erkennbar* ("not to-be-found").

Finally, the final lowering is quite pronounced, which means that closing-off of the yet open discourse structures can be inferred easily. This way, k'' will be closed off and the Continuation relation between k2 and k'' we already mentioned above can be constructed now, forming the SDRS k'. What remains to be done is the construction of the Cause relation between k1 and k' which we also had mentioned already, completing the representation.

EXPERIMENTAL VALIDATION

Up to now, recorded speech data from radio news have been hand-labelled and analyzed with regard to the discourse relations involved.

Recently, we have started resynthesizing texts with contours suggesting discourse relations differing from those found in the original texts, using a testing environment which allows for a free modification of pitch and duration of the recorded speech data before the resynthesis based on the PSOLA algorithm is performed [12]. The effects of these resynthesized texts on human listeners are being tested.

APPLICATIONS

Prosody is considered an important factor for both automatic natural language analysis and synthesis (or parsing and generation). The relation between parsing and generation has often been thematized in Natural Language Processing, specifically regarding grammar [13]. The role of prosody can be summarized as follows: in analysis, it helps to disambiguate, in synthesis, it allows for more natural utterances.

As for analysis, the usefulness of (prosodically derived) boundary information for disambiguating syntactically ambiguous utterances has been shown by other researchers [14,15]. At IMS the effects of pitch accents on pronoun resolution [16,17] are currently investigated, some of which have also an impact on discourse relations. As for ambiguous discourse structure, it seems that discrimination will be mostly relevant not so much for the kind of discourse relation, but mainly for the attachment points of these discourse relations (cf. [18,19]).

With regard to generation, whereas previous research has often focussed on conciseness, based on Grice's maxims [20], everyday utterances show a significant amount of redundancy. Taking prosody into account, it may be advantageous to code discourse relations prosodically and textually, e.g., by not only using cue words, but also marking them prosodically in a consistent way. Thus, even if the lexical clues alone would have allowed inferring the intended meaning, prosodic clues might enhance the textual clues and lead to more natural utterances.

The aspect of unwanted implicature has recently been noted in the generation literature [21]: When in a (written) report an entity is singled out by ascribing to it certain properties, the impression is created that for other non-mentioned entities this property does not hold. The example presented in [21] can be taken over to speech: If we say "Person A worked on task B. HE finished the work in time.", putting (erroneously) a pitch accent on "he", this will create the implication that there were other people involved in the project who did not finish in time.

Finally, during the last years much work in Natural Language Generation has been based on some sort of model for representing discourse relations, most notably RST [22,23]. With a specification of correspondences between a formal representation of discourse and a model of prosody we can hope to find a way to integrate Natural Language Generation and Speech Synthesis in an elegant way.

PRELIMINARY CONCLUSIONS AND FUTURE WORK

We have found some evidence for prosodic marking of discourse relations and sketched a framework for the investigation of this interdependency. The relation between prosodic phenomena and discourse relations is not a bijective function, rather it has to be seen as clues suggesting a certain structure of discourse, but not enforcing it.

The tone sequence model (TSM) has proven useful for describing some aspects of local discourse structure, and partially also for global discourse structure. What is still lacking is a formal description of other relevant parameters for global discourse structure like pitch range. Also missing is a formal description of the interaction of the various sorts of discourse clues, integrating the contribution of lexical and prosodic cues, among others. Asher has proposed to use nonmonotonic logic to capture the dependency of various discourse relations, but he did not talk about prosody. Maybe prosodic cues could be integrated into such a framework. Another possible candidate might be optimality theory [24], which has recently been shown to be a useful framework for the description of various aspects of prosody, but has – to our knowledge – not yet been applied to the description of interactions between prosody and discourse.

ACKNOWLEDGEMENTS

The research reported here forms part of the multi-national project *Scientific Cooperation in the European Network in Language and Speech* (Contract No. CHRX-CT93-0421). It was performed during a stay at the Institute of Natural Language Processing (IMS), Stuttgart, Germany, and sponsored by the EU's HCM programme. Many thanks to Grzegorz Dogil for fruitful discussions and useful hints, to Jörg Mayer and Stefan Rapp for help with labelling and graphics, and to all other members of IMS / Experimental Phonetics for the cooperative working atmosphere.

REFERENCES

[1] Grosz B.J., Sidner C.L. Attention, Intentions, and the Structure of Discourse, Computational Linguistics 12(3)175-204, 1986.

[2] Mann W.C., Thompson S.A.: Rhetorical Structure Theory: A Theory of Text Organization, in Polanyi L.(ed.), The Structure of Discourse, Ablex, Norwood, N.J., 1987.

[3] Asher N.: Reference to Abstract Objects in Discourse, Kluwer, Dordrecht, 1993.

[4] Scott D.R., Souza C.S.: Getting the Message Across in RST-based Text Generation, in Dale R., et al.(eds.), Current Issues in Natural Language Generation, Academic Press, New York, 1990.

[5] Rösner D., Stede M.: TECHDOC: A System for the Automatic Production of Multilingual Technical Documents, in Görz G.(ed.), KONVENS 92, Springer, Berlin, 1992.

[6] Féry C.: German Intonational Patterns, Niemeyer, Tübingen, 1993.

[7] Pierrehumbert J.B.: The Phonology and Phonetics of English Intonation, Ph.D. Thesis, MIT, Cambridge, MA, 1980.

[8] Kamp H., Reyle U.: From Discourse to Logic, Kluwer, Boston/Dordrecht/ London, 1993.

[9] Mayer J.: A Guide to Labelling within the Stuttgart Labelling System, Technical Report, IMS, Univ. Stuttgart, forthc.

[10] Huber D.: On the Discourse Function of Intonation, Proceedings ICPhS, Aix-en-Provence, 5:191-193, 1991.

[11] Dahlgren K., McDowell J., Stabler E.P.Jr.: Knowledge Representation for Commonsense Reasoning with Text, Computational Linguistics 15(3)149-170; 1989.

[12] Möhler G., Dogil G.: Test Environment for the Two Level Model of Germanic Prominence, Eurospeech '95, Madrid, 1995.

[13] Strzalkowski T.(ed.): Reversible Grammar in Natural Language Processing, University of California, Berkeley, CA, 1991.

[14] Lehiste I.: Phonetic Disambiguation of Syntactic Ambiguity, Glossa 7, 197-222, 1973.

[15] Ostendorf M., Price P., Bear J., Wightman C.W.: The Use of Relative Duration in Syntactic Disambiguation, in Proceedings of the DARPA Speech and Natural Language Workshop, Morgan Kaufmann, 1990.

[16] Mayer J.: Prosodische Disambiguierung von Anaphern im Diskurs, AG 10 Suprasegmentale Phonologie: Modelle und Prozesse, DGfS-Jahrestagung Göttingen, March 1-3, 1995.

[17] Dogil G.: Prosodic Cues to "Alternative" Semantics in DRT, unpublished talk held at the HCM Workshop on Discourse and Dialogue Prosody in Stuttgart, Feb 14-15, 1995.

[18] Pierrehumbert J.B., Hirschberg J.: The Meaning of Intonational Contours in the Interpretation of Discourse, in Cohen P.R., Morgan J., Pollack M.E.(eds.), Intensions in Communication, MIT Press, Cambridge, MA, 1990.

[19] Hobbs J.R.: The Pierrehumbert-Hirschberg Theory of Intonational Meaning, in Cohen P.R., Morgan J., Pollack M.E.(eds.), Intensions in Communication, MIT Press, Cambridge, MA, 1990.

[20] Grice H.P.: Logic and Conversation, in Cole P., Morgan J.L.(eds.), Syntax and Semantics 3: Speech Acts, Academic Press, New York, 1975.

[21] André E., Buchberger E., Cawsey A., Korelsky T., Not E., Rösner D., Teich E.: Report from Working Group 3 - Text Structure and Architecture in Natural Language Generation, in Höppner W., Horacek H., Moore J. (eds.), Principles of Natural Language Generation, Dagstuhl Seminar Report, Schloss Dagstuhl, Germany, 1995.

[22] Hovy E.H.: Generating Natural Language under Pragmatic Constraint, Lawrence Erlbaum, Hillsdale, NJ, 1988.

[23] Moore J.D., Paris C.L.: Planning Text for Advisory Dialogues: Capturing Intentional and Rhetorical Information, Computational Linguistics 19(4), 651-694, 1994.

[24] Prince A., Smolensky P.: Optimality Theory, Technical Reports of the Rutgers University Centre for Cognitive Science, TR-2, 1993.

THE FUNCTION OF INTONATION IN SPONTANEOUS AND READ DIALOGUE

J. C. Kowtko
CSTR & HCRC, University of Edinburgh, U.K.

ABSTRACT

Independently motivated analyses of intonation and discourse function are employed in a study of intonation function in task-oriented dialogue (taken from the HCRC Map Task Corpus [1]). Results from a comparison of the two analyses performed on single-word utterances in spontaneous dialogue show that intonation contour is a significant factor in utterance function. An examination of read-aloud dialogue reveals similar results with different patterns of correlation.

INTRODUCTION

Recent work on intonation in dialogue tends to follow one of two opposite approaches: it either describes very general discourse functions (e.g. [6] connecting, continuing and segmenting) or it identifies very specific discourse contexts (e.g. [3] on anaphor distribution and turn-taking). In order to make progress in this area, we need to combine these two approaches. This in turn requires an independent description of dialogue context as the basis for a robust account of intonational function. Such an independent description is the conversational games analysis outlined in [4]. It is employed here in a study of intonation function in single-word utterances taken from spontaneous and read-aloud dialogue.

Using the conversational games analysis to represent discourse function, and assuming that intonation plays a significant role in discourse function, we expect to find tune patterns by looking at move x in game y. We also expect to discover a difference between results in spontaneous and read-aloud dialogue. The nature of the difference interests us because we could possibly use the more easily constrained read-aloud dialogue to train speech recognisers.

DIALOGUE ANALYSIS

The analysis proposed by Kowtko *et al.* [4] involves interactional exchanges in dialogue, called *conversational games*. Games embody the linguistic interaction, e.g. initiation, response and feedback, which arises from non-linguistic goals. (See [7] for a discussion on the relationship between linguistic processes and underlying action goals.) A game consists of the turns necessary to accomplish a conversational goal. The components of games, *moves*, are defined in terms of speaker intention and dialogue function.

The data used in this study are based upon dialogues arising from the map task [1]: One person has a map with a route marked on it and has to tell another person how to draw the route onto a similar map. Neither participant can see the other's map.

The nature of the task is such that the speakers' intentions are clear to the analyst most of the time. Kowtko *et al.* report that one expert and three naïve judges agree on an average of 83% of the moves classified in two map task dialogues.

Six games appear in the dialogues: Instructing, Checking, Querying-YN, Querying-W, Explaining, and Aligning. They are initiated by the following moves: INSTRUCT (Provides instruction), CHECK (Elicits confirmation of known information—tests speaker's knowledge), QUERY-YN (Asks yes-no question for unknown information), QUERY-W (Asks content, *wh*-, question for unknown information), EXPLAIN (Gives unelicited description), and ALIGN (Aligns hearer's knowledge or beliefs—checks alignment of position in task).

Six other moves provide response and additional feedback: CLARIFY (Clarifies or rephrases given information), REPLY-Y (Responds affirmatively), REPLY-N (Responds negatively), REPLY-W (Responds with requested information), ACKNOWLEDGE (Acknowledges and requests continuation), and READY (Indicates intention to begin a new game).

Since the map task involves one participant instructing the other concerning how to draw the route, the conversations naturally consist of many Instructing games. The conversational games analysis allows for nesting of games and looping of response and feedback moves within games. The prototypical game consists of two or three moves: initiation, response, and optionally feedback. The large majority of games (84% from a sample of 3 dialogues, $n = 65$) match the simple prototype. Games that do not match the prototype are still well-formed, having extra response–feedback loops, nested games, or extra moves. Very few games (less than 2%) break down as a result of a misunderstanding or other problem.

Conversational game structure offers a taxonomy which specifies both the function and context of an utterance, as a specific move at a specific point within a specific game. This facilitates the study of the function of intonational tune since the tune reflects an utterance's conversational role.

INTONATION ANALYSIS

Single-word utterances which comprise whole conversational moves from the HCRC Map Task Corpus ([1] containing Scottish English) have been analysed in terms of intonation contour. A set of five intonational tunes was determined to best represent the data: High Level (Hi; level tune high in the speaker's local pitch range), Low Level (Lo; level tune low in the speaker's local pitch range), Fall (F; simple fall in pitch overall), Rise (R; simple rise in pitch overall), Fall-Rise (F-R; distinct falling pitch followed by a rise). The complementary sixth tune, rise-fall, did not appear in the data examined.

FUNCTION OF INTONATION

A study comparing intonation contour and dialogue function was performed on single words which individually comprise conversational moves and intonational phrases in the HCRC Map Task Corpus (120 from spontaneous dialogues and 120 from dialogues read aloud by the original participants using transcripts of the spontaneous dialogues). The data come from three whole dialogues. They involve 1 male and 5 female speakers who do not know each other, and total 12 minutes of spontaneous speech and 11 minutes of read speech. They sound fairly natural in the read-aloud condition and have good quality audio recordings.

All single-word moves in the three dialogues were considered as data points. The only words removed from the study are those for which the pitch trace failed (e.g. croaky, unintelligible utterances) and those which form partial intonational phrases.

The words used in the study are *almost, aye, ehm, mmhmm, no, okay, okey-dokey, right, rightee-ho, uh-huh, yeah, yes,* and *yup.* They appear as 6 of the conversational moves [4]: ALIGN, REPLY-Y, REPLY-N, REPLY-W, ACKNOWLEDGE, and READY. Each word is transcribed intonationally as high level (Hi), low level (Lo), rise (R), fall (F), or rise-fall (F-R). Their discourse function is represented by the framework of conversational games. The intonation of these words was compared with discourse function in terms of the move type and the game in which

it occurs.

Spontaneous Results

Results are shown with three intonation categories: rise + high level (R+Hi), low level + fall (Lo+F), and fall-rise (F-R). The motivation for this clustering is two-fold. Firstly, the level tunes may be phonetic variations of underlying pitch accent glides. Many of the low level tunes, for example, exhibit a detectable fall in pitch of approximately 8 Hz, but were categorised as level overall. Secondly, the results show some clustering of intonation categories, especially between low level and falling tunes.

From the spontaneous dialogues, significant correlations[1] emerge between intonation and discourse categories. Table 1 displays the results in terms of intonational tune associated with conversational move in a game. ALIGN moves significantly rise in pitch. The majority of REPLY-Y tunes fall, and the category is significantly falling and low level. REPLY-N moves, although small in number, exclusively fall in pitch. Most ACKNOWLEDGE moves are rising or high level. The two READY moves fall in pitch.

Looking at ACKNOWLEDGE moves in their game context adds some clarity to the pattern of tunes. In *Instructing* games, rises and high levels form a majority. Moves in *Querying-W* games show a similar significant pattern of rising tunes. Results in *Querying-YN*, *Checking*, and *Explaining* games are insignificant althought the first two tend toward falling and low level, and the third toward low level tunes.

These correlations indicate that discourse function, as defined by move type, is a principal factor in determining intonation contour.

Read Results

Results of the analysis of read aloud dialogue also show significant correlations between intonation and discourse

[1]Numbers which achieve significance at $p < .05$ by the Kolmogorov-Smirnov One-Sample Test are marked with daggers in the table.

Table 1: *Spontaneous Intonation v. Conversational Move*

Move	*R+Hi*	*Lo+F*	*F-R*	#
ALIGN	†7	1		8
REPLY-Y	6	†23		29
REPLY-N		4		4
REPLY-W	1			1
ACKNWL.	41	32	3	76
READY		2		2

categories. They reveal a pattern somewhat different from that in spontaneous speech. The tunes in some discourse categories have shifted toward low level and falling tunes. Table 2 shows a summary.

ALIGN moves significantly rise. REPLY-Y moves are significantly falling and low level. REPLY-N moves are all falling and low level tunes. The majority of ACKNOWLEDGE moves are falling and low level. READY moves fall in pitch.

Examining the moves in their game contexts reveals a pattern different from the overall one only in the case of ACKNOWLEDGE moves. All of the rising tunes except one occur within *Instructing* games. Here the data are shared between the rising and falling categories. Moves in *Querying-W* games tend toward falling and low level tunes. ACKNOWLEDGE moves in *Querying-YN* and *Checking* games are all falling or low level. The moves in *Explaining* games are exclusively level in pitch, and most of these are high level tunes.

Again, the significant correlations and clear trends found in the results of read dialogue suggest that intonation contour is an important factor in determining discourse function.

Discussion: Spont. v. Read

One of the most noticeable differences between spontaneous and read dialogue is a shift in the read condition toward low level and falling tunes in the response moves (REPLY-Y, REPLY-N, REPLY-W, and ACKNOWLEDGE). This shift is consistent with general knowledge about reading style and various

Table 2: *Read Intonation v. Conversational Move*

Move	*R+Hi*	*Lo+F*	*F-R*	#
ALIGN	†8			8
REPLY-Y	5	†24		29
REPLY-N		4		4
REPLY-W		1		1
ACKNWL.	28	47	1	76
READY		2		2

comparisons of the two modes of speech (e.g. [2] which states that at boundaries, read speech contains more falling tunes).

The increased presence of low level and falling tunes in read dialogue is most visible in ACKNOWLEDGE moves. There is a shift from a majority rise and high level pitch pattern to a fall and low level one. ACKNOWLEDGE moves in *Querying-W* games change from a significant pattern of rising tunes to a trend toward falling and low level tunes.

CONCLUSION

Discourse structure correlates with intonation contour, showing that intonation is a significant factor in dialogue function. The similar patterns in results from spontaneous and read dialogue encourage us regarding the use of read speech data to train speech recognition systems. They suggest that read-aloud dialogue could provide a partial substitute for that of spontaneous dialogue. We must, however, be careful in accepting read dialogue for analysis because of the increased number of low level and falling tunes in read speech. The increased number of these tunes could be related to the speaker's level of confidence while performing the task. In reading the transcript and acting out the map task dialogue, the participant may feel more confident than while originally doing the map task. This possibility is left to future research.

ACKNOWLEDGEMENT

Thanks are due to S. Isard and D. R. Ladd for their supervision on this work. Partial funding was provided by a UK Overseas Research Student Award.

REFERENCES

[1] Anderson, A H, M Bader, E G Bard, E Boyle, G Doherty, S Garrod, S Isard, J Kowtko, J McAllister, J Miller, C Sotillo, H Thompson, and R Weinert (1991), "The HCRC Map Task Corpus," *Language and Speech*, **34**: 351-366.

[2] Blaauw, E (1994), "The contribution of prosodic boundary markers to the perceptual difference between read and spontaneous speech," *Speech Communication*, **14**: 359-375.

[3] Hockey, B A (1992), "Prosody and the interpretation of cue phrases," *Proceedings of the IRCS Workshop on Prosody in Natural Speech*, University of Pennsylvania, IRCS Report No.: 92–37, 71-77.

[4] Kowtko, J C, S D Isard and G M Doherty-Sneddon (1992), "Conversational games within dialogue," Research Paper HCRC/RP-31, Human Communication Research Centre, University of Edinburgh.

[6] McLemore, C A (1991), *The Pragmatic Interpretation of English Intonation: Sorority Speech*, Ph.D. dissertation, University of Texas at Austin.

[7] Power, R (1979), "The organisation of purposeful dialogues," *Linguistics*, **17**: 107-152.

PERCEPTUAL CHARACTERISTICS OF VOICE QUALITY IN DUTCH MALES AND FEMALES FROM 9 TO 85 YEARS

J. van Rie & R. van Bezooijen
University of Nijmegen, Nijmegen, The Netherlands

ABSTRACT

This study provides a voice quality description of 180 Dutch speaking men and women from 9 to 85. The description was focused on voice characteristics such as pitch, tempo, and whisper. Computed were correlations among age and voice quality, and the mutual correlations among the voice characteristics. Finally, it was investigated how well age can be predicted on the basis of voice quality characteristics.

INTRODUCTION

Voice quality is the overall auditory colouring of an individual speaker's voice to which both laryngeal and supralaryngeal features contribute [1]. Disciplines such as speech pathology, psychology and psychiatry have studied this topic for a long time. Voice quality in normal speech, however, has been neglected, in foreign and Dutch studies alike. Voice quality studies that have been carried out are limited in sex [2, 3, 4, 5, 6, 7, 8], age [9, 2, 6, 8, 5], and voice quality characteristics described: mostly only pitch [2, 3, 6, 7, 4, 8]. Furthermore, most research has been conducted on the basis of reading texts. The aim of this study is to gain insight in the voice quality parameters in Dutch on the basis of spontaneous speech. The central questions are:

1. What is the correlation between age and voice quality parameters?
2. What is the relation between sex and voice quality parameters?
3. What is the mutual correlation between different voice quality parameters?
4. How well can age be predicted from voice quality parameters?

METHOD

Non-pathological speech of men and women from 9 to 84 was recorded. Eighteen age categories were distinguished: 9-10, 11-12, 13-14, 15--16, 17-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79, and 80-84. Five speakers were selected per sex and per age category. This resulted in 90 speakers for both sexes. Generally, each age category stretched over a period of five years. However, for the younger speakers (9 to 19) the interval was smaller because the breaking of the voice for boys, can be very sudden, whereas the exact moment can vary.

Because it is as yet unknown to what extent voice quality differs for different regional and urban varieties of Dutch, we restricted the description to speakers of Standard Dutch. The material consisted of 45 sec semi spontaneous descriptions of drawings depicting daily events or of spontaneous speech about topics like the speaker's eating habits.

The speech was scored on 21 voice quality parameters. The voice protocol was an extended version of the one described in [10]. The speech aspects covered by the scales were: prosodic aspects, temporal organization, articulation, and phonation. Two types of scales were used: 7-point scales and 4-point scales. The, bipolar, 7-point scales relate to intrinsic parameters of the speech signal that can take different values, e.g. pitch. They go from one extreme (-3) via a neutral reference point to the opposite extreme (+3). The, unipolar, 4-point scales rate features from absent (0) to strongly present (+3); these parameters can only depart from the neutral point of reference into one direction, e.g. harshness. The scales used in this description only cover the non-pathological area of speech. The scales are:

1. pitch (pit): impression of average pitch level, related to speaker's sex (7-point scale)
2. pitch range (prn): distance between the lowest and the highest pitch (7)
3. pitch variance (pva): frequency and/or strength of pitch changes (7)
4. stress (str): frequency and/or strength of accents (7)
5. tremulousness (tre): trembling quality (4)
6. fluency (fln): fluency degree (4)
7. pauses (pau): frequency and/or duration of silent intervals (7)
8. tempo (tem): speech rate (7)
9. tempo variance (tva): frequency and/or amount of rate changes (7)
10. effort (eff): loudness impression (7)
11. effort range (ern): distance between the softest and loudest utterance (7)
12. effort variance (eva): frequency and/or strength of loudness changes (7)
13. precision of articulation (art): extent to which the target positions of the distinct speech sounds are reached
14. whisper (whs): amount of escaping air through the glottis (4)
15. harshness (har): rough and rasping quality - aperiodic vibration (4)
16. creak (cre): discrete pulses can be perceived in phonation (4)
17. sonority (son): extent to which voice sounds superficial/sharp or warm/resonant (7)
18. tension (ten): impression of the muscle tension in the vocal folds (7)
19. audible breath (bre): amount of audible inadequate breathing (4)

The speech parameters were rated by the authors and a speech therapist. A consensus description was made: the scores were compared and, if possible, adjusted if the scores on a scale differed more than one scale position. The percentage adjusted scores for all raters per scale was less than 5%. On average, the percentage adjusted scores was 2.4%.

RESULTS AND DISCUSSION

Correlations between age and voice quality parameters

To gain insight into the coherence between the speakers' age and the 19 voice quality parameters Pearson product-moment correlations were calculated.

The older men and women get (table 1), the lower their perceived pitch becomes and the more creaky and sonorous their voices sound. Age related pitch has been studied rather extensively. It appears that fundamental frequency (F0) of babbling babies, is

Table 1. Correlations between the speaker's age and the 19 voice quality parameters. Significant r-values at .01 (two-tailed) are printed in bold.

	♂	♀	♂ + ♀
pit	**-.46**	**-.31**	**-.38**
prn	-.10	.21	.05
pva	-.21	.10	-.06
str	-.20	.20	-.01
tre	.23	.14	.19
fln	-.18	**-.28**	**-.23**
pau	.25	.14	.19
tem	**-.28**	-.19	**-.23**
tva	.25	**.30**	**.27**
eff	.01	.09	.05
ern	.11	.22	.16
eva	.07	.23	-.15
art	-.02	**.32**	.14
whs	.11	-.15	-.03
har	.01	.02	.02
cre	**.41**	**.41**	**.40**
son	**.45**	**.28**	**.35**
ten	.06	-.19	-.08
bre	**.40**	.23	**.31**

about 400 Hz [9]. Till puberty there is no difference between both sexes, F0 being about 200 to 250 Hz [8]. Then pitch lowers till adulthood is attained: F0 for adult males becomes approximately 100 Hz but shifts upward from about 65 years [3, 5] and F0 for adult females decreases to approximately 190 Hz with no further systematic change [11, 12]. This U-shaped curve is also observed in our data: till the age of about 20 perceived pitch decreases, till about 50 to 60 it stays more or less stable to shift upwards from about 60, especially for men. This U-curve masks the linear correlation till about 60 years and probably causes the low r-values. Calculating correlations omitting all speakers over 60, indeed causes higher r-values, viz -.69 for men (for whom the U-curve was observed), -.54 for women and -.61 for all speakers.

At the same time both creak and sonority correlate positively with age. This is not surprising, since for both sexes these parameters show a (fairly high) significant correlation with pitch (table 2). As for creak, the increase with age could be explained in physiological terms, related to a change in the vibration of the vocal folds. It might also be possible that older speakers adapt to socially based, stereotypical ideas about how the voice of an older person is expected to sound.

Table 2. Correlation matrix for men (upper right half) and women (bottom left half). Significant r-values at .01 are printed in bold. The decimal dot is omitted.

♂ / ♀	pit	prn	pva	str	tre	fln	pau	ten	tva	eff	ern	eva	art	whs	har	cre	son	ten	bre
pit	1.0	33	41	19	03	-16	-02	08	-28	03	-30	-12	00	08	38	-60	-68	-45	04
prn	25	1.0	74	46	02	02	05	07	25	13	29	25	-04	14	22	-17	-26	-27	-01
pva	32	76	1.0	48	-01	-01	-15	22	14	16	22	23	-10	09	25	-19	-31	-19	-11
str	05	37	44	1.0	01	-04	-10	17	09	40	34	39	14	02	25	-20	-36	-15	-23
tre	17	17	17	10	1.0	11	09	-11	-13	08	-09	-02	-11	-00	05	20	12	-08	20
fln	25	12	23	14	00	1.0	07	-08	-11	-05	11	-10	-25	01	-11	09	02	14	-18
pau	-02	06	-05	22	03	03	1.0	-38	07	-14	06	-10	-02	11	07	04	-06	-08	09
ten	10	10	16	-19	04	07	-50	1.0	12	15	12	23	-16	05	-10	-19	-12	12	-11
tva	-05	37	42	36	04	17	03	-01	1.0	14	47	46	-00	21	-05	08	13	-02	-17
eff	24	12	28	41	22	-05	01	04	19	1.0	28	29	07	-11	23	03	-13	-09	-02
ern	03	41	45	39	22	14	12	-01	56	32	1.0	73	11	04	-07	09	07	10	-06
eva	05	46	51	49	16	02	08	-00	59	47	78	1.0	16	07	03	-05	-07	-11	-08
art	-10	25	15	27	-07	-10	28	-26	19	33	14	22	1.0	-18	08	-11	01	11	-10
whs	-00	-09	-15	-04	-08	02	10	-03	-16	-32	-10	-13	-08	1.0	30	03	-34	-61	32
har	32	07	05	-07	29	21	11	04	-04	05	05	02	-39	05	1.0	-09	-31	-49	18
cre	-28	03	06	06	11	02	04	-07	31	16	26	25	13	-14	-05	1.0	42	10	13
son	-69	-24	-26	01	-13	-32	-04	-28	-04	-05	-07	-05	22	-16	-47	25	1.0	66	05
ten	-37	-07	-03	-08	-28	-12	-24	-03	-02	-03	-05	-07	15	-50	-59	05	55	1.0	-24
bre	-12	-02	-11	-04	10	02	03	-00	00	-09	-00	-06	08	31	-31	10	-03	-19	1.0

Interrelationship among the voice quality parameters

To get a clearer picture of the interrelationship among the different voice quality parameters, factor analyses were conducted. Factor analysis was succesful for the male speakers, but failed for the female speakers because communality of a variable exceeded 1.0. Therefore we present the correlation matrix (table 2).

Seven factors could be extracted for the male speakers, four of which exceeded an eigenvalue of 1.0. These four factors, explaining 45% of the variance, are a) laryngeal voice quality with high loadings of tension, whisper, harshness, and audible breath; b) prosodic variation with loadings of effort range, effort variance, and tempo variance; c) pitch variation with loadings of pitch range and pitch variation, and d) pitch with loadings of sonority, creak and pitch. These factors explained 18.4%, 13.1%, 8.2%, and 5.3% of the variation respectively. Inspection of the correlations for the women in table 2, reveals similar sets of correlating parameters.

Multiple regression

In order to know how well age can be predicted on the basis of voice quality parameter, we performed a stepwise multiple regression analysis. The powerfull predictors for the male speakers were: pitch, audible breath, pauzes, fluency, tension and sonority, which explained 60% of the age variance. In predicting women's age only 48% of the age variance could be explained by the predictors creak, fluency, articulation, harshness, audible breath, tempo variance and sonority.

ACKNOWLEDGEMENT

We express our gratitude to H. Heikens, R. van Hout, R. van Ameijde, L. de Bruin and H. Kraayeveld for lending us some of their speech material.

REFERENCES

[1] Laver, J. (1979), "The description of voice quality in general phonetic theory", In: *Work in progress*, Dept. of Linguistics, University of Edinburgh, vol. 12, pp. 30-52.

[2] Hollien, H. & Hollien, P. (1972), "A cross-cultural study of adolescent voice change in european males". In: Rigault, A. & Charbonneau, R. (eds.) *Proceedings of XIIth ICPhS 71*. The Hague: Mouton, pp. 332-337.

[3] Hollien, H. & Shipp, T. (1972), "Speaking fundamental frequency and chronological age in males", *JSHR*, vol. 15, pp. 155-159.

[4] Hollien, H. & Malcik, E. (1967), "Evaluation of cross-cultural studies of adolescent voice change in males", *Speech Monographs*, vol. 34, pp. 80-84.

[5] Mysak, E.D. (1959), "Pitch and duration characteristics of older males", *JSHR*, vol. 2, pp. 46-54.

[6] Duffy, R.J. (1970), "Fundamental frequency characteristics of adolescent females", *Language and Speech*, vol. 13, pp. 14-25.

[7] Hollien, H. & Paul, P. (1969), "A second evaluation of the speaking fundamental frequency characteristics of post-adolescent girls", *Language and Speech*, vol. 12, pp. 119-124.

[8] Michel, J.F., Hollien, H. & Moore, P. (1966), Speaking fundamental frequency characteristics of 15, 16 and 17 year-old girls", *Language and Speech*, vol. 9, pp. 46-51.

[9] Keating, P. & Buhr, R. (1978), "Fundamental frequency in the speech of infants and children", *JASA*, vol. 63, pp. 567-571.

[10] Laver, J., Wirz, S., Mackenzie, J. & Hiller, S. (1981), "A perceptual protocol for the analysis of vocal profiles", *Edinburgh University Department of Linguistic Work in Progress*, vol. 14, pp. 139-155.

[11] Helfrich, H. (1979), "Age markers in speech". In: Scherer, K.R. & Giles, H. (eds.) *Social markers in speech*. Cambridge: Cambridge University Press.

[12] Bezooijen, van R. (1993), "Verschillen in toonhoogte: natuur of cultuur?", *Gramma/TTT*, vol. 2, 165-179.

THE INFLUENCE OF SMOKING HABITS ON PERCEIVED AGE

Angelika Braun
Speaker Identification and Tape Analysis Section, Bundeskriminalamt, Wiesbaden, Germany
Toni Rietveld
Department of Language and Speech, University of Nijmegen, The Netherlands

ABSTRACT

Direct age estimates of 40 adult male speakers, 20 of them smokers and 20 non-smokers, were made by a group of 12 trained phoneticians and a group of 19 phonetically naive listeners from recorded speech samples. The results indicate that the expert listeners did not do significantly better than the untrained listeners. Smokers were assessed to be older than non-smokers of the same calendar age. The interaction of several phonetic variables with listener judgement was investigated. Syllable rate and HNR turned out to be the only significant predictors of perceived age.

INTRODUCTION

It emerges from previous research that listeners are able to make fairly accurate judgements about a male speaker's age from voice cues. Shipp/Hollien [1] found a correlation of r = 0.88 between calendar age and perceived age; Neiman/Applegate [2] calculated a correlation of r = 0.77 based on the data published in Ryan/Burk [3]; Horii/Ryan found a correlation of r = 0.76 [4]. Several factors have been shown to influence age perception accuracy to some extent, among them listener age [5], speaker age [2], the difference between the two [1], and listener sex. [6] There is evidence from speech production experiments which suggests that physiological condition may also be an important factor in perceiving the ageing voice [7, 8]. Speakers who were in good health were found to have younger-sounding voices.[9]

Cigarette smoking is definitely a factor which will not only affect physiological condition in general but also cause histological changes in the vocal apparatus. Despite well-documented effects on the vocal cords[10], there is a striking paucity of studies approaching the subject from an acoustical point of view, and these have all focused on Speaking Fundamental Frequency (SFF) [11,12, 13]. Generally, the F0 values for the non-smokers were found to be higher than those for the smoking group.

For the present study, the following questions were of interest: (i) whether or not a speaker's smoking habits influence his perceived age; (ii) which acoustic variables are good predictors of perceived age; (iii) whether or not trained listeners are better at estimating a speaker's age than phonetically naive listeners. The last question points to a potential forensic application of this study: One of the elements in speaker profiling, i.e. the analysis of an anonymous voice, is the assessment of a speaker's age group. It would be interesting to see whether this is done more reliably by phoneticians than by untrained listeners.

PROCEDURES

The recordings as well as the production data used in this study were available from a previous investigation [14]. Specifically, a total of 40 normal-speaking male subjects, 20 of them being smokers and 20 non-smokers, provided speech samples. Smokers ranged in age from 27 to 59 yrs with an average of 41.05 yrs (SD = 9.18). They had been smoking for an average of 21.4 yrs (range: 10-40 yrs; SD = 8.3). The average number of cigarettes smoked per day was 27.5, ranging from 20 - 40 (SD = 6.2). The non-smokers were between 25 and 58 years of age with a mean of 40.48 yrs (SD=10.89).

Subjects were first asked to read a standardized text (German version of „The North Wind and the Sun") which took approximately 45 sec. They then phonated the vowel /a:/ as steadily as possible for at least 3 sec at a comfortable pitch and loudness level. Only the text was used in the perception experiment.

Listeners

Two panels of listeners took part in the perception experiment. Group I consisted of twelve phoneticians, eight of them men and four women, who had extensive experience in the forensic analysis of anonymous voices. The age range for this group was 29-62 with a mean of 40.7. Group II consisted of 19 university students with no particular training in auditory phonetics. This group ranged in age between 20 and 32 years (mean: 23.3). All listeners reported normal hearing.

The text passages read by the 40 speakers were randomized and presented to the two panels of listeners through a high-quality recorder-speaker system in a quiet room. Listeners were informed that all speakers were male adults. They were instructed to listen to each sample and make a direct age estimation for each speaker.

RESULTS AND DISCUSSION

A production study had been carried out on the basis of the same data [14], investigating the following variables: speaking fundamental frequency, jitter, shimmer, and harmonics-to-noise ratio (HNR). For the purpose of the present study, syllable rate was measured in addition because there are findings which indicate that this parameter forms an important clue for listeners [3, 15]. The production study on the data used here revealed that shimmer and HNR were more effective in discriminating the two groups than speaking fundamental frequency and jitter.

The results of the listening experiment are summarized in Table 1:

Table 1: Means and standard deviations of speakers' average chronological age and the age perceived by the two listener groups

speaker group	chron. age	s.d.	perc.age$_{exp.}$	s.d.	perc.age$_{non\text{-}e.}$	s.d.
all speakers (N=40)	40.77	9.94	41.37	9.59	40.59	11.06
smokers (N=20)	41.05	9.18	44.14	10.64	43.79	12.12
non-smokers (N=20)	40.48	10.89	38.60	7.71	37.40	9.11

An analysis of variance was carried out on the differences between the estimated and the calendar age of the 40 speakers. The ANOVA was of the 'repeated measures' type, with one between-subject factor: the two listener groups, and one within-subject factor: smokers vs. non-smokers. Only the second factor turned out to be significant: $F(1,29) = 112.84$, $p < 0.001$. This means that the calendar age of the smokers was overestimated by both groups of listeners, and that of the non-smokers was underestimated. This finding seems to be in line with findings reported by Ringel/Chodzko-Zajko [9] pertaining to speakers who are in good physiological condition. Even though we did not test the physical fitness of our speakers, it seems fair to assume that smokers of the type recorded here (i.e. at least one pack per day for a minimum of 10 years) be in less than perfect health.

Furthermore, statistical analysis (Pearson correlation; one-tailed) reveals high correlations between speakers' cal-

endar age and perceived age for the trained (r = 0.699) as well as the untrained (r = 0.680) listeners ($p < 0.001$ for both). This basically supports the results reported in previous studies, although the correlation is not quite as high. Looking at both groups of speakers separately, it emerges that the correlations between perceived age and chronological age are much higher for the smokers than for the non-smokers in both listener groups (0.892 and 0.572, respectively, for the expert group, and 0.903 and 0.518, respectively, for the student group). A possible explanation for this finding is that the degenerative process in the larynx which is induced by smoking may have served as a cue for listener judgements. In order to investigate this question further, the correlation between smokers' chronological ages and the number of years for which they had been smoking was calculated. The result is 0.907, which demonstrates that the older smokers in this study have also smoked for a longer period of time. This finding is confirmed by the calculation of a partial correlation between calendar age and perceived age in which the factor "smoking time" was factored out. In this case, correlations between chronological age and perceived age drop to 0.650 for the expert group and 0.577 for the student group. These results suggest that duration of smoking has a distinct influence on listener judgements and largely contributes to the higher correlation for smokers. This finding indirectly supports the results of a study by Ramig/Scherer/Titze [7] which is the only one in which listener judgements did not correspond to the chronological ages of speakers. The authors explain this result by the fact that their speakers were specifically chosen to have good physical condition and that "These age ratings may have been related to listeners' expectancies of age-related characteristics of voice" [p.6]. In other words: listeners judge biological age rather than chronological age, and as soon as these two do not run parallel in a speaker, listeners can no longer resort to stereotypes, and their estimates become less systematic.

No statistical difference with regard to the correlations was found between the performances of the two listener groups, i.e. the expert listeners did marginally but not significantly better than the naive listeners. The same applies to the overall correctness of the judgements. The average difference between perceived age and chronological age was 6.5 years for the non-experts and 5.9 for the expert group. Both groups were more correct about estimating smokers' ages than those of non-smokers, the experts erring by 4.7 and 7.1 years, the naive listeners by 4.7 and 8.4 years respectively. This is well within the margin which is usually given in a forensic report. A possible explanation for the lack of a difference between the groups is that the design of the listening experiment was very different from forensic real-world conditions in several respects. There is also the possibility that age estimation is a task which does not require phonetic, let alone forensic phonetic skills but is based instead on the everyday experience (or even: stereotypes) of any listener within a speech community.

Regression analyses were carried out with chronological age and perceived age as dependent variables in order to investigate which production parameter would best explain the results. The following predictors were examined: F0, jitter, shimmer, HNR, and syllable rate. Of these, only syllable rate and HNR proved to be significant predictors for both calendar age and perceived age (5%-level). This finding supports previous research [3, 15] where "rate of reading" was found to be among the most efficient predictors of perceived age. Here, it was also found to predict chronological age. HNR has not been studied as a predictor for perceived or calendar age, but the results obtained here are no surprise in view of the fact that HNR is a good indicator of various voice pathologies [16].

With regard to the questions asked at the outset of the study it can be concluded that smoking does in fact affect age estimation in that smokers are judged to be significantly older than non-smokers of the same age. Furthermore, listeners can be demonstrated to make systematic use of the variable "smoking time" in order to assess the chronological age of a speaker. Syllable rate and HNR constitute the only variables with significant value as predictors for age estimation.

The finding that perception seems to be geared to biological age rather than chronological age has implications for age estimation in the forensic domain, because there, obviously, the latter is called for. Thus, it is advisable to use utmost care and to indicate an age span or even only general descriptions like "very young", "middle-aged" etc. rather than attempting direct age estimates for forensic purposes

REFERENCES

[1] Shipp, Th. & Hollien, H. (1969), "Perception of the Aging Male Voice", *JSHR*, vol. 12, pp. 703-710.

[2] Neiman, G.S. & Applegate, J.A. (1990), "Accuracy of Listener Judgements of Perceived Age Relative to Chronological Age in Adults", *Folia Phoniatr*, vol. 42, pp. 327-330.

[3] Ryan, W.J. & Burk, K.W. (1974), "Perceptual and Acoustic Correlates of Aging in the Speech of Males", *J Comm Disord*, vol. 7, pp.181-192.

[4] Horii, Y. & Ryan, W.J. (1981), "Fundamental Frequency Characteristics and Perceived Age of Adult Male Speakers", *Folia Phoniatr*, vol. 33, pp. 227-233.

[5] Huntley, R., H. Hollien, Th. Shipp (1987); "Influences of Listener Characteristics on Perceived Age Estimations", *J. Voice*, vol.1, pp. 49-52.

[6] Hartmann, D. (1979), "The perceptual identity and characteristics of aging in normal male adult speakers", *J. Commun Disord*, vol. 12, pp. 53-61

[7] Ramig, L.A., R.C. Scherer, I.R. Titze (1985), "The Aging Voice". *Voice Foundation Symposium 'Care of the Professional Voice'*, Denver, June 1985. 10p.

[8] Ramig, L.A. & Ringel, R.L. (1983), "Effects of physiological aging on selected acoustic characteristics of voice", *JSHR*, vol. 26, pp. 22-30.

[9] Ringel, R.L. & Chodzko-Zajko, W.J. (1987), "Vocal Indices of Biological Age". *J Voice*, vol.1, pp. 31-37.

[10] L.J. Wallner (1954), "Smoker's Larynx". *Laryngoscope*, vol. 64, pp. 259-270.

[11]H.R. Gilbert & G.G. Weismer (1974), "The Effects of Smoking on the Speaking Fundamental Frequency of Adult Women". *J Psycholinguistic Research*, vol. 3, pp. 225-231.

[12]Sorensen / Y. Horii (1982), "Cigarette Smoking and Voice Fundamental Frequency". *Journal of Communication Disorders* 15, pp.135-144.

[13]C.H. Murphy / P. Doyle (1987), "The Effects of Cigarette Smoking on Voice-fundamental Frequency". *Otolaryngol Head Neck Surg*, vol. 97, pp. 376-380.

[14]Braun, A. (1994), "The influence of cigarette smoking on vocal parameters", *Proc. ESCA Workshop on Automatic Speaker Recognition, Identification, Verification*, Martigny, pp.161-164.

[15]Ptacek P.H. & Sander, E.K. (1966), "Age Recognition From Voice", *JSHR*, vol. 9, pp. 273-277.

[16]E. Yumoto / W.J. Gould / Th. Baer (1982), "Harmonics-to-noise ratio as an index of the degree of hoarseness". *JASA*, vol. 71, pp. 1544-1550.

IDENTIFICATION OF MALE AND FEMALE VOICE QUALITIES IN PRE-PUBESCENT CHILDREN

Kate Moore
University of Helsinki
Departments of English and Linguistics

ABSTRACT

This paper addresses the issue of gender-based language differences in children, specifically the ability of adult English and Finnish speakers to identify four to six-year-old Finnish children by voice samples alone. Four listener groups were presented two discrimination tests. All groups scored at or below chance level with shorter utterances (Test 1). The Finnish listeners' accuracy increased with longer utterances (Test 2), but even the best group mean scores (49-59% correct responses) were below levels reported in the literature for children speaking other languages.

INTRODUCTION

The differences between adult male and adult female voices are often easily recognized, due in part to vocal tract characteristics that develop after puberty. However, evidence collected from pre-pubescent children's speech suggests that certain aspects of voice quality may be learned--those which can not be attributed to larynx and vocal tract size--and that girls and boys can be thus differentiated by speech long before biological changes occur [1 & 2]. This raises important issues for prosodic research, namely, how children learn to manipulate gender-related speech cues in their formative language-learning years. Moreover, if these gender cues are learned, one would expect a difference in the way they are manifested in various language and cultures.

In 1992, Karlsson and Rothenberg reported that in their study of Swedish, Finnish, and English boys and girls, the gender of Finnish children was most difficult to recognize from voice samples.[3]. Whereas their listener groups provided correct answers 70% of the time for the other languages, they were only able to correctly identify Finnish boys and girls at slightly better than chance and at chance levels respectively. Karlsson and Rothenberg conclude that this is evidence that at least part of gender-specific speech is learned, and that the Finnish language, with its lack of gender cues in Finnish pronouns, may reflect and reinforce lack of identification of gender at early ages in that culture.

The objective of this study was to conduct a follow-up study of the Karlsson and Rothenberg findings, and to test the degree of accuracy by American and Finnish speakers in identifying the sex of a four- to six-year-old child by voice samples alone. Furthermore, another aim was to determine if there are prosodic features of the children's voices that correlate with these gender judgments.

METHOD AND MATERIALS

Test material

Recordings of 13 pre-school Finnish children were taken from a larger data base of 80 pre-adolescent Finnish children living in Helsinki. Voice samples of these four-, five- and six-year old boys and girls were isolated for two discrimination tests. The first test included ten utterances, ranging from two to eight syllables; the second test, containing ten utterances of eight syllables or longer, was devised to determine if utterance length affected the accuracy of the listener's judgments. The subjects were balanced for sex; five boys' and five girls' voices were included in each test.

Subjects

There were three phases of the investigation, each with a different set of subjects as listeners. The adult listener groups were not balanced for sex. In the first phase of the study, 31 adult listeners--18 Finnish-speakers and 13 American-English speakers--gave their judgments on the first test. Of these adult speakers, 11 Finnish-speakers and 11 American-English speakers participated in the second test. In the second phase of the study, twenty-five Finnish kindergarten teachers gave judgments for the two tests. And, finally, in the third phase, 55 Helsinki grade school children (25 girls and 30 boys) between the ages of nine and twelve years old, provided judgments for both tests.

Reliability test

Written transcripts of Test 1 and Test 2 were presented to five native Finnish speakers to evaluate possible gender-specific vocabulary or phrases that could bias judgments (four Finnish language editors--two male and two female--and one male speech therapist). No test item was considered to be gender-biased by the five Finnish speakers.

Equipment

The subjects were recorded using a Sony Professional WM-D3 tape recorder with a Sony ECM-144 microphone. The tapes prepared for the adult listener discrimination tests were copied using two Sony Professional WM-D3 tape recorders. The two tests were played to the adults on a Sony Walkman cassette player. The comprehensive school children listened to the tests in groups.

RESULTS

Gender judgments

All listener groups had difficulty differentiating Finnish girls' from Finnish boys' voices in the two tests. Results of the first phase of the study are presented in Table 1:

Table 1. Mean number of correct identifications (out of ten) with standard deviations indicated in parentheses:

Voice Discrimination Tests Phase I		
Listeners	Test 1	Test 2
Finnish adults	4.06 (1.47)	4.91 (1.68)
American adults	5.08 (1.27)	4.64 (1.77)

As is shown in the table above, American speakers were slightly better in identifying Finnish boys' and girls' voices on Test 1 than were Finnish listeners. The Americans scored a mean of 5.08 correct answers out of 10, (SD 1.27, Range 3-7), whereas the Finnish listeners' mean was 4.06 correct answers (SD 1.47, Range 2-7). While the English listeners' scores were at chance level, the Finnish speakers did not achieve a 50/50 split in judgments. For the American listener group, an increase in utterance length (Test 2) did not lead to better performances; on the contrary, they achieved a mean score of 4.64 (SD 1.77, Range 2-9) on Test 2, in comparison to Finnish listeners, whose Test 2 accuracy slightly increased over their Test 1 results, with a mean of 4.91 correct identifications (SD 1.68, Range 3-9) for Test 2. Figure 1 is an illustration of the Finnish adult speakers' success in identifying boys and girls voices in Test 2:

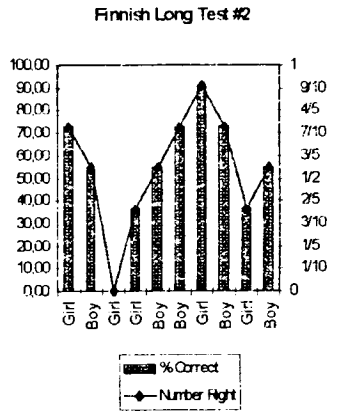


Figure 1: Percentage of correct identifications by Finnish speakers for Test 2. The horizontal axis represents (from left to right) the order and the gender of the ten voice samples. The vertical axis presents the percentage of correct identification for each test voice.

Phase II of this study dealt with testing Finnish-speaking kindergarten teachers. This group was selected because it was on the other end of the

scale of listeners than the American English listeners who did not understand Finnish. The kindergarten teachers communicated daily with Finnish preschoolers, and would thus be the most familiar with children's speech. Nevertheless, familiarity or expertise in relating to young Finnish children did not seem to be an advantage for Test 1. In fact, the kindergarten teachers had the fewest correct identifications of all listener groups on Test 1. Scores for these teachers are shown in Table 2:

Table 2. Finnish-speaking Kindergarten teachers' mean number of correct identifications (out of ten), with standard deviations indicated in parentheses:

Voice Discrimination Tests Phase II		
Listeners	Test 1	Test 2
Kindergarten teachers	3.2 (1.33)	5.92 (1.67)

The Finnish kindergarten teachers had the lowest mean score of all listener groups for Test 1 with a mean of 3.2 correct identifications (SD 1.33, Range 0-6). The teachers' accuracy improved significantly on Test 2, attaining a mean score of 5.92 correct responses (SD 1.67, Range 0-8), the highest mean number of correct identifications for all listener groups.

Another listener group selected for familiarity with children's voices was comprehensive school children between nine and twelve years old. Again, here the assumption was that children who are sure of their own gender identification, and who are in daily contact with children, would themselves be able to correctly identify voice qualities that differentiate boys from girls. At least for Test 1, this did not prove to be case, as is evident in Table 3:

Table 3. Finnish School children's mean number of correct identifications (out of ten), with standard deviations indicated in parentheses:

Voice Discrimination Tests Phase III		
Listeners	Test 1	Test 2
Finnish children	3.93 (1.48)	5.67 (1.50)

An examination of the data in Table 3 suggests that Finnish children of this age group had similar scores on the first test to Finnish adults, correctly identifying the gender of children 39 % of the time (Mean 3.93, SD 1.48, Range 1-7). The children's mean score improved, however, on Test 2 to 56% correct judgments (Mean 5.67, SD 1.50, Range 2-8) The listener groups of Phase II and III (kindergarten teachers and comprehensive school children) had a significant increase in correct identifications in Test 2. Thus, teachers and children found longer utterances easier to judge correctly, suggesting that Finnish intonational or prosodic patterns that are associated with gender require longer utterances than eight syllables for correct identifications. A summary of the four listener groups that participated in this study's three phases are shown in Table 4.

Table 4. Mean number of correct identifications (out of ten) with standard deviations indicated in parentheses:

Voice Discrimination Tests Phases I, II, and III		
Listeners	Test 1	Test 2
Finnish adults	4.06 (1.47)	4.91 (1.68)
American adults	5.08 (1.27)	4.60 (1.77)
Kinder-teachers	3.2 (1.33)	5.92 (1.67)
Finnish children	3.93 (1.48)	5.67 (1.50)

DISCUSSION

The data collected for this study suggests that pre-pubescent Finnish children's voices do not provide prosodic cues to identify gender, and is a

corroboration of Karlsson and Rothenberg's findings that Finnish children's voices carry less gender-specific information than do children's voices in other languages. However, there are so many variables yet unaccounted for concerning Finnish children's prosodic patterns, that stronger generalizations and conclusions about Finnish children's gender recognition and speech patterns would be at this point premature. In fact, very little is known about male versus female voices in Finnish.

Karlsson and Rothenberg discuss whether the construction of the particular language influences the child in learning to speak in a more gender-specific way. In the case of Finnish, the authors point to the lack of the he/she distinction of the third person pronoun in Finnish as a possible contributing factor. Here, the implication is that categories in language at the lexical level affect our classification and conceptualization of people, and in turn this classification is reflected in voice quality. While this should not be ruled out *a priori*, proving the connection is far from realized at this point. Since, as Karlsson and Rothenberg point out, there are many other ways to indicate gender, ranging from the frequent use of gender-specific personal names, dress, gestures, play interests, and other non-verbal behavior, we do know enough about the total picture of gender-specific awareness as reflected in speech, to make valid correlations.

As a native English speaker living in a Finnish-speaking environment and studying Finnish children's speech, I would like to make one general cultural observation that could affect these results, and which had not been mentioned in previous studies: the tendency of Finns to avoid interacting with strangers. Finns are often judged by members of other cultures to be withdrawn, shy, or reticent (the "silent Finn"), when in fact, this type of behavior in Finland is most predominant in public behavior amongst unfamiliar people. Testing young children thus presents a problem, as the voices reflect the interview situation: unknown adults question, observe and record, and the children, responding to the video-cameras and tape recorders and to the new people, interact shyly. I am currently trying to determine if there is a prosodic profile in Finnish that reflects this attitude, and that does not differentiate along gender lines. In fact, as partial verification of this, adult Finnish speakers were asked to describe the children's voices in Tests 1 and 2. The most frequent adjectives provided were "shy", "cautious", or "scared" for the first discrimination test, and this is precisely the test where all the Finnish listener groups had the most difficulty making gender identifications.

My conference paper will present more information on the prosodic variables in Finnish children's speech that could possibly skew the data to draw conclusions that gender information is not carried by Finnish children's voice quality.

ACKNOWLEDGMENTS

The voice discrimination data was collected by Kate Moore and Ulla Ström. Thanks to Jonita Mikkola for collecting Finnish children's responses in the lower comprehensive schools of Munkkivuori and Pihlajisto, and to Niina Kerppola, Eeva Naritz and Mari Saaralainen for providing the responses of the Helsinki kindergarten teachers.

REFERENCES

[1] Weinberg, B. and Bennett, S. (1971), "Speaker Sex Recognition of 5- and 6-Year-Old Children's Voices," *J. Acoust. Soc. Am.*, vol. 50, pp.1210-1213.

[2] Weinberg, B. and Bennett, S. (1979), "Sexual characteristics of preadolescent children's voices." *J. Acoust. Soc.Am.*, vol. 65, pp. 179-189.

[3] Karlsson, I. & Rothenberg, M. (1992), "Intercultural variations in gender-based language differences in young children," *STL-QPSR* No. 1, pp. 1-17.

THE PERCEPTION OF GENDER DIFFERENCES IN THE SPEECH OF 4½ - 5½ YEAR OLD CHILDREN

Moray J. Nairn
Queen Margaret College, Edinburgh, Scotland

ABSTRACT

Listeners judged the gender of different samples of the speech of 89 4½ - 5½ year old children. Sentences yielded the highest mean rate of identification (76.23%) and isolated vowels yielded the lowest (65.91%). Gender differences in listener response were also reported. The results reflect previously reported rates of accuracy from other countries and are interpreted in a theoretical framework which assumes that gender identification is a universal perceptual ability.

INTRODUCTION

The perceptual distinction between the voices of the adult male and female is in little doubt. When judges have been asked to classify subjects into male or female groups on the basis of voice alone, studies have reported success rates which approach 100% [1]. What is perhaps more surprising is the ability of judges to correctly classify prepubertal children at above chance levels of accuracy in the same way.

The issue of listeners' ability to identify the gender of prepubertal children from their speech has been addressed by several empirical studies [2,3,4,5,6,7]. A number of sample types have been employed in previous research, ranging from extracts of spontaneous speech (with selection criteria designed to minimise potential cues from utterance content) to isolated vowels spoken in a whisper. Among the experiments which used spontaneous speech samples, Weinberg and Bennett [2] demonstrated that judges were able to correctly identify the gender of 5 and 6 year old American children with a success rate of 74%. Meditch [5] achieved a similar level of accuracy using the speech of a small group of children as young as 3 years. Sentences have also been widely used as a sample type. Sachs et al [3] found a success rate of 81% with recorded sentences from a group of children with a rather large age range (4 to 14 years); the success rate dropped to around 66% when the samples consisted of isolated vowels only [4]. Günzburger et al's [7] results followed a similar pattern with 7 and 8 year olds, they found that the average recognition rate of 74% for sentences dropped to 55% when listening instead to isolated vowels. Bennett and Weinberg [6] had listeners judge the gender of 73 six and seven year old children using various types of speech sample. Overall, the identification rate was around 68%, with the success rates for the isolated vowels slightly lower than for the sentences. The identification rates of a quarter of the subjects were above 97% for the isolated vowels; however a significant number of the children were not judged consistently as either male or female. This emphasises the need to consider the distribution of not only mean results but also the identification scores of individual subjects. While the gender of many subjects may be very consistently identified, others may be ambiguous and some may even be consistently identified as having the opposite gender. With regard to possible listener bias in favour of one or the other gender, Meditch [5] found that listeners made significantly more 'male' guesses (i.e. gender judgements in favour of males). However Bennett and Weinberg [6] found no overall bias towards either gender.

The general conclusion that may be drawn from these studies is that perceptible gender differences do exist in the speech of many prepubertal children. Adult listeners are able to correctly identify gender with a success rate of somewhere around 70% from samples of normal speech. Even with samples of isolated vowels the success rate (at around 66%) is not much reduced. However, the identification rate is by no means as high as it is with adult subjects. This experiment is part of a larger study which attempts to address some of the issues surrounding gender identification using, for the first time, Scottish children who are, in addition, younger than many of the subjects used in previous studies.

METHOD

Eighty nine child subjects (46 boys and 43 girls), screened to exclude speech and hearing impairments, were recruited from two Edinburgh primary schools. English was the only language spoken in their homes and the accent used was the local Scottish accent of English. The methodological techniques and broad purpose of the study were explained to parents and children, but they were not informed of the study's relation to gender as it was felt that this might insert bias into the children's responses.

Three sample types were selected:-

Isolated vowels: The children sustained the vowels /i/, /a/ and /o/ at a comfortable pitch and level. These vowels were selected because it was felt that they represented enough of a contrast between high, low, front and back articulations to highlight any spectral gender difference which may exist.

Isolated sentences : The two sentences chosen were : *Rover is a big, brown dog* and *Rover has got a bone in his mouth*. These sentences were elicited by associating each with a picture and then showing the picture to the child who would then speak the sentence. This method avoided verbal or intonational cueing which often accompanies mirrored stimuli.

Spontaneous speech : A period of quasi-spontaneous speech was elicited from each child by means of the 'Bus Story' [8]. This procedure involves the researcher narrating a story which is illustrated by colour pictures viewed by both child and researcher. The child is then asked to re-tell the story from memory using the pictures. The resulting speech sample has the advantage of being of a very similar topic across all children but leaves the final choice of wording open to each individual child.

Recording equipment consisted of a Realistic® PZM Microphone and an AIWA HD-S1 Digital Audio Tape recorder fitted with a HDA-1 A/D converter. The data was recorded directly onto Maxell DM90 DAT tapes at a sampling rate of 48kHz with 16-bit Digital to Analogue (quantization) rate.

The master tapes were edited down to give three experimental tapes, one sample type per tape. On the vowel and sentence tapes, each child's vowels or sentences were kept together as a unit although the order of presentation of children on the tape was randomised.

The panel of listeners consisted of 8 male and 8 female subjects with an age range between 18 and 33 years. All were members of the academic staff and students of Queen Margaret College, Edinburgh and reported themselves to be native speakers of English and free of hearing disorders.

All of the adult subjects judged all of the children's speech samples. The adult judges recorded their responses on an answer sheet by circling 'M' or 'F' (for boy and girl respectively) in the appropriate box. Each judge received one answer sheet which was divided into three sections (one section for each type of speech sample - vowels, sentences, passage).

The judges made one response to each child's set of vowels, one response to each child's set of sentences and one response to each child's spontaneous speech. Therefore there were 3 gender responses to each child. As there were 89 children, this realised a total of 267 gender judgements per judge and a grand total of 4272 gender judgements in the whole experiment.

The purpose of the experiment was explained to the judges before beginning the task and they were each given written and oral instructions on how to record their responses. The adult subjects were seated in a quiet room and listened to each tape through headphones which connected directly to a Sony DTC-690 digital audio cassette player. Judges marked their estimate of the gender of each child on the response sheet. A short pause was included between tapes to allow the judges to rest. The gender of each judge was recorded along with their responses to permit analysis of the effect of the sex of the listener on gender identification.

RESULTS

Bias statistics revealed that one female judge was heavily biased to respond 'male' and so was removed from the analysis. Tables 1-3 show the proportion of correct gender responses made by remaining judges. The differences between these correct gender judgement rates varied between the three speech sample types. Paired two-samples t-tests were

carried out between the correct judgement rates of each speech sample type. Both the sentence and passage samples yielded significantly higher identification rates than the vowel sample ($p<0.05$). There was no significant difference between the identification rates of the sentence and passage samples. All identification rates are significantly above chance ($p<0.05$). In addition, the female judges were better at correctly identifying gender than the male judges ($p<0.001$) and girls were better identified than boys (girls - 76.37%; boys - 67.79%; $p<0.01$).

A total of 1859 (46.42%) 'male' responses and 2146 (53.58%) 'female' responses were made. A greater proportion of female to male responses was made in every condition (judges' sex × children's sex × sample type).

Table 1. Correct gender identification rates from VOWEL *samples*

	Male Judges	Female Judges	Average
Boys	59%	65%	61.96%
Girls	68%	72%	70.22%
Average	63.36%	68.82%	**66.08%**

Table 2. Correct gender identification rates from SENTENCE *samples*

	Male Judges	Female Judges	Average
Boys	71%	72%	71.49%
Girls	80%	82%	81.17%
Average	75.43%	77.22%	**76.33%**

Table 3. Correct gender identification rates from PASSAGE *samples*

	Male Judges	Female Judges	Average
Boys	67%	73%	69.93%
Girls	77%	78%	77.72 %
Average	71.81%	75.85%	**73.83%**

DISCUSSION

The results of this experiment go some way towards answering certain relevant questions and appear to raise some of their own. It is clear, for example, that the gender recognition rates found in this study conform closely to the levels of accuracy reported in previous research from around the world (U.S.A. - [2]; Sweden - [9]; etc.). This finding is interesting in that it suggests a universal ability on the part of the human listener to perceive acoustic gender differences. Existing experimental evidence seems to indicate that the differences in larynx size and vocal tract length which largely account for the adult vocal sex-difference are absent (or at least less influential) in pre-pubertal children. It has been suggested that processes of socialisation, which are responsible for emphasising existing physiological differences, and even separate gender dialects are in use. If this latter proposition, that gender is signalled vocally by learned speech characteristics, is indeed the case some obvious questions present themselves: are there universal aspects of the gender-marking and recognition process; at what age do these dialects emerge and what exactly is the method of gender-marking used by males and females?

Karlsson and Rothenberg [9] found that recognition rates of American and Chinese listeners were highly correlated with the rates displayed by Swedish and Finnish listeners and suggested that "at least some aspects of the gender differentiation in Swedish are not language-specific." [p14]. If the ability of listeners to determine gender in pre-pubertal children is a feature which is present universally, then reported differences in gender recognition rates may be accountable by differences in methods used by children from different cultures to signal gender. It seems unlikely that there are large-scale typological differences in the linguistic features employed across different languages to mark speaker gender, however one possibility is that there is a pool of features (phonetic, phonological, morphological etc.) from which languages are supplied - the precise features present in any given language depending on the configuration of the grammar (i.e. the parameters of Universal Grammar set during language acquisition). In Hebrew, for example, the correct conjugation of verbs depends on the sex of the speaker whilst in Japanese, boys and girls are encouraged to use different forms for the first-person pronoun although the spoken language is syntactically unmarked for gender. Also, listeners find it harder to identify Finnish boys from girls than Swedish boys from girls. The suggestion is that as there is no grammatical gender in Finnish it may be that young children acquiring the language "are less aware of sex-differences...than are Swedish children and therefore less motivated to develop speech patterns [based on] speaker sex." [9]

The age at which these gender dialects start to emerge cannot be answered by the present study, however in the light of these results using 4½-5½ year old children it is clear that there is some information relating to the child's gender present in the voice by the fifth year of life. To ascertain what form this gender information might take in any given language would require an in depth acoustic study of the voice. Such an investigation is currently being carried out by the author using the same children from this study. Previous research has tended to be inconclusive, however the weight of opinion seems to be that fundamental frequency, unlike the situation in adults, is not a major factor in gender marking in English.

Turning now to the differences in listener accuracy between the three different speech sample types, the fact that the passage and sentence conditions yielded significantly higher rates than the vowel condition may be attributed either to the fact that there was more phonetic information present in the former or to the fact that there was a qualitatively different type of information available (intonation, formant transitions, etc.). What is more surprising is that the sentence sample gave a higher correct recognition rate than the spontaneous speech sample (although not significantly so). One possible interpretation is that the sentence sample involved the children speaking exactly the same words every time whereas the passage sample only controlled for the topic of conversation but allowed the children the freedom of choice of words. Therefore, strictly speaking, listeners were not comparing like with like in the passage condition, they were forced to extrapolate each child's speech patterns to a certain extent in order to make mental comparisons. This extra step in the recognition process might adversely influence levels of accuracy relative to the sentence sample.

In summary, listeners were highly successful in determining the gender of 4½-5½ year old Scottish children. The identification rates achieved were very similar to previous comparable studies. The result that girls were better perceived than boys, and female judges were better then male judges appears to contradict some of the previous findings. Further work is needed to supply the missing answers.

REFERENCES

[1] Coleman R.O. (1971) The perception of maleness and femaleness in the voice and its relationship to vowel formant frequencies, *Proceedings of the 7th Int. Congress of Phonetic Sciences.*

[2] Weinberg B and Bennett S.(1971) Speaker sex recognition of 5 and 6 year old children's speech, *Journal of the Acoustical Society of America* **50**:1210-3.

[3] Sachs J, Lieberman P, Erickson D. (1973) Anatomical and cultural determinants of male and female speech. in *Language attitudes: current trends and prospects.* ; Shuy R, Fasold R, editors. Washington D.C. Georgetown,

[4] Sachs J. (1975) Cues to the identification of sex in children's speech, in *Language and sex: difference and dominance.* ; Thorne B, Henley N, editors, Rowley, Mass.

[5] Meditch A.(1975)The development of sex-specific speech patterns in young children, *Anthropological Linguistics* **17**:421-33.

[6] Bennett S and Weinberg B.(1979) Sexual characteristics of preadolescent children's voices, *Journal of the Acoustical Society of America* **65**:179-89.

[7] Günzburger D, Bresser A, and Ter Keurs, (1987) Voice identification of prepubertal boys and girls by normally sighted and visually handicapped subjects, *Language and Speech* **30**(1):47-58.

[8] Renfrew, C. (1969) *The Bus Story - A Test of Continuous Speech,* C.E. Renfrew, Old Headington, Oxford.

[9] Karlsson, I. and Rothenberg, M. (1992) Inter-cultural variations in gender-based language differences in young children, *Quart. Prog. and Status Report 1/1992. Speech Transmission Lab. Institute of Technology, Stockholm.* pp.1-17.

AN OPTIMISED PARALLEL FORMANT SPEECH SYNTHESIZER

J R Andrews and K M Curtis
Department of Electrical and Electronic Engineering, University of Nottingham, UK

ABSTRACT

The problem that is addressed in this paper is the production of high quality natural-sounding multi-gender speech for different languages. The development of the Nottingham Parallel Seven Formant Speech Synthesizer is described, presenting the reasons for the choice of structure, component parts and demisyllable synthesis units. Results of the synthesis of German speech sounds and an utterance are given.

HIGH QUALITY SPEECH

Speech is regarded as the most natural form of communication known to man. The human-computer interface would be drastically revolutionised if speech was the medium chosen for the transfer of information. For example, the computer keyboard could be replaced by commands being issued and received by speech. This requires high quality speech synthesis and recognition systems, which are reliable, robust and can produce naturally sounding speech and recognise normal human speech.

The acoustic theory of speech production [1] provides a sound basis from which this problem can be tackled, in the frequency domain using formants.

In order to produce high quality speech, a flexible well-designed synthesis model must be sought. The most significant models are the five formant cascade and parallel KLSYN88 [2], developed by Klatt *et al* and the five formant parallel synthesizer [3] of Holmes *et al*. Both produce male and female speech and have been used in commercially. The CNET PSOLA synthesis system [4] uses time waveforms/windowing techniques and produces natural-sounding speech.

Diphones and demisyllables are the most widely used units for synthesis. PSOLA uses diphones for its synthesis unit. The German "HADIFIX" synthesis system [5] is based on a combination of diphones and demisyllables. These units have not yet been applied to a high quality formant speech synthesizer.

The synthesizer must have the ability to produce sounds present in most european languages and be flexible enough to produce male, female and child's speech. Therefore, it must incorporate a high level of control and be driven by highly accurate formant and excitation source data.

The synthesizer presented here forms part of a complete synthesis system [6]. It is a highly-programmable parallel formant synthesizer designed to synthesize multi-gender speech sounds of a high quality. A parallel structure provides the most flexible structure for synthesizing all types of sounds. The mapping of the synthesizer onto a network of processors is performed using a novel data/control flow methodology, thus enabling maximised processor utilisation and real-time performance. The synthesizer can model up to seven formants, has a maximum bandwidth of approximately 10 kHz and can use natural voiced sources.

NOTTINGHAM SYNTHESIZER

The synthesizer developed here consists of seven resonators connected in parallel. A novel parallel configuration was chosen for modelling the vocal tract, which is the most flexible design for synthesising all types of sounds. The choice of a solely parallel structure also reduces the complexity of the design. It has been shown [2] that a combination of a cascade and a parallel structure can be used for the production of all sounds.

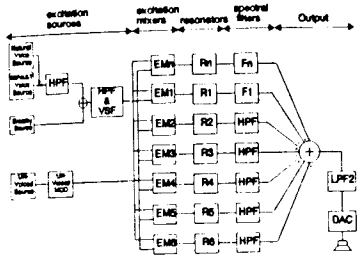


Figure 1. The Nottingham Parallel Seven Formant Speech Synthesizer.

This is because each resonator structure is suited to the production of a certain set of sounds. However, with accurate control of the resonator parameters, high quality sounds can be produced by a parallel branch only. This level of control is achieved by a constant update of all the synthesizer's parameters from sample to sample.

The aims of the design were to provide accurate control of all synthesizer parameters and to keep all specifications as dynamic as possible, in order to provide a highly adaptable synthesizer and thus to produce high quality natural sounding speech. Figure 1 shows the final components of the Nottingham Parallel Seven Formant Speech Synthesizer. The three main components are the excitation sources, the bank of resonators (representing vocal tract transfer function), and the output filter. In developing the final structure shown below other synthesizer designs were researched and tested [7].

The most significant parts of the synthesizer design are the default voice source, the novel use of individual excitation mixers and resonators. The choice of demisyllable synthesis unit will also be described.

Default Voice Source

The default model of the voiced source in the synthesizer is a complex design incorporating a number of important spectral features. The equation used in this model was based on the voiced source model employed in the KLSYN88 formant synthesizer [2].

The voiced source contains provision for dynamic variation of; amplitude of voicing (AV in dB); fundamental frequency (F0 in Hertz); open quotient, OQ in % of the total pitch period T0, this is the ratio of open glottis time to its pitch period.; period-to-period flutter (or jitter), FL in % of the fundamental frequency F0 value; and Shimmer SH, the period-to-period variation in the amplitude of voicing, A.

The first three parameters are essential for a high quality voice and the last two provide a degree of naturalness. The basic equation which describes the one pitch period of the waveform is of the form shown below.

$$v(nT) = a\left((nT)^2 - \frac{(nT)^3}{(\frac{OQ}{100}) * T0}\right) - DC; \quad 0 \le nT < (\frac{OQ}{100})T0$$

$$= -DC \qquad ;(\frac{OQ}{100})T0 \le nT < T0$$

where a and DC are constants determined by AV, OQ and T0.

Excitation Mixers

Figure 1 depicts each resonator branch as a combination of three distinct parts, the excitation mixer (EM_n, EM_1...EM_6), a formant resonator (R_n, R1,..., R6) and a spectral weighting filter (FN, F1 or HPF). Each resonator branch provides one formant resonance. R_n provides the low frequency component and nasal formant. Resonators R1 to R6 can each provide a formant, with R6 providing the highest formant (up to the bandwidth of the synthesizer's current implementation).

The excitation mixer combines both the voiced and unvoiced excitation for the resonator in its branch. It allows for individual dynamic control of both types of excitation for each resonator. This is unlike other common fixed synthesizers (Holmes and Klatt) and provides an extra degree of freedom in the amount of excitation in each formant.

The excitation mixer requires two parameters A_{Vi} (dB) and A_{Ni} (dB) per speech segment, for specifying voiced and unvoiced source amplitudes. The output of the EM module is defined below.

$$EM_i(nT) = a'\left\{A_{Vi}(S_{gi}(nT)) + A_{Ni}(S_{ni}(nT))\right\}$$

where $EM_i(n)$, $S_{gi}(n)$ and $S_{ni}(n)$ are the outputs of the excitation mixer, voiced source branch and unvoiced source branch respectively. a' is a factor derived from the formant frequency and bandwidth in the resonator equation.

Resonator

The synthesizer utilises a 2nd order resonator to generate the formant. The basic resonator (Rn, R1-R6) requires three parameters to specify its acoustic properties, the formant frequency, F_i Hz, amplitude and bandwidth, BW_i Hz. The formant amplitude is controlled in the excitation mixer. The resonator only requires the F_i and BW_i.

The output of the resonator is:

$$R_i(nT) = EM_i(nT) + B_i\{EM_i(nT-T)\} + C_i\{EM_i(nT-2T)\}$$

where n is the sample number, $EM_i(n)$ $EM_i(n-1)$ and $EM_i(n-2)$ represent the excitation's output at sample (n), (n-1) and (n-2). The coefficients B_i and C_i are calculated directly from F_i and BW_i.

Demisyllable Synthesis Units

The demisyllable was chosen as the speech synthesis unit to ensure synthesis with a very high degree of naturalness. A systematic approach [6] is used to define a demisyllable database. Synthesis is based on three types of demisyllables; (consonant-vowel cluster) and final demisyllables (vowel-consonant cluster) and vowel-vowel clusters. These were required to adequately cover all the coarticulation effects present in european languages. The approach adopted ensures that high quality and uniformity is maintained in the synthesis elements. The synthesis units are formant coded using the analysis system [6].

The temporal minima of the coarticulation effects coincide with the boundary of the demisyllable, therefore, only a small number of relatively simple concatenation rules are required. Each demisyllable is regarded as a number of formant-coded segments. The segments are concatenated by simple linear interpolation.

SYNTHESIZER IMPLEMENTATION

The implementation aspect of the synthesizer design is just as important as the structure of the synthesizer itself. Not only is a high quality natural-sounding speech synthesizer required, but the ability to produce sounds in real-time and as efficiently as possible is also needed.

The choice of both synthesizer design and implementation lead to the choice of the transputer parallel processor and processing language Occam [7].

As the synthesizer design is a complex algorithmic problem, a novel technique [8] of identifying and modelling both the flows of data, and flow of control variables was developed.

The synthesizer solution lead to the construction of a multi-transputer architecture to give a highly versatile and novel speech synthesizer, capable of the easy introduction of further formant generators. see figure 2 below.

The flow of control parameters is a pipeline originating in the "monitor" transputer which calculates the parameters from user input data.

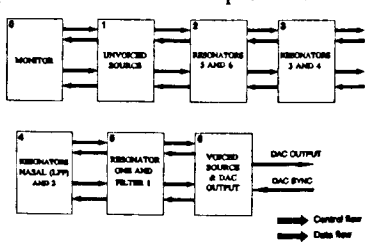


Figure 2. Synthesizer implementation.

The results of the synthesis of an utterance spoken by a male German are now presented.

RESULTS

An utterance was selected for synthesis using the German database of demisyllables. The utterance was "Das Stoßgebëf ist ohne Sinn". The utterance consists of the following demisyllables; /da/ /as/ /_Sto:/ /o:s/ /g@/ /_@/ /@_/ /bE:/ /E:f/ /_I/ /Ist/ /_o:/ /o:_/ /n@/ @_/ /zI/ /In/.

All the demisyllables underwent formant analysis as described in reference [6]. The each demisyllable is described by a number of sets of formant parameters.

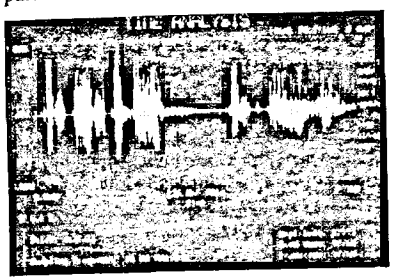


Figure 3. Time waveform of natural sentence /das Sto:sg@bE:f Ist o:n@ zIn/.

CONCLUSIONS

The synthesized utterance contained all types of speech sound. The waveform shows the advantage of demisyllables as units for speech synthesis. The demisyllables were synthesized from sets of formant coded segments and linear interpolation was carried out between the sets of parameters. This produced smooth transitions from segment to segment. However, there were some problems encountered in the production of the synthesized waveform. This was in the control of the unvoiced source model. As can be seen from the synthesized waveform, the amplitude settings for the unvoiced/mixed voiced segments were too high.

Work is under way in developing a more precise strategy for the control of the parameters fed to the synthesizer. Further work is also being carried out to develop a more natural sounding voice source which can be easily implemented into the synthesizer's structure. This should allow the modelling of different voices, accents, and manners of speaking. The synthesizer has shown greatest flexibility in the description of demisyllables.

REFERENCES

[1]Fant G. *Speech Sounds and Features*. MIT Press 1973.

[3]Klatt D H and Klatt L C, "Analysis, Synthesis, and Perception of Voice Quality Variations among Female and Male Talkers", J.Acoust.Soc.Am Vol 87 No.2, Feb 1990, pp 820-856.

[4]Holmes J N, "A Parallel formant synthesizer for machine voice output", in *Computer Speech Processing*, Prentice Hall UK, 1985 pp 163-187.

[5]Hamon C, Moulines & Charpentier F, "A Diphone Synthesis System Based on Time-Domain Prosodic Modifications of Speech", ICASSP'89, pp 8-11.

[6]Portele T, et al, "Hadifix: A Speech synthesis system for German", ICSLP'92, Banff, pp 1227-30.

[7]Andrews J R and Curtis K M, "A Comprehensive Analysis and Synthesis System, and Synthesis Methodology for the Production of High Quality Speech", International Symposium on Speech, Image Processing and Neural Networks, Hong Kong, April 1994, pp 591-594.

[9]Curtis K M, Asher G M, Pack S E & Andrews J, "A Highly Programmable Formant Speech Synthesizer Utilising Parallel Processors", ICSLP'90, Kobe, Japan, 19.14.1-19.14.4.

[11]Curtis K M and Andrews J R, "Control Flow: A Technique for Optimising Processing Architectures", Proc of 9th IASTED, Austria'91, p322-5.

A Finite Automaton Translator of Text to Phonemes for the Portuguese Language

G. L. de Campos and D. T. Chbane

Polytechnic School of the University of São Paulo, São Paulo, Brazil

ABSTRACT

A system for text-to-speech conversion for the Portuguese language is described. The system includes a 320,000 words vocabulary, encompassing all valid Portuguese words. Using information in this vocabulary and a set of rules, the system is able to translate a general text in the corresponding phoneme sequence.

INTRODUCTION

The use of speech as a form of computer output has growing importance as part of user-friendly man-machine interfaces and for specific applications. However, its usefulness has been impaired by the lack of synthesis systems able to generate speech from an unrestricted text, providing a natural and pleasant sounding.

The process of voice synthesis from text is comprised of the following steps:

1 - Text to phonemes translation. In this context, the word phoneme comprises phonemes and diphones, including their durations. Diphones are peculiar sequences of phonemes that, due to coarticulation effects, are very difficult to simulate precisely when the phonemes are considered alone. It happens more with plosive-vowel sequences. This translation is most a problem of lexical analysis, although some rudimentary syntactical analysis is required, since a relatively large class of words has different pronunciation when they belong to different grammatical classes. This happens more frequently with verbs and substantives (go[o]sto, substantive, and go[ ]sto, verb, for instance).

2 - Pitch contour determination. Speech with constant pitch, resulting in a robot like voice, is very irritating and a serious system must generate a pitch contour such that each phrase sounds as naturally as possible. Since the pitch conveys non-syntactical and emotional information, it is not generally possible to produce really natural sounding utterances, but acceptable pitch contour can be generated based on the general grammatical structure of a phrase. The result of this phase can further change the duration of the phonemes already determined in phase 1. Much research is still needed, but present results are encouraging.

3 - Speech synthesis. This phase is being implemented using a technic somewhat similar of what is done on the successful CELP vocoders. The results will be published elsewhere. [1]

The synthesis process is shown in figure 1.

This paper concentrates on algorithms and techniques for the step 1.

Through this document, symbols between brackets are phonemes represented according to the IPA (International Phonetic Alphabet).

GENERAL CONSIDERATIONS

Converting an unrestricted text into speech requires the capacity of converting any sentence in the language; therefore, every word of the language must be recognized. This requires the availability either of a general algorithm able to convert any word into a sequence of phonemes, or a vocabulary comprising all the words of a language, or some suitable combination.

Practice had shown that the first approach is impossible for most languages, including English and Portuguese. Most systems use the third approach, based on the assumption that a large vocabulary is difficult to generate, to maintain and, above all, to access.

The present system uses the third approach, referring all words to a vocabulary. In Portuguese, the problem of conversion from text to phonemes is simpler than in English, since most letters can be unambiguously converted to a phoneme, but there are exceptions that requires using a dictionary, since these letters can be mapped to a small (2 to 4) set of phonemes, in most cases by historical reasons. Once a dictionary is required, it is worthwhile to consider the problems usually associated with the use of a full vocabulary.

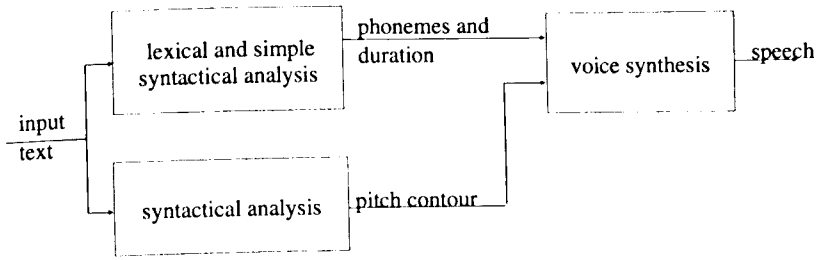


Figure 1 - Modules of the text-to speech system

Contrary to the general opinion, it is easier to generate a full vocabulary, provided there is a standard one. All that is required is to transcribe it to a machine readable form, either with OCR equipment or by hand transcription; in any case, a mechanical operation. To use only the exceptions requires the determination of the exceptional cases, which is a tedious but intellectual operation. Maintenance and access issues are related: if there is an automatic procedure that converts the transcription of the vocabulary to a compact representation, and fast access procedures, these issues can not be considered a problem. Such procedures exists, and consist of converting the vocabulary to a (possibly augmented) finite state automaton, [3], [5]; in the latter, a representation was obtained with a compactation in the range of then 1 bit per word, and very fast accesses.

Taking this in consideration, it was decided to use a full representation of the vocabulary. Section 3 details the characteristics and implementation of the vocabulary.

The vocabulary is not enough, however, since the uttering of a word frequently depends on context. There are three conceptually different dependency categories. The first is essentially grammatical, and has to do with individual words; it occurs when the same word may represent different grammatical entities, like being a substantive or a verb. It can be somewhat easily handled by a simple grammatical analyzer.

The other categories deal with the sentence level, and determine the pitch contour of the utterance. The second can be related with the structure of the sentence, and can roughly be determined by a more elaborate syntactical analysis. The third category depends on semantics: a speaker can change the meaning of an utterance by changing the stress and intonation of certain words; it is impossible to determine differences in this category using only the text.

Vocabulary structure

The vocabulary was produced by transcription of the "Vocabulário Ortográfico da Língua Portuguesa", which is the official list of the Brazilian Portuguese language. It contains the word and grammatical category (categories if the word belongs to more then one). For some words, it contains also phonetic information about some vowels that can be open or closed. As usually happens in dictionaries produced by hand, there is a lot of difficulty in using these information, since there are many deviations from the standard syntax, specially lack of punctuation and comments in unexpected places. Overall, it contains about 320.000 entries, including only the infinitive of 22,600 verbs, that in Portuguese change with tense and person. If expanded, this would add more than a million entries; fortunately, most verbs follow a regular pattern of change, and this is not necessary. This verbs can be automatically expanded during their inclusion in the automaton representing the vocabulary.

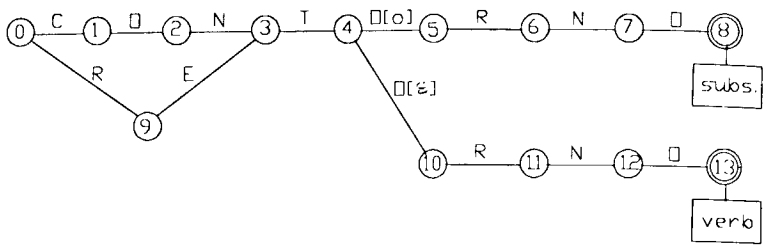


Figure 2 - Example of an augmented automaton This automaton recognizes the words reto[o]rno (substantive), reto[ ]rno (verb), conto[o]rno (substantive), conto[ ]rno (verb). See text for details.

Internally, the vocabulary will be represented by an augmented finite state automaton. The language accepted by the automaton will be the set of all words present in the Vocabulary, plus expanded verbal forms. Figure 3 presents an example automaton. In this figure, ignoring for now the boxes, the state numbered 0 is the initial state and the states marked with a double circle are possible final states. This representation is quite efficient, since it offers an implicit way of sharing prefixes and suffixes [5]. It also offers the possibility of representing new words produced by proper derivation. An example automaton is shown in figure 2.

The automaton is augmented by grammatical and phonetic information. This allows the conversion of the input string as the automaton is making transitions (possibly non-deterministically). When a letter may have multiple phonetic transcriptions, the incoming symbol is expanded to include all the phonetic alternatives, denoted by the phonetic symbol appended to the letter. In this case, the automaton follows all the alternatives simultaneously. Each final state is augmented by the grammatical category of the recognized word. In figure 2, the boxes marked "verb" and "subs." (substantive) show examples of this situation.

state	attribute	link
0		
1		
2		
3		
4		
5		
6		
7		
8	substantive, masculine, singular	
9		
10		
11		
12		
13	verb, present, singular, 1st person	

letter	state
C	1
O[o]	2
N	3
T	4
O[o]	5
R	6
N	7
O[o]	8
E[e]	3
R	11
N	12
O[o]	13

letter	state
R	9
O[ɛ]	10

Figure 3 - Automaton corresponding to the words in figure 2

When a letter may represent more than one phoneme, the preceding state will have as many transitions as the number of possible phonemes. In the example of figure 2, state 4 has two transitions associated with the letter O; one corresponds to the phoneme [o], and the other to [ ]. The other transitions with the letter O are unique, since all of then corresponds to the phoneme [o].

In the internal data structure, every state is represented by a tuple (*state, attribute, pointer*). The field *state* contains the number of the state; in the final minimized version, it is determined by the relative position of the other fields, and does not exist explicitly. The *attribute* field is also one byte long, and contains grammatical or other complementary information. The field *pointer* is three byte long and points to a transition list, formed by pairs (*letter, state*) indicating the next state corresponding to each incoming letter. In this table, the *letter* field is one byte long, and contains the letter, augmented by the phonetic variant, if the letter may correspond to several phonemes (in Portuguese, there are at most four variants; this occurs with the letter X). Figure 3 shows the structure of the automaton corresponding to example in the figure 2.

The algorithm mounting the automaton makes initially a trie (a special form of a tree of letters, described in [2]; the name trie comes from reTRIEval). After constructing it, formal methods are applied for minimizing the resulting automaton, by the reduction of equivalent states.

Words are included in the automaton directly from the transcription of the "Vocabulário Ortográfico", except in the case of verbs, that are automatically conjugated according to the usual rules of the Portuguese grammar. The representation of letters is always phonetic, and a word appears as many times as its possible variants, either in phonetic representation or grammatical category. The phonetic translation of each word is determined either from information contained in the vocabulary or by the application of a set of rules designed to overcome some difficult aspects of the Portuguese language phonetics.

Conclusions

This paper presents a system for converting written text to phonemes in Portuguese. Although the conversion is simpler than in some other languages as English, it shows some complexity and a vocabulary is essential for a large number of words. The system will be used as part of a text-to-speech system under development.

It is important to note that the automaton, augmented by the phonetic and grammatical information described in this paper, can be used also for voice recognition, since the phonetic information is an integral part of the state changing mechanism.

The transcription of the vocabulary and its codification is an important part of the work. This vocabulary will be put in the public domain once certain legal aspects are clarified.

REFERENCES

[1] Campos, G. L. and Gouveia, E. B. *Voice Synthesis by CELP Technology*, to be published.

[2] Fredkin, E., *Trie Memory*, Comm. of the ACM, **3(9)**:490-500, Sept 1960.

[3] Gross, M., *The Use of Finite Automata in the Lexical Representation of Natural Language*, in M. Gross and D. Perrin, editors, Electronic Dictionaries and Automata in Computational Linguistics, 34-50, Springer-Verlag, Berlin, 1989. Lecture Notes in Computer Science, vol 377.

[4] Hertz, S. R. et al, *The Delta Rule Development System for Speech Synthesis from Text*. Proc. of the IEEE, **73(3)**:737-793, Sept. 1987

[5] Lucchesi, C. L., and Kowaltowsky, T., *Applications of Finite Automata Representing Large Vocabularies*, Software - Practice and Experience, **33(1)**, Jan 1993.

NEURAL NETWORK SOLUTIONS FOR IMPROVING ENGLISH TEXT TO SPEECH TRANSCRIPTION

P.R. Gubbins and K.M. Curtis
Parallel Processing Specialist Group, Department of Electrical and Electronic Engineering, University of Nottingham, England

ABSTRACT

This paper describes the application of a novel hybrid architecture to the text to speech mapping problem. The hybrid architecture combines the two previously prominent techniques used in this field: a rule base and a neural network. It aims to improve on the success rate of the neural network solution whilst reducing processing and greatly speeding up learning rate. A further neural network application which evaluates the rule base part of the hybrid system is also discussed.

INTRODUCTION

The first stage in the high quality synthesis of speech from unrestricted text is the process of converting from a code representing language graphically to one which represents the sounds that make up the signals we can decode through our auditory system.

The fact that we are dealing with unrestricted text is important when considering the techniques we are to use. A synthesis system that only needs to reproduce words from a limited vocabulary will be able to store sound code for whole or large portions of words. As the required vocabulary of a system increases however, the storage capacity and processing for the retrieval of such large blocks of sound code becomes impractically large. The longer and hence more complex the unit of sound that is stored the greater the number that is needed to be stored to produce all combinations of speech sounds in a language. For a system working from an unrestricted vocabulary it is only practical to store the most fundmental units of sound code, i.e. phonemes. The transcription from text to phonemes is therefore the problem we are faced with.

The historical derivation of the English Language makes it a particularly difficult case for transcription from text to phonetic code. English is derived from several widely variant sources with words and therefore sounds coming from Germanic, Latin, and Scandinavian based languages. The problem is compounded by the fact that these sounds must be encoded graphically by a relatively sparse alphabet, lacking accents, which in other languages multiply the number of vowel sounds that can be represented. The result of this feature of english is that in many cases a relatively wide context of letters has to be considered around a target letter in order to correctly predict the resultant phoneme.

APPROACHES TO TRANSCRIPTION

Previous research has concentrated on two main methods for transcription of text to phonetic code. The first of these is a rule based algorithm. In an attempt to handle the complexity in English described above this approach has led to a three layered algorithm. The alternative approach that has been widely researched, notably with the "NETtalk" system[1], is the use of a neural network to learn and then repeat the mapping from graphical to phonetic code.

Rule based algorithm

Before other techniques were developed systems for converting from graphemes to phonemes were based on a set of rules that indicated which particular phoneme a letter (grapheme) should represent depending on the context of the other letters surrounding it. Much of the later work on rule based techniques has been based on the set of rules developped by Elovitz et al [2]. A three layered rule based system was developed at Nottingham University [3] as an extension to the Elovitz letter to sound rules. This system used three parallel, hirearchical algorithms to analyse the text input and produce phonemic output. The algorithms are, in order of precedence, an exceptions dictionary, a morphological decomposition algorithm and letter to sound rules. Morphological decomposition involves recognising common letter groups (morphs) at the beginning and end of words and applying phonemic and stress information to the pronounciation of the whole word where it can be deduced from the presence of a particular morph. Examples of morphs are the premorph "con-" and the postmorph "-ation".

Neural Network Solution

As research into neural networks developed the text to phoneme transcription task was seen to be a suitable task to apply a neural net. solution to. Artificial neural networks (ANN's) are good at picking up statistical regularities but are not impeded by occaisional inconsistencies. The NETtalk system was developed [1] and has become the basis for further research.

NETtalk is a software simulation of a three layered structure, the three layers being input, hidden and output. The network has as its input a seven letter window with the central letter being the target and the remaining six being the context. This translates to 203 (29x7) binary 'nodes' in the input layer where each of the 7 input letters is set as one of the 26 letters of the alphabet or one of 3 punctuation symbols. The output layer consists of 26 binary units representing 26 articulatory features. Each phoneme is produced by a unique combination of these articulatory features. Between these two layers is the hidden layer of 120 units. Each input unit is connected to every hidden unit and each hidden unit in turn to every output unit, giving a total of 27,000 connections. Each connection has a weight value associated with it and each node a transfer function which evaluates the sum of its weighted inputs and produces a resulting output. The network learns by being presented with an input and also the correct output. The weights in the network are adjusted by means of a back propogation algorithm for each presentaion of an input and correct output pattern. Eventually, after very many presentations (e.g. 20,000 words) the weights in the network will reach stable values and the network will be able to generalise i.e. produce an output pattern for any given input pattern.

THE NEED FOR AN IMPROVED SOLUTION

Both of the above approaches, rule base and neural net, have their limitations. In the case of the rule base the problem is the need to produce a set of rules that are correct in all cases for a very complex language or to compensate for deficiencies in the rules by means of the exceptions dictionary. In the latter case the exceptions dictionary tends to become prohibitively large. In the case of the neural network solution the problem is in controlling the learning such that apparent anomalies in pronounciation but nevertheless valid cases are catered for. Possible solutions are to increase the size of the input layer to give wider context or to increase the size of the hidden layer or indeed the number of hidden layers. These solutions lead to an increase in the already very large processing overhead. A further consideration is that although the number of connections provides no difficulty in software implementation it renders a hardware implementation completely impractical.

HYBRID ARCHITECTURE SOLUTION

Research at the University of Nottingham has shown that the use of a novel hybrid architecture can provide a solution to mapping problems with the potential to improve results whilst reducing the processing overheads [4],[5].

The hybrid architecture combines a rule base and a neural network in a parallel sturcture. Figure 1 shows a block diagram of the structure. Rather that the neural network learning the correct output pattern for a given input pattern it learns the difference between the output given by the rule base and the correct output pattern. In effect the neural network learns to correct the output of the rule base. When generalising the two units work concurrently, each being presented simultaneously with the same input vector.

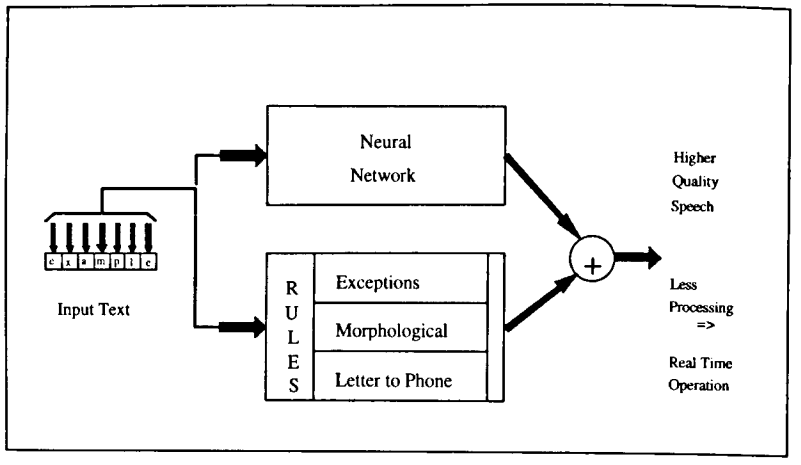


Figure 1. Block Diagram of The Hybrid Text to Phoneme Transcription Architecture

The system output is obtained by summing the output of the rule base and the ANN. The network complements the rule base in mapping the areas that it doesn't model. This can be said to model the way that we learn the relationship ourselves in that we are taught some rules when we initially learn to read, but do not consciously refer to them as our skill progresses, rather our brain develops subconscious patterns relating letter combinations to sounds. The contribution of the neural network allows the rule base to be very much cut down, therefore overcoming the need to specify an exhaustive set of rules. Similarly, due to the presence of the rule base the learning process of the neural net is greatly accelerated when compared with a purely network solution. As well as reducing the learning overhead this leads to the possibility of reducing the size of the network and hence the number of connections. As mentioned above this is a desireable situation when considering a hardware implementation of a text to speech system.

SELF ORGANISING NEURAL NETWORK USED TO EVALUTE RULE BASE REDUCTION

One of the major problems to be faced when investigating the use of the hybrid architecture for the text to phoneme transription task is deciding on the most appropriate way to cut down the rule base so as to maximise the performance of the hybrid system. This leads to a secondary use of neural networks in this research. A self-organising network, using a radial basis function algorithm, is used to examine the output of the rule base system. Whereas supervised networks such as NETtalk learn to map inputs to outputs by means of many correct examples of the mapping, unsupervised or self organising networks simply group inputs within the space provided by the network according to similar properties[6],[7]. Analysis of the way in which the output phonemes of the hybrid system group themselves in such a network will enable the selection of the best way to cut down the original rule base to give the optimum results in the combined system. It has been found that certain mapping phenomena, such as the difference between initial 's' and 'z' sounds, are differentiated only very late in the learning process of the neural net alone solution [8]. With the knowledge of such effects the choice of which rules are to be retained will enable these phenomena to be counteracted.

CONCLUSIONS

Extensive investigations are currently being carried out into the effectiveness of the hybrid architecture with many different reduced rule base and reduced size neural network combinations. Initial results indicate that a system with a very greatly reduced rule base and the original sized network can acheive correct transcription rates as good as NETtalk (91% after training on a set of 20,000 words) after substantially fewer training iterations (circa 50%). It is envisaged that once results of the self organising network have been applied to the system the success rate will increase. Trials involving reducing the network size have yet to be initiated.

The application of either a rule based or a neural network based solution to text to speech transcription requires a great deal of initial work to be carried out either in the form of producing a set of rules or in preparing an extensive training database of correct phoneme transcriptions. The use of a hybrid sytem as described potentially reduces these tasks as only simplified rules are required and training data requirements are reduced. A further feature of the english language to be considered is its very wide usage around the world. This leads to widely variant pronounciations as the language has followed separate development paths in different nations. The reduced requirements of the hybid solution put forward will facilitate the application of speech synthesis systems to different regional and national accents and indeed different languages.

REFERENCES

[1] Sejnowski, T.J. and Rosenberg, C.R., (1987), "Parallel netwoks that learn to pronounce English text", *Complex Systems 1*, pp145-168.

[2] Elovitz, H.S., Johnson, R., McHugh, A and Shore, J.E., (1976), "Letter-to-sound Rules for Automatic Translation of English Text to Phonetics", *IEEE Trans. Accoustics, Speech and Signal Processing*, 24(6)

[3] Asher,G.M., Curtis, K.M., Andrews, J and Burniston, J, (1990), "A Parallel Multialgorithmic Approach for an Accurate and Fast Text to Speech Transcriber", *ICSLP(90)*, Kobe, pp813-816.

[4] Burniston, J.D., Curtis, K.M. and Craven, M., (1992), "A hybrid rule based/ rule following parallel processing architecture", *PACTA'92*, Barcelona.

[5] Burniston, J.D. and Curtis, K.M., (1994), "A hybrid neural network/ rule based architecture for diphone speech synthesis", *ISSIPNN'94*, Hong Kong.

[6] Kohonen, T, (1989), *"Self-Organisation and Associative Memories"*, Springer Verlag.

[7] Wilde, S.A. and Curtis, K.M., (1994), "A Transputer Based Mixed Supervised/Unsupervised Neural Network For Speech Recognition", *Transputer Applications and Systems '94*, IOS Press.

[8] Craven, M.P., (1993), *Inter chip communication in an analogue neural network utilising frequency division multiplexing*, PhD Thesis, University of Nottingham

TRANSCRIBING NAMES WITH FOREIGN ORIGIN IN THE ONOMASTICA PROJECT

Joakim Gustafson

Department of Speech Communication and Music Acoustics, KTH, Stockholm, Sweden

ABSTRACT

This paper studies the problem of transcribing foreign names. The transcriptions of first names in five languages have been studied to show examples of how this problem has been dealt with in the Onomastica Multi-Lingual Pronunciation Dictionary of European names.

The paper describes this dictionary and the methods used to do the automatic transcriptions for the Swedish part.

INTRODUCTION

Names have a different morphology and phonology compared to ordinary words. This is the reason why the normal letter-to-sound rules used in general text-to-speech systems are inadequate for the transcription of proper names. To deal with the name pronunciation problem, name transcription rules and a name dictionary have to be developed. The objective of the Onomastica project is to produce such rules and a dictionary of European names that will be published on a CD-ROM. This paper will present the problems encountered in the work on this project, and how these have been solved. The transcriptions of first names in five languages are examined to illustrate the problem. The Swedish name transcription system will be presented as well.

THE ONOMASTICA DATABASE

The objective of the ONOMASTICA project, funded by the LRE-programme, is to build a quality controlled, multi-lingual pronunciation dictionary of proper names in Europe. The project covers eleven languages: Danish, Dutch, English, French, German, Greek, Italian, Norwegian, Portuguese, Spanish and Swedish. Transcription of up to 1.000.000 names per language will be produced in a semi-automatic way.

The ultimate pronunciation dictionary should include a carefully verified transcription of each name, but due to the limited resources only a subset of the name list can be transcribed and verified manually. The names are transcribed in three different quality bands, where the first band includes transcriptions judged to be correct for some owners of the name. The second band gives transcriptions that are acceptable to a native speaker/listener. The third band contains names that have been transcribed automatically, without manual checking. The names in bands I & II were chosen according to their frequency in the telephone directory, so that a cumulative coverage of at least 80% was obtained.

The Swedish database, see Table 1, consists of the whole Swedish telephone directory, containing 4.5 million subscribers. The names that occurred more than five times were selected for transcription in band I, obtaining a cumulative coverage from close to 95 % for surnames to 100% for place names (almost all places have more than five subscribers).

Table 1. The Swedish Name Database

Name category	# of names	names with frequency >5
Surnames	228048	46859
Place names	6373	6120
Titles	27055	5370
Street names	65196	39822
First names	60850	10479

THE TRANSCRIPTION SYSTEM

The existing KTH text-to-speech system has been modified and upgraded to cope with proper names. (See Figure 1.) [3]. First the origin of the name is determined to simplify the work for the automatic transcriber [5]. Since the system is designed to imitate a Swedish person attempting to pronounce a foreign

name, it is not certain that the origin tags will be etymologically correct. However, the goal is that they should make the same decisions about language origin as people with ordinary language knowledge would do. To date, 23 tags for origin have been included. The tagging is done using the KTH text-to-speech system with phonological rules that recognise patterns that are specific to different languages [2].

Depending on the origin, each name is sent to a different set of grapheme-to-phoneme modules. The Swedish names are first sent through a TwoLevel-morphology analyser (TWOL) [4] with general Swedish morphs augmented with 1200 name-morphs and a name-lexicon with names occurring in the Stockholm telephone directory compiled during a previous project [2]. The morphology approach is especially suitable for names in Sweden because they are often multi-morphemic. From the morphology analyser morphs with stress and boundary markers are obtained. A set of phonological rules merges these into complete transcriptions. The names that were not transcribed by TWOL are processed by the ordinary Swedish letter-to-sound rules adjusted for names. The foreign names are first run through language-specific letter-to-sound rules with language specific phonemes. These phonemes are then mapped to the closest Swedish equivalents.

All names were manually corrected by the same person in order to obtain consistency. Different tools were used in this process ranging from UNIX scripts to the KTH text-to-speech system. The method of correcting the transcriptions using both orthography, transcription and synthesised speech has proven to be both fast and efficient [3].

NORMALISING SPELLING OF NAMES

People with ordinary names sometimes make their names more unusual and "interesting" by spelling them in an unorthodox way. To address this problem we received a list from the Swedish PTT with different spellings of the same names.

The use of different spellings seems to be more popular in Sweden than in the other four languages examined in this paper. In Swedish only 81% of the first names have a single spelling compared to 97% in Italian, where a sequence of names is used to make the name unique. Swedish first names have up to 24 different spellings, Italian names have up to 6. The different spellings do not always follow ordinary orthographic conventions. One practice that has been observed is the insertion of "h", Bhlom, another the use of "x" instead of "ks" in names ending with "son", for example, Ericxson. Some other popular replacements are: s→z, k→q, k→c, å→aa, ö→oe, ö→eu, i→ie, f→ph, v→fv, v→w.

The spelling of the names must be normalised in order to simplify the automatic transcription.

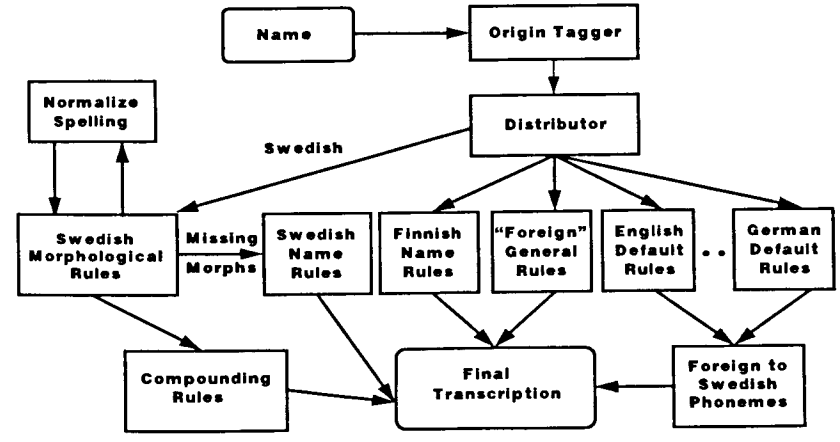


Figure 1. The KTH-system for transcription of names.

A STUDY OF THE FIRST NAMES IN FIVE LANGUAGES

To exemplify the problems of transcribing names with foreign origin the databases containing first names from Great Britain, France, Germany, Italy and Sweden were examined. These varied in size from 10000 in Sweden to 35000 in Italy. (See Table 2)

Table 2. The number of transcribed first names in the five databases.

Sw	En	Fr	Ge	It
10461	16111	12383	31979	35013

The difference in size of the databases could be adjusted by selecting the 10.000 most frequent names, but since the frequency only was available in the Swedish database the databases could not be truncated. All the transcribed names are used in the study.

The structure of names differs from common words, since names often move with people across borders, and adjust to the new language. Table 3 shows the mean number of letters and phonemes in the first names and in the 10.000 most frequent common words [1].

Table 3. Mean number of letters and phonemes in First Names (FN) and Common Words (CW) in the languages.

	letters in FN	letters in CW	phonemes in FN	phonemes in CW
Sw	7.4	7.4	5.6	6.9
En	7.0	7.1	5.6	6.0
Fr	8.9	7.6	6.3	5.2
Ge	8.1	8.7	6.2	7.8
It	10.7	7.4	9	6.9

The names were transcribed in different phonetic alphabets with broad transcriptions. To be able to compare the transcriptions done in the different languages, they were converted from the various phonetic alphabets to IPA. But since broad transcriptions were used the actual realisation of individual phonetic symbols will vary from language to language.

The most common phonemes in each language's first names are shown in Table 4. The table shows that the most common phoneme is [a] in all languages, except for English where it is the [ə].

Table 4. The most common phonemes in the transcriptions of first names, with percentage figures to the right.

Sw		En		Fr		Ge		It	
a	12	ə	9	a	13	a	10	a	15
n	7	ɪ	8	i	10	t	7	o	11
r	7	n	8	ʀ	8	n	7	e	9
l	7	æ	7	l	7	r	6	i	8
ɪ	7	r	6	n	6	l	6	n	8
s	6	l	5	ɛ	5	ɪ	5	r	8
e	6	i	4	m	5	ə	5	l	6
t	4	s	4	s	4	k	4	t	5
m	4	ɛ	4	d	4	ɐ	4	m	4
k	4	m	4	t	4	s	4	j	4

In all languages, except Italian, the ten most common phonemes cover about 60% of all occurring phonemes. In Italian they cover 77%. Italian has the least number of phonemes (28) but the largest number of phonemes per name (9). Swedish and English, however, have the largest number of phonemes (about 40), but the smallest number of phonemes per name (about 5.5). If you pick the names from each country that contains as many as possible of these phonemes you get the following names:

Sw Nils-Einar [nɪlsejnar]
En Alexander [ælɪgzɑndər]
Fr Alexandrine-Marthe [alɛksãdrinmaʀt]
Ge Weichselgärtner [vaɪksəlgɛrtnɐ]
It Vittorio-Emanuele [vitorjoemanuele]

The databases altogether contain 88.000 different names. 79.000 of these only occurred in one country, 981 occurred in all five. The length and stress markers were removed from the transcription of these common 981 first names and the transcriptions were compared. Table 5 shows that the most similar languages are Swedish-German and French-Italian, and those that are most dissimilar are German-Italian. In Italian foreign names often get an Italian spelling, for example Jesus is spelled Gesu in Italy.

Table 5. Number of names that get the same transcription in the language-pairs

	Sw	En	Fr	Ge	It
Sw	-	115	121	201	113
En	115	-	116	115	102
Fr	121	116	-	102	193
Ge	201	115	102	-	87
It	113	102	193	87	-

Table 6. The pronunciation of an initial J in first names, number of occurrences. The likely origins of the names are indicated within the parentheses.

Swedish	English	French	German	Italian
j 458	dʒ 642	ʒ 858	j 558	j 188
ɧ 31 (Fr,Sp)	j 50 (Sw,Ge)	dʒ 37 (En)	dʒ 29 (En)	dʒ 79 (En)
ʂ 12 (Fr)	ʒ 12 (Fr)	j 14 (Sw,Ge)	ʒ 24 (Fr)	i 19 (Fr)
dj 3 (En)	h 6 (Sp)	x 9 (Sp)	x 4 (Sp)	x 17 (Sp)

When examining the databases it was noticed that the letter J in initial position got quite different transcriptions in the five languages (See Table 6). The different ways it can be pronounced seem to be dependent of the likely origin of the name. The names that are considered to be of a certain origin get the pronunciation of "J" that is most common in that language, or is mapped to the closest one in the native language. English names are mostly transcribed with [j] in Swedish, but in some cases the [dj] have been used to imitate the English [dʒ]. Spanish names like Juan [xwan] has been transcribed with the same phonemes in German, French and Italian, but [ɧuan] in Swedish and [hwan] in English.

CONCLUSIONS

The transcription of foreign names presents some problems. There are a number of factors that influence the realisation of a foreign name:

- the level of education in foreign languages
- the phoneme inventory and prosody of the foreign name is frequently adapted to the language spoken
- the context in which it is produced, such as the receiver of the message.

The work on ONOMASTICA has shown that there are a number of decisions that have to be made, such as:

Q1 How to transcribe a foreign name if you don't know the origin

A1 Use the same pronunciation rules for foreign names as for native:
The Swedish name Greger ['gre:gər] gets the Dutch transcription ['xrexər].

Q2 How to deal with foreign phonemes that do not exist in the native language.

A2 a) Map the foreign phonemes to the closest native:
The English name Winston ['wɪnstən] is transcribed ['vɪnstɔn] in Swedish.

A2 b) Enlarge the native phoneme inventory:
The English phonemes [ð] and [θ] are added to the Swedish inventory to get Heather['hɛðər] and Keith['ki:θ].

Q3 How to deal with foreign graphemes.

A3 a) If the realisation of the grapheme in the foreign language is known use the closest native phoneme:
The Swedish town Göteborg [jø:tebɔrj], is transcribed [jɛtɛbɔrg] in Greek, where the Swedish way to pronounce "ö" is known, but it has to be mapped to the closest Greek phoneme.

A3 b) If the realisation of the grapheme is unknown map it to the closest native grapheme:
Göteborg is mapped to Goteborg in Spanish and it is transcribed [gɔtɛβɔr].

ACKNOWLEDGEMENT

The work on the Swedish part of the Onomastica project has been supported by grants from NUTEK.

REFERENCES

[1] Carlson, R. Elenius, K. Granström, B. Hunnicutt,S (1985): "Phonetic and orthographic properties of the basic vocabulary of five European languages" STL-QPSR 1/1985 pp 63
[2] Carlson, R. Granström, B. Lindström, A (1990): "Automatic generation of name pronounciation for a reverse dictionary service," Report, Dept. of Speech Com. Music Ac , KTH.
[3] Gustafson, J.(1994): "Onomastica - Creating a multi-lingual dictionary of European names", w. papers 43, Lund Univ. Dept. Ling. pp 66-70.
[4] Koskenniemi, K (1983) "TwoLevel Morphology: A general computational model for word form recognition and production" Dept. of General Ling. , University of Helsinki
[5] Vitale, T (1991):"An algorithm for high accuracy name pronunciation by parametric speech synthesizer" Computational Linguistics, Vol. 17, No. 3, pp.257-76

THE DELTA SYSTEM WITH SYLLT: INCREASED CAPABILITIES FOR TEACHING AND RESEARCH IN PHONETICS

Susan R. Hertz
Eloquent Technology, Inc., 24 Highgate Circle, Ithaca, New York 14850, U.S.A. and Cornell University

Elizabeth C. Zsiga
Georgetown University, Washington D.C., U.S.A.

ABSTRACT

Sylit is a partial phone-to-speech program designed for use with the Delta System, a sophisticated software tool for teaching and research in phonetics. From a string of phonetic symbols representing a CVC or VCV utterance, Sylit creates a multi-tiered utterance representation (*delta*) from which parameter values for a Klatt synthesizer are automatically derived. The deltas can be modified either interactively with simple commands, or automatically with built-in or user-defined Delta language procedures. Sylit can also quickly implement stepwise changes to a delta to generate stimulus continua or matrices.

INTRODUCTION

The Delta System is a flexible software tool for natural-language processing, with specialized features for speech synthesis. It runs on PCs (Windows or DOS), and Sun and SGI workstations (UNIX). The system includes a linguistically-oriented programming language called Delta [1], and an interactive environment called DeltaTools, both designed for ease in building, manipulating, and synthesizing from nonlinear, "multi-stream" utterance representations (*deltas*). Deltas can be built either with a program expressed in the special Delta programming language, or with interactive DeltaTools commands (or a combination of both). While the system makes manipulating deltas easy, building deltas from scratch can be difficult, especially for users inexperienced in speech synthesis, since they may not know exactly what parameter values to insert. Sylit (derived from Eloquent Technology's more complete text-to-speech program, Eloquence) was designed to alleviate this difficulty [2].

The input to Sylit is a string of phones representing a CVC monosyllable or VCV disyllable in General American English (where C is any stop or fricative and V is any simplex vowel). The output is a delta containing phonological units (e.g., phoneme symbols and associated features), phonetic units (e.g., bursts), and quantitative parameter values (e.g., formant target frequencies) for a Klatt synthesizer [3]. For example, to synthesize the word *toe*, the user would enter it phonetically as:

(1) `to`

A portion of the delta (slightly simplified for space reasons) that the program would construct is shown below:

(2)

```
phone: |t   |   |o          |   |
trans: |    |tr |           |tr |
F1:    |300 |   |550 |      |400 |
F2:    |1600|   |1250|      |850 |
AV:    |0   |0  |58          |0  |
AH:    |0   |63 |0           |45 |
Ms:    |60  |80 |0   |210 |0|60 |
       +----+---+----+----+-+---+
       1    2   3    4    5 6   7
```

This type of representation is called a delta because it consists of "streams" of information of the user's choice. The vertical bars, called *sync marks* (or synchronization marks), coordinate units across streams. All vertical bars in the same column represent the same sync mark. An important property of the delta data structure is that it can combine abstract linguistic information with numeric information in a single integrated representation. The above delta fragment contains abstract linguistic streams representing phone and transition units (see below) as well as streams for first and second formant target values in Hz (`F1` and `F2`), voicing amplitude values in dB (`AV`), aspiration amplitude values in dB (`AH`), and a timing stream called `Ms` that coordinates the units in the other streams. There are also other streams not shown. Although not visible in the above representation, the tokens in the `phone` stream have associated features, such as `consonant`, `vowel`, and, `stop`, which are used by the Sylit program in determining the appropriate acoustic values. From the information in the delta, Sylit generates a set of parameter values for a Klatt synthesizer.

Consider first the `F2` information in the delta. A second formant value of 1600 Hz during the phoneme [t] is followed by an 80 ms transition to an F2 value of 1250 Hz at the beginning of the phone [o]. The values between adjacent sync marks will be interpolated over the specified duration to produce the final synthesizer values. There is no voicing or aspiration (represented by 0 dB for the parameters `AV` and `AH`) during the [t], and there is 63 dB of aspiration amplitude during the transition from the [t] to [o]. At the beginning of [o], voicing comes on and aspiration turns off. The theoretical basis for the phone and transition structure of the deltas constructed by Sylit (and by the more complete program Eloquence) is motivated in [4, 5].

Deltas are constructed by Sylit using context-sensitive rules expressed in the Delta programming language. When Sylit is operating, these rules are visible in one window. In another window, the user can issue interactive commands to trace the operation of the program, to watch the derivation of a delta, to manipulate the delta and synthesize from it, to automatically generate a sequence of deltas (and resulting speech files) differing along one or more dimensions, and much more. Several of these capabilities are illustrated in the sections that follow.

MANIPULATING DELTAS INTERACTIVELY WITH DELTATOOLS

With simple DeltaTools commands, users can easily modify any aspect of the delta, listen to the result, and play selected utterances back to back. For example, a user might modify formant or fundamental frequency trajectories, frequency or duration of burst noise, or the timing of voicing onset relative to stop release. Such manipulations might be used, for example, to illustrate perceptual effects to students, to test perceptual hypotheses, or to test the effect of synthesis rules before incorporating them into a program.

The following commands illustrate one way to make the transition from [t] to [o] in the delta in (2) voiced rather than aspirated:

(3)
```
delta insert [AH 0] 2.3
delta delete AV 2.3
delta delete AV 3
```

The first command replaces the 63 dB aspiration token between sync marks 2 and 3 (the transition from [t] to [o]) with 0 dB. The second deletes the 0 dB `AV` value in the transition. The final command deletes sync mark 3 from the `AV` stream, to extend the 58 dB voicing value to start at the beginning of the transition. Note that the commands refer to the sync marks by number. Though not shown here, sync marks can also be given mnemonic names. The following delta fragment shows the effect of the above commands. Only the relevant streams are shown:

(4)

```
phone: |t   |   |o          |   |
trans: |    |tr |           |tr |
AV:    |0   |58 |           |0  |
AH:    |0   |0  |0          |45 |
Ms:    |60  |80 |0   |210 |0|60 |
       +----+---+----+----+-+---+
       1    2   3    4    5 6   7
```

The user can position values anywhere in the delta, whether there is an existing sync mark at the desired time point or not. For example, consider the following commands, which when applied to the delta in (2) would cause voicing to overlap aspiration, starting 30 ms before the end of the transition from [t] to [o], rather than at the beginning of the [o]:

(5)
```
delta delete AV 2.6
delta insert [AV 58] (3-30).6
delta delete AV 2
```

The first command deletes the two `AV` values (0 and 58) between sync marks 2

and 6. The second positions the value 58 to start 30 ms before sync mark 3 in the AV stream and to end at sync mark 6. The third command deletes sync mark 2 from the AV stream, extending the 0 dB token from the beginning of the delta to the point where voicing begins in the transition. The following delta fragment shows the effect of these commands when applied to the delta in (2):

(6)

```
phone: |t    |      |o          |   |
trans: |     |tr    |           |tr |
AV:    |0        | 58          |0  |
AH:    |0    |63    |0          |45 |
Ms:    |60   |50|30|0    |210|0|60 |
       +-----+--+--+-----+---+-+---+
       1     2  3  4     5   6 7   8
```

Note that the 80 ms time token in the transition has automatically been divided into two tokens, 50 and 30, to position the voicing value at the specified point. In general, sync marks are placed in the Ms stream anywhere an acoustic event begins or ends, and the time tokens are divided accordingly.

MANIPULATING DELTAS AUTOMATICALLY WITH PROCEDURES

One of the most onerous tasks in speech perception work can be creating a synthetic stimulus continuum, a series of synthetic stimuli that vary along one or more parameters. With built-in Syllt procedures, users can automatically generate sequences of synthetic utterances that differ systematically along one or more dimensions.

Assume, for example, the user wishes to create a VOT continuum, beginning with a fully aspirated transition, changed to fully voiced in 20 ms steps, starting with the delta shown in (2). Rather than create such a continuum with a large number of interactive commands of the sorts illustrated above, the user can create such a continuum with two simple commands:

(7)
```
set_pointers(2,3)
vot(58,63,80,20)
```

The first command invokes a procedure that delimits the stretch of the delta to be manipulated (the stretch between sync marks 2 and 3). The second command invokes the procedure `vot` to create the continuum; it specifies the desired voicing value (58), the aspiration value (63), the total duration of the specified stretch (80), and the step size (20). The `vot` procedure then creates a series of deltas and accompanying parameter and speech files with the /t/ gradually changing to /d/. The first three deltas in the series are shown below. Note the changes in the AV, AH, and Ms streams during the transition between [t] and [o].

(8)

```
phone: |t    |   |o          |   |
trans: |     |tr |           |tr |
AV:    |0    |0  |58         |0  |
AH:    |0    |63 |0          |45 |
Ms:    |60   |80 |0    |210|0|60 |
       +-----+---+-----+---+-+---+
       1     2   3     4   5 6   7

phone: |t    |      |o          |   |
trans: |     |tr    |           |tr |
AV:    |0    |0  |58|58         |0  |
AH:    |0    |63 |0 |0          |45 |
Ms:    |60   |60 |20|0    |210|0|60 |
       +-----+---+--+-----+---+-+---+
       1     2   3  4     5   6 7   8

phone: |t    |      |o          |   |
trans: |     |tr    |           |tr |
AV:    |0    |0  |58|58         |0  |
AH:    |0    |63 |0 |0          |45 |
Ms:    |60   |40 |40|0    |210|0|60 |
       +-----+---+--+-----+---+-+---+
       1     2   3  4     5   6 7   8
```

Each of the deltas and accompanying parameter and speech files is automatically named sequentially and saved. The deltas can be recalled for subsequent manipulation with the Delta System, and the speech files can be played later in an experimental setting. A log file keeps track of the name of each delta, and what it contains.

While the above continuum varies in just one dimension, Syllt can also create a two-dimensional matrix. For example, the user could create a vowel space by varying F1 and F2. The following commands, when applied to the delta in (2), change F1 for the entire vowel from 300 to 800 Hz in 250 Hz increments:

(9)
```
set_pointers(3,7)
matrix(F1,300,800,250)
```


In the next step, a second dimension is created, varying F2 from 1500 to 1800 Hz in 150 Hz steps for each of the stimuli created by the previous commands:

(10)
```
dimension 2
matrix(F2,1500,1800,150)
```

Syllt includes procedures for creating continua for VOT, F1, F2, F3, and duration, but the user can easily use these as models to write procedures for other parameters.

MODIFYING SYLLT

Syllt is structured into three main modules: (1) "abstract linguistic" rules, which insert abstract structure such as phones and transitions, (2) "phone" rules, which fill in the acoustic and durational values specific to particular phones, and (3) "default" rules, which fill in any values that remain constant over the whole utterance (such as values for F4 and F5). The user can quickly learn the structure of these modules by tracing the operation of the rules using DeltaTools commands, and watching how the deltas are changed by them. Syllt is also accompanied by extensive documentation, containing a complete description of the structure of the program, and a number of hands-on tutorials to aid the user in learning how to write rules in Delta, trace Delta programs, manipulate deltas, etc. All source code for Syllt is provided.

Users can modify the structure of Syllt for different needs. For example, for teaching purposes, an instructor might want to suspend the application of the phone rules, so that just the abstract linguistic and default rules apply. For a given input string of phones, the program would then create a template into which students could insert the phone-specific values interactively as they learn about different acoustic cues. Students can also write their own rules for filling in acoustic values, and incorporate them into the program.

Users might also wish to modify the existing Syllt modules—to create deltas for a different language or to add new phone types, for example. They might also wish to add new modules to Syllt—perhaps a filter that modifies the values in the delta to create a different voice quality.

CONCLUSION

The Delta System with Syllt provides increased capabilities for teaching and research in phonetics. Teachers will find the program useful for demonstrating the importance of different acoustic cues and for giving students a head start on their own synthesis projects. Researchers will appreciate the natural-sounding utterances that are automatically generated, the easy and precise control over acoustic parameters, and the speed with which stimulus continua can be created. Syllt gives the Delta System added power and flexibility to meet the needs of a wide variety of users.

ACKNOWLEDGEMENTS

The authors thank Jerry Lame and Katherine Lockwood for useful comments on the manuscript. This research has been supported in part by Grant DC00758-03 from NIDCD to Eloquent Technology, Inc.

REFERENCES

[1] Hertz, S. R. (1990), "The Delta programming language: an integrated approach to non-linear phonology, phonetics, and speech synthesis", *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech*, J. Kingston and M. Beckman (eds.), Cambridge University Press, 215-257.

[2] Hertz, S. R., E. C. Zsiga, and M. K. Huffman (1994), "Syllt for building deltas: simple speech synthesis for teaching and research," *J. Acoust. Soc. Amer.* 95, No. 5, Pt. 2, 2815.

[3] Klatt, D. H. and L. C. Klatt (1990), "Analysis, synthesis and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Amer.* 87, 820-857.

[4] Hertz, S. R. (1991), "Streams, phones, and transitions: toward a phonological and phonetic model of formant timing", *J. Phon.* 19, 91-109.

[5] Hertz, S. R. and M. K. Huffman (1992), "A nucleus-based timing model applied to multi-dialect speech synthesis by rule", *Proc. Int. Conf. Spoken Lang. Proc* 2, 1171-1174.

AN INTEGRATED PHONOLOGICAL-PHONETIC MODEL FOR TEXT-TO-SPEECH SYNTHESIS

Jill House, Department of Phonetics and Linguistics, University College London
Sarah Hawkins, Department of Linguistics, University of Cambridge

ABSTRACT

We propose a model for text-to-speech synthesis (TTS) in which the units of phonological structure, correctly identified, determine phonetic interpretation in a non-arbitrary manner. A notion of dominance is used in the phonetic interpretation, which involves the interrelationship of all levels of structure in spectral, temporal and intonational domains.

1. PRELIMINARIES

1.1 Objectives

Our main aim is to maximise the intelligibility and naturalness of British English TTS using structures and procedures which are constrained in a principled, non-arbitrary way. Since a further objective is to design a model with multi-lingual potential, we must define structures which have quasi-universal applicability, while allowing that the detailed properties of the units identified, and of the processes they enter into, may be language-specific.

1.2 Theoretical considerations

Many current theories of phonology emphasise the discovery of non-arbitrary universal principles and language-specific parameter specifications; this represents a move away from an over-powerful re-write rule format in favour of rich structural representations. We too aim to identify a hierarchical structure whose components constitute all those units which enter into linguistic contrasts. In the phonetic interpretation we exploit a notion of **dominance** which involves the inter-relationship of all levels of structure in spectral, temporal and intonational domains.

Our model draws on the strengths of relational, structure-based systems such as YorkTalk [1]; an important difference is that we explicitly address phone-sized segments, within the structures of which they form part, in our phonetic interpretation. This allows us to make straightforward use of the segment-based transcriptions provided by the lexicon.

1.3 Technical considerations

The proposed model is adaptable in principle to both concatenative and formant synthesis systems, though details of the implementation would differ. Input to the model is assumed to be the output of a parser (e.g.[2]), which takes initial responsibility for building prosodic structure.

2. PROSODIC COMPONENT

2.1 The parser interface

The job of the parser is to derive a hierarchical prosodic structure using morphological and grammatical category information stored in a large pronunciation lexicon. Following [3], its grouping strategies are further motivated by principles of verb-balancing and verb-adjacency. To a large extent, the parser provides us with the required units of phonological structure, in tree format, with labelled nodes partially marked for prominence and boundary strength. Lexical look-up, supplemented where necessary by spelling-to-sound rules, supplies a segmental transcription, complete with lexical stress assignment to individual syllables, and an indication of the boundaries of phonological words.

The first task of the prosodic component is to check the well-formedness of the tree structure supplied by the parser. In practice this means using metrical principles to assign strong/weak values to nodes, supplementing the partial prominence orderings, and assigning segments to syllabic constituents. The completed structure is not subsequently altered.

2.2 The prosodic hierarchy

Following e.g. [4], we propose that utterances should be organised into constituent units arranged in a prosodic hierarchy. For us, the components of such a hierarchy include: *intonational phrase > pitch accent group > foot > syllable > syllabic constituents > skeletal positions (segments).* Units at the top of the hierarchy are made up of constituent units from the lower levels. One unit at each level acts as the dominant **head** constituent. The head typically constrains the phonetic interpretation of all constituents on the same level, and also of those units at lower levels which it directly dominates in the hierarchy. During the implementation, which proceeds top-down, information about the headedness of constituents is retained as an index that reflects the dominance hierarchy used in computing spectral, temporal and intonational parameters.

We recognise that intonational phrases are themselves loosely organised into "utterances", which in TTS typically correspond to sentence-length chunks of text, and that these chunks may be further organised into prosodic paragraphs. However, at levels above the intonational phrase the dominance hierarchy is not applicable.

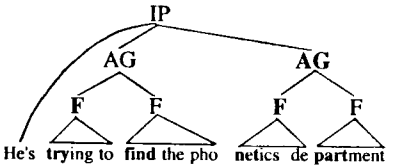
Constituents in the prosodic hierarchy have the following properties in English:
intonational phrase (IP): the domain for a well-formed intonation contour; all lower level constituents must be organised within an IP. The IP consists of one or more pitch accent groups, and the last of these constitutes the head.
pitch accent group (AG): this consists of an accented syllable (both stressed and pitch prominent), together with any unaccented syllables following it. The AG may contain several feet, of which the first (containing the accented syllable) is the head.
foot (F): like the pitch accent group, this unit is left-headed, consisting of an obligatory strong syllable (the head) and any weak syllables following it. For our purposes the foot is not bounded:

(1) $_{F}$[**prop**erties are de] $_{F}$[**creas**ing in] $_{F}$[**value**]

successive feet in (1) contain 5, 3 and 2 syllables respectively.

The organisation of IPs into AGs and feet is represented in (2). Constituents shown in **bold** are heads of the units of which they are daughters. Note that unstressed syllables at the beginning of an IP need not be organised into feet or pitch accent groups, but are dominated directly by the IP itself.

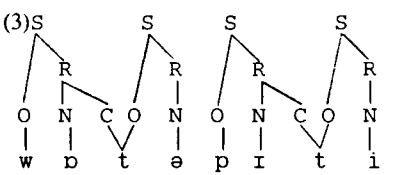
(2) $_{IP}$[He's $_{AG}$[$_{F}$[**try**ing to]$_{F}$[**find** the pho]]$_{AG}$[$_{F}$[**net**ics de]$_{F}$[**part**ment]]]



syllable (S): all segments are organised into syllabic constituents: obligatory **rhyme (R)** (head), and optional consonantal **onset (O)**.
syllabic constituents: the rhyme contains an obligatory **nucleus (N)** (head), normally filled by a vowel, and optional consonantal **coda (C)**. Onset, nucleus and coda may all branch. Dominance relationships between the segmental components of these constituents are expressed in terms of an index of coarticulation resistance (see 3.1).

Phrase-internally, consonants may be *ambisyllabic*, belonging simultaneously to the coda of one syllable and the onset of the next, proving phonotactic constraints are not violated.

(3)



Lexical stress further constrains ambisyllabicity word-internally: the /t/ in "PITy"(3), following a stressed vowel, is ambisyllabic, whereas the /t/ in "preTEND" would be assigned only to the onset of the second syllable. This reflects a different realisation of /t/ in such contexts in many varieties of English. In connected speech, ambisyllabicity applies more widely, to fill the empty onsets of vowel-initial words, regardless of stress: the /t/s in both "what a" (3) and "what ELSE" may be considered ambisyllabic.

2.3 The phonological word

One unit that is not included explicitly in our prosodic hierarchy is the **phonological word (PW)**. This important unit, created in the parser, involves modifying lexical representations by attaching weak,

monosyllabic function words as clitics to the last foot of a preceding word. Feet that are internal to the PW are prime sites for lenition processes (e.g. the reduction of "want to" to "wanna"). However, the phonological word is partly independent of other units in the hierarchy, since word boundaries and foot/AG boundaries need not coincide:

(3) This # im$_F$[PORtant] # $_F$[PRINciple] # $_F$[MUST be # re$_F$[MEMbered] #
(# = boundary of phonological word)

PW boundary information is passed on at the parser interface to ensure appropriate phonetic interpretation.

3. PHONETIC INTERPRETATION

We have three aims for the phonetic interpretation. First, the computed output values of each parameter should vary fairly continuously within a given range, so that the synthetic signal mimics the gradient parameter values of natural speech. Second, like the phonological structures, the phonetic principles we use should be language-independent; differences in parameter values reflect differences due to language, accent, and speech style. Third, for a given speech style, we seek a single control structure to account as far as possible for coarticulation, allophonic variation, timing/rhythm, and connected speech processes, in the belief that these are different facets of the same general process. For example, we claim that the syllable defines a unit of major importance for all these aspects; particular properties associated with the syllable will have different importance for, say, traditional coarticulation and timing. The work that will unify control of these four processes is not complete, so these are preliminary ideas.

For clarity in this short paper, we first discuss the control of coarticulation in its traditional sense, and then briefly indicate links with the other factors.

3.1 Coarticulation

Coarticulation is traditionally defined as the influence of one speech sound upon another, where a "sound" is a phone-sized unit. This definition typically includes obligatory effects due to aerodynamic and biomechanical properties of the vocal tract e.g. CV transitions, and certain non-obligatory processes that are often seen as easing articulatory demands, such as spread of lip rounding or nasalization. The definition usually excludes optional processes in which sounds mutually influence one another, such as assimilatory connected speech processes and controllable allophonic variation. With rare exceptions [5], explicit mention of timing is also excluded.

Most traditional coarticulatory effects for a given speech rate and style can be achieved through knowledge of the properties of the current syllable and of those adjacent to it. Within-syllable coarticulation accounts for CV and VC effects, while coarticulation across adjacent syllables (but not across a pause) accounts for V-to-V coarticulation and C ambisyllabicity.

We expect to achieve coarticulatory effects in a way similar to YorkTalk, by systematically computing parameter values in an order that reflects the domain of influence of each syllabic constituent, both within and between syllables. One difference from YorkTalk is that each terminal element (allophone) is associated with an index of coarticulation resistance [6] that affects the extent to which that phone influences the parameter values for the entire syllable. This coarticulation resistance index is incorporated into the dominance hierarchy.

3.2 The dominance hierarchy

While the domain of coarticulatory influence need be only as local as adjacent syllables, we need access to further information about position in the prosodic hierarchy and PW boundaries to compute the parameter values. The dominance hierarchy expresses the degree of influence that different factors have on syllabic constituents. Conceptually similar to [7]'s "acme articulations", it includes, but is broader than, "coarticulation resistance" [6]. Since the dominance hierarchy is expressed structurally, all information from the prosodic tree can contribute to it. Each node in the tree, from IP downwards, is associated with a weight, sensitive to the headedness of the constituent at each level. The weight associated with each terminal element is the product of all these weights on the structural components. For each relevant acoustic parameter in the syllable (e.g. aspiration amplitude, vowel duration, etc), this "terminal" weight is scaled appropriately and used in calculating the output value for that parameter. Other factors contributing to parameter specification include word or morphemic status, number of syllables, and position of the current syllable in the word.

The dominance hierarchy is intended to produce the gradient output that we aim for, and to be pivotal in bringing together coarticulation, allophonic variation, timing, and connected speech processes.

3.3 Allophonic variation, timing, and connected speech processes

Since we seek a general structure appropriate for different languages, accents, and speaking styles, we aim to standardise the control processes as far as possible, and to avoid rule proliferation. To this end, we aim to maximise the overlap in structural description and control parameters between coarticulation, allophonic variation, timing, and connected speech processes.

Since "coarticulatory processes" differ between languages, it seems that most aspects of coarticulation are under speaker control. We can thus conceptualise a continuum which extends from inevitable biomechanical and aerodynamic consequences (many traditional coarticulatory processes) at one end, to arbitrary, language-specific but highly controlled effects (many cases of allophonic variation) at the other. For example, CV transitions are evidence of coarticulation in the traditional sense, while differences in, say, the realisation of /u/ in the context of word-medial /r/ vs /z/ [8] represent allophonic variation of a more arbitrary type that can nonetheless be seen, and hence modelled, as coarticulatory in origin.

The structures relevant to (traditional) coarticulation and timing control are distinct in some ways but not in others. For example, the foot contributes directly to timing but not, we think, to coarticulation; the syllable, on the other hand, is a fundamental unit for both,, and desired coarticulatory effects can often be produced by appropriate adjustments to timing.

Similar patterns of convergence and difference exist for the other aspects of phonetic realisation. Assuming segments are defined in terms of structure, allophonic variation (e.g. aspiration amplitude and duration) largely reduces to gradient effects due to properties already defined by phonological structure, and hence represented in the dominance hierarchy, or to variations due to changes in speech rate or style (such as /t/ flapping). Some connected speech processes must be modelled for all fluent speech, but the degree and quality of these processes is style-dependent. Processes such as assimilations can be seen as having their origin in ease of articulation and hence in traditional coarticulation, but some cases may be analysed as involving structural differences.

4. SUMMARY

Our proposed model closely integrates a prosodic hierarchy and an acoustic dominance hierarchy, which together determine TTS parameters. They ensure a gradient output, properly constrained, which will increase naturalness and acceptability.

Partly supported by Telia Promotor Infovox AB

REFERENCES

[1] Local, J. (1992), "Modelling assimilation in a non-segmental rule-free phonology", in G.J. Docherty and D.R. Ladd (eds), *Papers in Laboratory Phonology II*, Cambridge: CUP, 190-223.

[2] Youd, N.J. & A. Slater (1994), "Parsing for prosody", unpublished ms.

[3] Bachenko, J. & E. Fitzpatrick (1990), "A computational grammar of discourse-neutral prosodic phrasing in English", *Computational Linguistics* 16(3), 155-170.

[4] Nespor, M. & I. Vogel (1986), *Prosodic Phonology*, Dordrecht: Foris.

[5] Harris, K.S. & F. Bell-Berti (1981), "A temporal model of speech production", *Phonetica* 38, 9-20.

[6] Bladon, R.A.W. & A. Al-Bamerni (1976), "Coarticulation resistance in English /l/", *Journal of Phonetics* 4, 137-150.

[7] Kelly, J. (1989), " On the phonological relevance of some non-phonological elements", in T. Szende (ed.) *Proc. of the Speech Research '89 (Hungarian Papers in Phonetics* 21), 56-59.

[8] Hawkins, S. & A. Slater (1994), "Spread of CV and V-to-V coarticulation in British English: implications for the intelligibility of synthetic speech", *Proc. ICSLP 94*, 1, 57-60.

GENERATING PROSODIC STRUCTURE FOR RESTRICTED AND "UNRESTRICTED" TEXTS

*Anders Lindström** *Merle Horne*** *Tomas Svensson****
*Mats Ljungqvist** *Marcus Filipsson***

** Telia Promotor Infovox AB, P.O. Box 2069, S-171 02 Solna*
E-mail: {anders.lindstrom|mats.ljungqvist}@infovox.se
*** Dept. of Linguistics, Lund University, Helgonabacken 12, S-223 62 Lund*
**** Dept. of Linguistics, Stockholm University, S-106 91 Stockholm*

ABSTRACT

A new text analysis component for the generation of prosodic structure in Swedish text-to-speech conversion is described. It combines shallow syntactic analysis and prosodic parsing algorithms with referent tracking and treatment of lexicalized phrases. It is explained how this approach can be used in restricted and unrestricted texts.

INTRODUCTION

Text-to-speech (TTS) conversion is performed in three stages. In the first stage, the text is analyzed and a symbolic linguistic representation, including pronunciation and prosodic structure, is produced. In the second stage, this symbolic representation is realized in terms of a synthesis specification[1], and in the third and last stage, this specification is used to drive the synthesizer.

One of the most important functions of the first stage of TTS conversion is to produce a proper prosodic description. Such a description of an unrestricted text passage cannot be derived from syntactic information alone, but also depends on semantic and pragmatic factors. A current trend is to use shallow syntactic analysis, such as that produced by applying prosodic grouping algorithms to the output of a part-of-speech tagger. For "discourse-neutral" English texts, it has been shown to be possible to obtain improved prosodic phrasing using this approach [1].

[1]The specification can typically be the formants, bandwidths, amplitudes etc. of a formant synthesizer, or unit selection information plus prosodic parameters of a concatenative synthesizer.

In previous work [2], we have shown that an important aspect of text analysis for TTS conversion, at least in restricted texts, involves keeping track of the coreferential status of lexical items, so that already mentioned items or concepts can be deaccented when re-encountered in the text.

In this paper, we propose a new text analysis component (TAC) for the generation of prosodic structure in Swedish TTS conversion. The structure is a hierarchy of the prosodic constituents "prosodic word", "prosodic phrase" and "prosodic utterance", coupled with information regarding coreferential status (affecting degree of prominence) and boundary strength (later realized as boundary tones, degree of final lengthening and pause length) [3, 4]. We combine shallow syntactic analysis, referent tracking, prosodic parsing algorithms and treatment of lexicalized phrases, and show how this approach can be used both in restricted and unrestricted texts.

The structure of the TAC will be outlined, and the tagging and referent tracking modules will be examined. This work is based on what has previously been reported [2, 3, 5], and the overall goal is to produce a system for high-quality text-to-speech conversion not only in Swedish, but also in other languages [6].

TEXT ANALYSIS

As pointed out earlier [5], the first stage of TTS conversion should not be regarded as a simple "text pre-processor", but rather as a quite complex knowledge-based system, with many highly specialized knowledge sources, that are all located *within* the TTS system, and are allowed to share knowledge with each other. Several linguistic levels of analysis are required to produce the desired prosodic structure. Pragmatics, semantics, and syntax are necessary components in such a scheme.

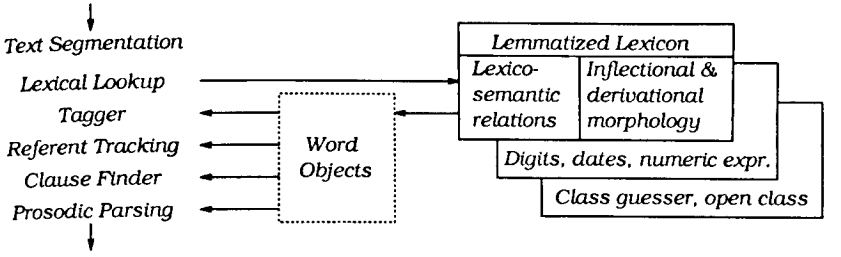


Figure 1: *System architecture of the text analysis component* (TAC)

An outline of our system structure is shown in Figure 1. First, the input text is segmented into sentences and words. The words, or word sequences, as is the case when dealing with lexicalized phrases, are looked up in the lexicon, which is an implementation of a large, lemmatized pronunciation dictionary [7], further equipped with lexico-semantic relations. The result of lexical lookup is stored in word objects in terms of morphological and semantic tags. The sequence of word objects is disambiguated by the tagger. This is a hybrid tagger, where rule-base constraints, dealing with lexicalized phrases, are applied both before and after probabilistic tagging. The resulting, disambiguated sequence of word objects is accessed by a module for referent tracking [2], which makes use of a memory function combined with the lexico-semantic relations stored in the dictionary, to keep track of and mark already mentioned items or concepts. The resulting structure of words, tags and "new/given" status information is used by a pattern matching component to delimit clauses. This information is used, then, to construct prosodic phrases, within which prosodic words are defined [3].

Part-of-speech tagging

The probabilistic part of the tagger used for part-of-speech (POS) disambiguation is an adaptation to Swedish of the Xerox POS tagger [8], a first order Hidden Markov Model based tagger. A tagger of this type has a limited scope of two tokens, making decisions based on transition and symbol probabilities. The texts used for training have been derived from the Stockholm-Umeå Corpus (SUC) [9]. The tagset used is a subset of the SUC tagset consisting of 43 tags.

Choice of tagset content and size is crucial to performance, since the tagset provides the description of the language, with which the tagger is to detect distributional distinctions. We have found that tagset content should reflect as many of the distributional distinctions as possible, and yet not reflect ambiguities that the model cannot resolve, given its inherent limitations.

When constructing a tagset to allow the shallow syntactic analysis required, we found that the language description obtained using a small tagset is not necessarily the best for resolving ambiguities resulting from applying this tagset to a given language. Improvements can be made

Table 1: *Tagging accuracy when A: Tagged & evaluated with 27 tags, B: Tagged & evaluated with 43 tags and C: Tagged with 43 tags & and evaluated when mapped onto the 27 tag subset.*

	Correctly tagged tokens [%]		Tags per word
	Total	Ambiguous	
A	95.1	86.7	1.65
B	94.8	86.7	1.68
C	96.1	90.2	1.65

by splitting and relabeling original tags, in order to separate lexical items with conflicting distributional properties, while avoiding excessive increase in perplexity. The results shown in Table 1 demonstrate the considerable improvements in tagging accuracy that can be achieved by tagging with a larger set (43 tags) and mapping onto a smaller set (27 tags).

Referent tracking

We have previously shown that, when applied to texts from a known, restricted domain, referent tracking algorithms can benefit from taking into account several types of lexico-semantic relations that may indicate co-reference, e.g. morphological identity (both inflectional and derivational), (partial) synonymy, hyponomy-hyperonymy ("is-a" relations) and meronymy-holonymy ("has-a" relations) [2]. In addition to this, some concepts can be considered as pragmatically "given" in a restricted domain, such as *per cent* in stock market texts. When confronted with a word in running text, the referent tracking algorithm consults its "memory"[2] of processed entities to determine whether the current word has been recently mentioned or not, and assigns a status of either "given" or "new" to the word.

Restricted vs. "unrestricted" text

By "unrestricted" texts we mean texts, which can in fact consist of sections, which are narrow in domain, but where these domains and the section boundaries are *not known* to the TTS system in advance. Because of the modular architecture used [5], going from a restricted domain to either another restricted domain, or to unrestricted text, is achieved by inserting or changing knowledge only in a few well-defined places in the system.

For better coverage on unrestricted text, more effort has to be spent on modelling different textual conventions, such as date formats, fractions, abbreviations etc. This is achieved by extending the regular grammar that performs text segmentation and tokenization on the one hand, and, on the other hand, adding to the lexicon corresponding "methods" dealing with those conventions. In this way, the form and internal structure of dates, e-mail addresses etc. can be exploited, and dealt with using (sometimes) more appropriate methods than lexical listing.

As regards lexico-semantic relations, necessary for referent tracking, in the unrestricted case it is possible, even without knowledge of topic structure (see Discussion), to recognize coreference using morphological identity relations from the lemmatized lexicon.

The identification of lexicalized phrases in unrestricted texts [10], which is important both in order to increase the naturalness of synthetic speech and to ease the burden on the probabilistic tagger [5], has been taken even further in a restricted domain than is otherwise possible: When applying the TAC to texts from the stock market domain [3], lexical items are marked with semantic tags such as share-name (*Atlas Copco, Ericsson* etc.), share-type (*A* or *B*) and share-modifier (*bundna (bound)* or *fria (free)*). These tags are later used in the first stage of tagging, to match patterns, e.g. share-name share-type share-modifier, and construe lexicalized phrases, like *Atlas Copco B fria*, which behave like lexical items both grammatically and prosodically.

When it comes to POS tagging, a Hidden Markov Model based tagger, such as the Xerox tagger, is especially easy to adapt to new domains, since it only requires raw, untagged text, a tagset and a lexicon to be retrained.

DISCUSSION

In our work on a restricted domain, the lexico-semantic relations were coded manually on top of the lexicon, but results from the fields of lexicography and information retrieval indicate that lexico-semantic relations could be automatically inferred from other sources, such as corpora, "ordinary" dictionaries, or thesauri. While thesaural information can be directly useful for restricted texts, its use may be a little less straightforward in unrestricted texts, because of problems with polysemy. A solution to this could be to use word sense disambiguation techniques [11]. Related techniques [12] could be used to find the topic (or section) boundaries that are not known in advance, but which are relevant in order to determine a better domain for the referent tracking algorithms.

Different parts of this system place different demands on tagger output, sometimes conflicting with the tagset required by the tagger for optimum accuracy. In our experience, it is virtually impossible to meet the tagger-external demands, while also meeting tagger-internal ones. Therefore, for the sake of tagging accuracy, one should choose a tagset that meets the internal demands only, and leave unresolvable ambiguities to other modules.

Further work is needed in several of the areas mentioned in this paper. Particularly, the area of lexicalized phrases needs to be further studied from both grammatical and prosodic points of view. The proposed text analysis component (TAC) also needs to be formally evaluated.

ACKNOWLEDGEMENT

We thank Telia Research and the HSFR/NUTEK Language Technology Programme for supporting this work, Gunnar Eriksson, Dept. of Linguistics, Stockholm University, for his work on modifying the tagger for Swedish, and Janne Lindberg, Dept. of Linguistics, Stockholm University, for sharing his results on lexicalized phrases.

REFERENCES

[1] J. Bachenko and E. Fitzpatrick. A computational grammar of discourse-neutral prosodic phrasing in English. *Computational Linguistics*, 16:155–167, September 1990.

[2] M. Horne, M. Filipsson, M. Ljungqvist, and A. Lindström. Referent tracking in restricted texts using a lemmatized lexicon: Implications for generation of intonation. In *Proc. of the European Conf. on Speech Technology*, volume 3, pages 2011–2014, Berlin, 1993. ESCA.

[3] M. Horne and M. Filipsson. Generating prosodic structure for Swedish text-to-speech. In *Proc. of the 3rd Intl. Conf. on Spoken Language Processing*, pages 711–714, Yokohama, 1994.

[4] M. Horne and M. Filipsson. Computational modelling and generation of prosodic structure in Swedish. In *Proc. of the 13th Intl. Congr. of Phonetic Sciences*, Stockholm, 1995.

[5] A. Lindström and M. Ljungqvist. Orthographic processing *within* a speech synthesis system. In *Proc. of the 3rd Intl. Conf. on Spoken Language Processing*, pages 1683–1686, Yokohama, 1994.

[6] M. Ljungqvist, A. Lindström, and K. Gustafson. A new system for text-to-speech conversion, and its application to Swedish. In *Proc. of the 3rd Intl. Conf. on Spoken Language Processing*, pages 1779–1782, Yokohama, 1994.

[7] P. Hedelin and D. Huber. A new dictionary of Swedish pronunciation. In Kjell Morland and Kari Sørstrømmen, editors, *Proc. of the Scandinavian Conf. in Computational Linguistics*, pages 105–117. Norwegian Computing Centre for the Humanities, Bergen, Norway, 1991.

[8] D. Cutting, J. Kupiec, J. Pedersen, and P. Sibun. A practical part-of-speech tagger. In *Proc. of the 3rd Conference on Applied Natural Language Processing*, pages 133–140, Trento, Italy, 1992.

[9] E. Ejerhed, G. Källgren, O. Wennstedt, and M. Åström. The linguistic annotation system of the Stockholm–Umeå corpus project. Technical report, Dept. of General Linguistics, Umeå 1992.

[10] J. Lindberg. Detektering av lexikaliserade fraser för text-till-talkonvertering. In *Proc. of The 9th Conf. of Nordic and General Linguistics*, Oslo, 1995. Forthcoming.

[11] D. Yarowsky. Word sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proc. of the 14th COLING*, pages 454–460, Nantes, France, 1992.

[12] J. Morris and G. Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17:21, 1991.

[2] The memory size is rather arbitrarily chosen to be the most recent 60 words in the input text.

IIIGH INTELLIGIBILITY AND NATURALNESS CHINESE TTS SYSTEM AND PROSODIC RULES

Chu Min and Lu Shinan
Institute of Acoustics, Academia Sinica, Beijing, China, 100080

ABSTRACT

This paper presents a new waveform concatenation Chinese TTS system based on TD-PSOLA method. It can produce clear and natural Chinese speech. The intelligibility and naturalness of the above system are 94.1% and 7. 8 respectively. The flowchart of the system are given. The prosodic rules of this TTS system, which are based on the acoustic analyses of broadcast-speech, are discussed in details.

1. INTRODUCTION

Researches on synthetic Chinese disclose that only when both the segmental and supra-segmental features of the synthetic speech are similar to those of the natural one, the synthetic speech will sound intelligible and natural[1]. Segmental features of Chinese syllables are relatively stable, but prosodic features such as the pitch contour and the duration, which are main factors that affect the naturalness of Chinese, often change greatly in continues speech. The TD-PSOLA[2] method can modify the pitch, duration and intensity of waveforms with little distortion of dynamic spectrum, so it is very suitable for synthesis Chinese. The system discussed in this paper are based on the TD-PSOLA method and utilizes mono-syllables as synthetic units.

There are many kinds of Chinese dialects in use today. Even for Standard Chinese, people from different age groups or from different educational backgrounds speak differently. The speech uttered in CCTV news broadcast is the first prototype of our TTS system. A library for prosodic rules, which includes tone and sandhi patterns of lexical items, duration model for words, stress patterns for words and sentence intonation models, are built up according to acoustic analyses of broadcast-style speech. By scanning the input text on the word level and the sentence level, the prosodic rules assign pitch contours and durations to syllables of a sentence comprehensively.

2. FLOWCHART OF THE SYSTEM

The flowchart of the system is in Fig.1. The system starts from Text Scan Module, where the input text is decomposed into sentences, breathing groups and words, pronunciations of constituent Chinese characters are decided, prosodic markers are separated from the text and the order of syllables, words or phrases, breathing groups and sentences are registered. Then, pitch contours and durations of syllables of a sentence are assigned. After that, syllable waveforms and their pitch marks fetched from the syllable library are conveyed with the corresponding target pitch contour of the syllable to the PSOLA module. At last, the system concatenates the synthetic waveforms, inserts pauses in proper position, and conveys them to the D/A converter. Then, very clear and natural continues synthetic speech can be heard. The system showed in Fig.1 is carried out on a 386 PC computer. All the additional hardwares needed are a sound blaster and a speaker. The system can run on real time.

3. LIBRARY OF PROSODIC RULES

We recorded six-hour broadcast-speech and analyzed them in the length of about half an hour. We also analyzed half an hour's recording materials of several

chosen texts reading by a male and a female, both of whom are senior students studying announcing arts in the Institute of Broadcast at Beijing. On the basis of the work already done, a library of prosodic rules, which includes tone and sandhi patterns of lexical items, duration distribution model for words and phrases, stress patterns for words and phrases and sentence intonation models, are built up .

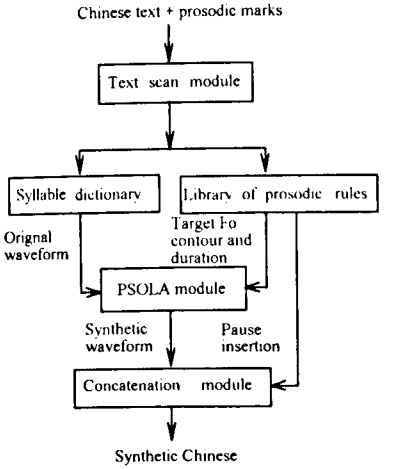


Fig.1. Flowchart of the TTS system

3.1 Tone and sandhi patterns of lexical items

Though the four Chinese tones have rather stable tone patterns in the case of isolated syllables, they undergo various modifications in continues speech due to coarticulations. Yet, for words or phrases with the same tone combination, the overall tone-sandhi patterns are almost unchanged.

In the library of prosodic rules, tone-sandhi patterns for disyllabic, trisyllabic and polysyllabic items are stored. Since the majority of Chinese lexical items are disyllabic or trisyllabic ones, there are 20 tone-sandhi patterns for disyllabic items and 120 patterns for trisyllabic ones. For items having more than three syllables, tone-sandhi patterns are formed by concatenating monosyllabic tone patterns by special rules. The library stores three monosyllabic tone patterns for each of the four Chinese tones according to its position (at beginning of a word, at end of a word and others) . All tone patterns have been normalized into the same tonal pitch range of normal stressed words.

3.2 Syllable duration distribution model for words and phrases

It is very helpful for improving the naturalness of synthetic speech to study the distribution of syllable durations in words or phrases. It is found that there are two kinds of duration patterns for disyllabic word . If the accent is on the first syllable, the duration of the first syllable is longer; otherwise, the last syllable is longer. In trisyllabic words, the middle syllable is the shortest, while the last syllable is the longest. In words with more than three syllables, the syllable duration is often alternated between long and short. Besides when the number of syllables in word increases, the duration of each syllable decrease. The syllable duration is also affected by its position in sentence. The last syllble of a breathing group, a subsentence or a sentencee is always lengthened.

3.3 Stress patterns for words and phrases and the sentence intonation model

The system turns Chinese text to speech sentence by sentence. A sentence is usually decomposed into several subsentences and a subsentence into breathing groups, a breathing group into words or phrases. Y. R. Zhao first used the concept of the range of pitch to describe the tone and the intonation in Chinese[3]. Shen Jiong proposes the term of the tonal pitch range[4], which is adopted in this paper. It has many advantages to describe the intonation of a tonic language such as Chinese by the movements of the up-line and the base-line of tonal pitch ranges. The movement of the up-line indicates the stress levels of words and the movement of base-line

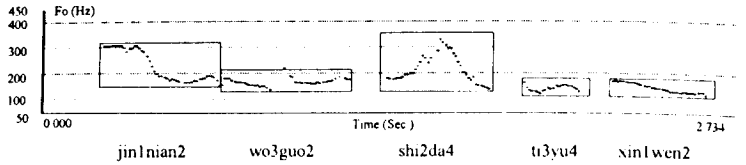


Fig.2.Tonal pitch range of the sentence of "Jin1nian2 wo3guo2 shi2da4 ti3yu4 xin1wen2"

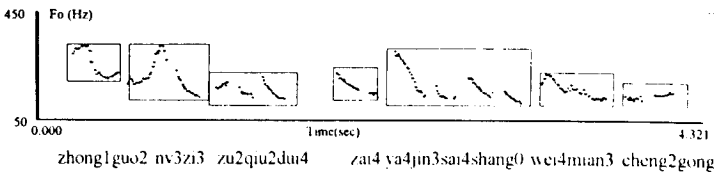


Fig.3. Movements of the up-line and the base-line of sentence of "Zhong1guo2 nv2zi3 zu2qiu2dui4 zai4 ya4jun3sai4shang0 wei4mian3 cheng2gong1."

reflectes the rhythm. At the end of sentences with different mood, such as indicative mood, interrogative mood and exclamatory mood, the position of the two line are different.

In acoustic studying, the tonal pitch ranges of words are found to be the basic blocks of the intonation of a sentence. The up-line of the stressed word always moves upwards to make the size of the stressed tonal pitch range larger than the size of the not stressed one. In Fig 2, the stressed words are "jin1nian2" and "shi2da4" . The base-line of tonal pitch ranges have a trend of moving downwards in a breathing group, and they rise at the beginning of a new breathing group. In Fig.3, there are two breathing groups in the sentence. The base-line falls in the first group, rises at "zai4" and falls again in the second breathing group. The two-line intonation model is used in our TTS system.

There are altogether five stress levels for words and phrases, which are heavy, secondary, normal, weak, and light. The up-line moves up and down according to the stress levels of words. Sometimes Chinese expresses stressing by expanding the duration. The system proposes a series of symbols for users to modify the durations of words when necessary.

The rising and falling of base-lines and inserting proper pauses in speech are corresponding to the overall rhythm. Breathing groups usually express relatively independent meanings, and are units of rhythm. The base-line falls in a breathing group until reaching the end of the group. After a little period of pausing, a new breathing group begins. The base-line rises and another falling period starts Movements of the upline and the base-line at the end of a sentence are also the main means for expressing the mood of the speech. At the end of indicative speech, the base-line falls and the up-line falls more greatly; At the end of interrogative speech, the base-line rises and the up-line almost unchanges; At the end of exclamatory speech, the base-line falls and the up-line rises.

The system inserts 10ms pause between words or phrases, 100ms pause between breathing groups, 300ms pause between subsentences. At the end of indicative sentences there are 500ms pauses and at the end of interrogative and exclamatory ones the pauses are 700ms.

When there is no prosodic symbols, the system can decide the tonal pitch range and the duration automatically. By using a few prosodic symbols to change the stress level or the duration of some words in texts, the synthetic speech will be improved in naturalness.

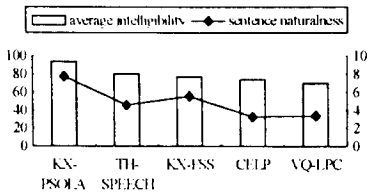
4. RESULT OF EVALUATION

The system took part in a formal evaluation of speech quality of synthetic Chinese which is held by the specialist group of the State High Technology Development Project of China in May,1994. The evaluation includes intelligibility and naturalness evaluations. Here gives out some results in Fig.4. KX-PSOLA is the proposed system in this paper. KX-FSS is another TTS system of our laboratory, which is a formant synthesizer and uses the same prosodic rule library as KX-PSOLA. TH-SPEECH is a another waveform concatenation system, and CELP and VQ-LPC are two systems using LPC method. There are altogether 16 listeners. The average intelligibility and the naturalness of the synthetic speech of KX-PSOLA are 94.1% and 7.8 (the naturalness of natural speech is 10.) respectively.

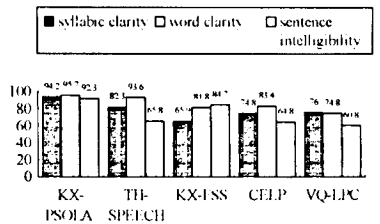
In Fig.4 (a), the average intelligibility of the two waveform concatenation systems KX-PSOLA and TH-SPEECH are higher than others, and the sentence naturalness of KX-PSOLA and KX-FSS, which have a good prosodic model, are higher. Both the average intelligibility and the naturalness of KX-PSOLA are the highest. In Fig.4 (b), though the syllable and word clarity of KX-FSS is lower , the sentence intelligibility of KX-FSS is higher. This is due to the contribution of the prosodic rules.

5.CONCLUSION

Waveform concatenation synthetic technique based on TD-PSOLA method is suitable for synthesis Chinese. Proper



(a)



(b)

Fig.4. Some results of evaluation

controlling of speech prosody is the key point for improveing the speech quality. Doing more reseach work on the prosody of Chinese and improving the prosodic rules are our main work of next stage.

REFERENCES

[1] Lu Shinan, Qi Shiqian and Zhang Jialu (1993), "Experimental study on the naturalness of synthetic speech", *Chinese Journal of Acoustics*, Vol.12, No.3, P.256-264

[2]F. Charpentier, M. Stella (1986), "Diphone ynthesis using an overlap_add technique for speech waveforms concatenation", *Proc. Int. Conf. ASSP*, P.2015-2018

[3]Y. R. Zhao (1933) ,"Tone and intonation in Chinese", *Shi Yu Suo Ji Kan*, Vol.4, No.2, P.121-134

[4]Shen Jiong (1983), "Pitch range of tone and intonation in Beijing dialect", Working papers in experrimental phonetics, P. 73-120 (In Chinese)

ARTICULATORY SYNTHESIS USING A STOCHASTIC TARGET MODEL OF SPEECH PRODUCTION

Gordon Ramsay and Li Deng
Dept. of Electrical & Computer Engineering, University of Waterloo, Canada.

ABSTRACT

A stochastic target model for articulatory synthesis is described, where articulator motion is modelled by a linear system driven by random target functions and modulated by a finite-state Markov model. States in the model represent overlapping phonological units, while probability distributions for the associated target regions represent systematic articulatory variation. Simple examples of random synthetic speech are given.

INTRODUCTION

Speech recognition and synthesis have traditionally been approached through entirely different methodologies; the former based largely on *trainable models* which reflect the statistical characteristics, but not the mechanisms, of real speech; the latter based on *rule-driven systems* which incorporate a great deal of a-priori knowledge about speech mechanisms, but which lack the ability to adapt automatically to a particular corpus of data. In neither of these approaches is it usually possible to model articulatory phenomena directly, due to the lack of an appropriate articulatory representation.

In a previous presentation, a framework for articulatory speech recognition was outlined, based on a stochastic target model of speech production constructed around explicit articulatory and acoustic models of the vocal tract [1]. It was shown that the parameters of the model can, in theory, be trained automatically from a corpus of acoustic data using the EM algorithm.

In this paper, it is shown that the model can also be used for speech synthesis by sampling the underlying probability space. The resulting output incorporates a degree of non-deterministic but systematic variation, reflecting some of the possibilities for compensatory phenomena observed in real speech.

Simple examples of $(VCV)^+$ utterances are generated to demonstrate that the model is capable of producing plausible articulator and formant trajectories automatically.

MODEL STRUCTURE

Assume an underlying probability space $(\Omega, \mathcal{F}, P)$, and let $S = \{S_m : m \in \mathcal{N}\}$ be a finite-state Markov chain taking values in $\mathcal{S} = \{s_i : i = 1 \ldots N\}$, with transition matrix $\Pi = [\pi(i,j) : i,j \in \mathcal{S}]$, where $\pi(i,j) = P(S_{m+1} = j | S_m = i)$ and $\sum_j \pi(i,j) = 1$. Each state in the Markov chain represents a phonological symbol or, more generally, a combination of overlapping symbols, while any path through the state structure generates the symbol sequence for a particular utterance.

To describe the temporal characteristics of each phonological sequence, define a second process $T = \{T_m : m \in \mathcal{N}\}$. Each T_m represents the number of time frames spent in state S_m, and is assumed for convenience to be Poisson-distributed with parameter $\mu_\tau(S_m)$ drawn from a set $\mathcal{T} = \{\mu_\tau(i) \in \mathcal{R} : i \in \mathcal{S}\}$ according to the Markov state. The T_m are conditionally independent given S.

Each phonological symbol is assumed to possess a number of underlying physical correlates, which may be articulatory, acoustic or perceptual in nature. The fundamental modelling assumption is that the set of correlates for each symbol can be projected onto an equivalent *target region* in a Euclidean space of articulatory parameters $\mathcal{X} = \mathcal{R}^p$. The target region associated with any individual symbol can then be modelled as a distribution function on $\mathcal{X}$, giving the probability that any particular vocal tract configuration in $\mathcal{X}$ is capable of realizing the phonetic correlates associated with that symbol. Every time a transition occurs in the Markov chain, a new target configuration is chosen according to the target distribution for the new state, and held constant until the next state transition occurs.

To represent this, define a target process $U = \{U_m : m \in \mathcal{N}\}$ taking values in $\mathcal{X}$, where the U_m are independent conditioned on S, and each U_m is Gaussian-distributed with mean $\mu_u(S_m)$ and covariance matrix $\Sigma_u(S_m)$, selected from a set of target parameters $\Theta = \{(\mu_u(i) \in \mathcal{R}^p, \Sigma_u(i) \in \mathcal{R}^{p \times p}) : i \in \mathcal{S}\}$ according to the Markov state S_m. The extension to arbitrary continuous distributions on $\mathcal{X}$ is straightforward by approximation using Gaussian mixtures, and more complicated parameterizations are clearly possible where the target distributions interact or are made to vary with time.

The processes S,T,U then describe the generation of a random distribution of vector-valued target functions in articulatory space for a class of phonological state sequences.

Now, let $X = \{X_n : n \in \mathcal{N}\}$ be a random process on $\mathcal{X}$ representing the articulatory state, and let $Y = \{Y_n : n \in \mathcal{N}\}$ be a measurement process generating observations of X in an acoustic space $\mathcal{Y} = \mathcal{R}^q$. Assume that the initial state X_1 is distributed as $N(\mu_1 \epsilon \mathcal{R}^p, \Sigma_1 \epsilon \mathcal{R}^{p \times p})$ and define zero-mean Gaussian i.i.d. processes $V = \{V_n : n \in \mathcal{N}\}$ and $W = \{W_n : n \in \mathcal{N}\}$ to represent unmodelled perturbations in $\mathcal{X}$ and $\mathcal{Y}$, with covariance matrices $\Sigma_{vv} \epsilon \mathcal{R}^{p \times p}$ and $\Sigma_{ww} \epsilon \mathcal{R}^{q \times q}$ respectively.

Assume furthermore that X evolves in time according to the linear difference equation (1) driven by U (cf. [2][3]), and that Y is generated from X through a memoryless non-linear transformation $h : \mathcal{X} \rightarrow \mathcal{Y}$ as seen in (2).

Here d is the order of the system, and the matrices $A_i(j) \in \mathcal{R}^{p \times p}$ are selected from a set of system parameters $\mathcal{A} = \{A_i(j) : i = 1 \ldots d, j \in \mathcal{S}\}$ according to S_m. Since each control state S_m influences T_m frames of the articulatory process X, a random index function $J : \Omega \times \mathcal{N} \rightarrow \mathcal{N}$ is needed to cross-reference points in (S,T,U) and X.

$$X_{n+1} = \sum_{j=1}^{d-1} A_j(S_{J(n)}) X_{n+1-j} + A_d(S_{J(n)}) U_{J(n)} + V_n, \quad (1)$$

$$Y_n = h(X_n) + W_n. \quad (2)$$

This completes the description of the overall model structure. The function $h(.)$ represents the articulatory-acoustic mapping, and can be approximated using a codebook of points simulated from an acoustic model of the vocal tract. Provided that the phonetic correlates chosen for each phonological state can be expressed in terms of quantities which can be measured from model simulations, the corresponding target distributions can easily be derived from the codebook by defining an appropriate normalized cost function on articulatory space. Initial estimates for the duration parameters and the time constants of the state recursions can be measured from acoustic or articulatory data.

SIMULATION RESULTS

The model can now be used for speech synthesis by randomly generating sample paths from the probability space, using a Monte-Carlo technique. A state path through the Markov chain is first selected according to the transition matrix. Once the state path has been chosen, corresponding durations and target points are generated as a sequence of independent random variables with distributions determined by S. The articulatory state is Gaussian conditioned on (S,T,U), with mean and covariance that can be calculated recursively from the initial distribution for X_1, the target sequence U, and the sequence of system matrices defined by the Markov state path. Once the distribution of X is known, synthetic speech

can be obtained by generating a single random sample path of X and passing the result through the acoustic model.

Figure 1 shows a simple state structure representing $(VCV)^+$ utterances for the vowels /a/,/i/,/u/ and consonants /r/,/d/,/w/. Associated with each state is a target region representing the appropriate oral stricture. For the vowels, the region is characterized by a set of constraints on the first two formants, for example $/a/ = \{600 < F1 < 1000, 1000 < F2 < F1 + 500\}$, $/i/ = \{F1 < 400, F2 > 2000\}$, $/u/ = \{F1 < 400, F2 < 1000\}$ (in Hz), together with a requirement that the formant energy be greater than a sonorant threshold. For consonants, a mixture of acoustic and articulatory correlates are used. /r/ is defined by $\{F3 - F2 < 400, F2 < 1900\}$, /d/ by a closure along the alveolar ridge with $\{1600 < F2 < 1900\}$, while /w/ requires protrusion of the lips with $\{F1 < 300, F2 < 500\}$. The states are intended to represent combinations of abstract units, but these need not necessarily be segmental (cf. [4]).

Six parameters of a version of Mermelstein's model [5], shown in Figure 2, were chosen to form the dimensions of the articulatory space. A small codebook of 25000 entries generated from a finite-difference solution of the wave equations was used to provide measurements of the formants, average acoustic energy, and constriction location for a uniform distribution of points on $\mathcal{X}$.

The articulatory image of each target was then constructed by fitting a single Gaussian distribution to the class of all points satisfying the appropriate definition. Figure 3 shows two different projections of the sample distribution for /d/, illustrating some of the correlation patterns which arise automatically from this technique.

Figures 4 and 5 show spectrograms of two typical utterances produced from sample paths of X using an articulatory synthesizer, together with the parameter traces for X (−) and U (···). Realistic formant trajectories have been produced using only the statistical properties of target regions derived from relatively abstract and flexible phonetic specifications.

CONCLUSIONS

A stochastic target model for articulatory synthesis has been outlined, based on Monte-Carlo simulation of a Markov-modulated linear system. The model permits a compact articulatory representation of speech in terms of a relatively small number of statistical parameters which, in theory, it should eventually be possible to train from a corpus of acoustic data. In conjunction with existing filtering algorithms, the same modelling framework may also be used for speech recognition. Simple examples of synthetic VCV utterances have been constructed, and demonstrate that the model is indeed capable of reproducing many of the characteristics of real speech, although the quality does not at present approach that of formant synthesis. Future work will concentrate on improving the underlying model and adapting it to real data.

REFERENCES

[1] Ramsay G., Deng L., (1994) "A Stochastic Framework for Articulatory Speech Recognition," *JASA* 95 (5) Pt.2 Abstract 2aSP19.

[2] Saltzman, E. L., Munhall, K. G. (1989), " A dynamical approach to gestural patterning in speech production," *Ecological Psychology* 1 (4) pp. 333-382.

[3] Shirai K., Honda M., (1976) "Estimation of articulatory motion," in *Dynamic Aspects of Speech Production*, University of Tokyo Press.

[4] Browman C.P., Goldstein L., (1992) "Articulatory phonology : an overview," *Phonetica* 49 pp. 155–180.

[5] Rubin, P., Baer, T., Mermelstein, P. (1981), "An articulatory synthesizer for perceptual research," *JASA* 70 pp. 321-328.

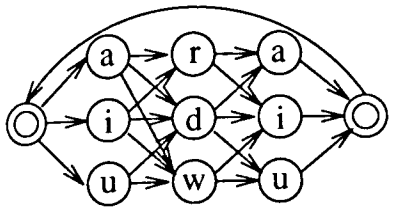


Figure 1: $(VCV)^+$ state structure.

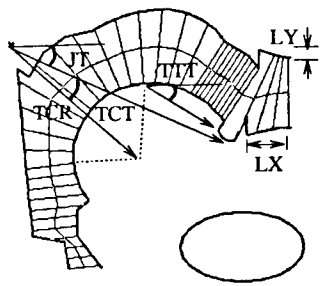


Figure 2: Articulatory model.

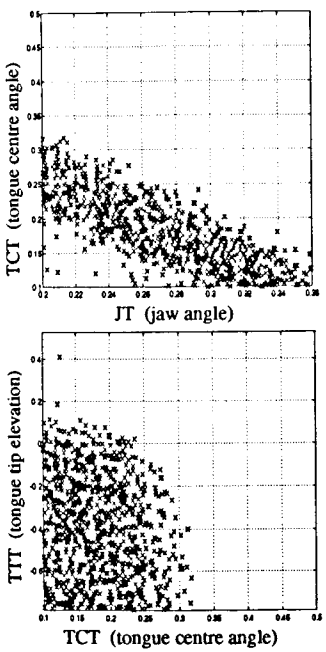


Figure 3: /d/-target projections.

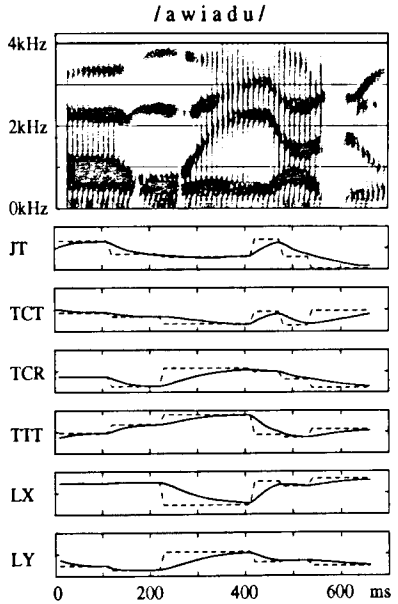


Figure 4: Random synthetic speech

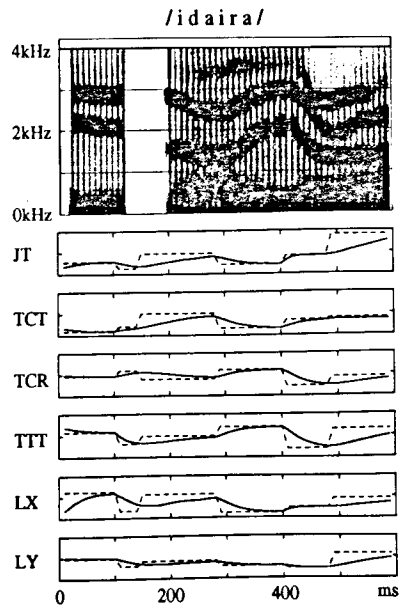


Figure 5: Random synthetic speech

PHONETIC REALISATION OF PROSODIC BOUNDARIES IN SYNTHETIC SPEECH

Angelien Sanderman
Institute for Perception Research / IPO, Eindhoven, The Netherlands

ABSTRACT

In this research different sets of prosodic boundary rules were developed and their acceptability was evaluated. The results show that the rule set distinguishing 5 levels of boundary strength produced synthetic speech that is as acceptable as its natural counterpart.

INTRODUCTION

Although synthetic speech is often quite intelligible, it sounds unnatural and is not always easy to comprehend [3]. To improve the quality of synthetic speech it is important to provide it with appropriate prosody. The research reported on here is concerned with the demarcative function of prosody at the sentence level: it studies how prosody is used to group words into phrases.

From previous research [4, 2] we know that speakers use pause, melodic marker and declination reset systematically to indicate various degrees of prosodic boundary strength between words. The perceptual relevance of these features is evident from the fact that listeners systematically assign different perceptual boundary strength-values (PBS) on a 10-point scale to word boundaries differing in such prosodic characteristics. They can even do so when the speech is made unintelligible, so that they can not rely on syntactic or semantic information. In other words, listeners use the phonetic cues used by the speaker to determine the degree of disjuncture in the flow of speech.

The results of previous experiments were used to develop boundary rules for synthetic speech to generate well-phrased utterances. These rules describe how to realize phonetically the prosodic boundaries of three different strengths (zero, minor and major). The results of this pilot experiment showed that listeners had a significant preference for the realizations *with* the boundary rules over the realizations *without* the boundary rules [5].

The decision to distinguish three levels of boundary strength was made on the basis of practical considerations and was inspired by ideas derived from prosodic phonology [1]. In this research we want to explore whether distinguishing more than three boundary levels will lead to a further improvement of synthetic speech quality.

METHOD

Rules

To develop boundary rules, we used the results of previous experiments. From these results we know that pause, melodic marker and declination reset are important prosodic cues in the phrasing of utterances. From these studies, we can predict with a 90 % success rate what the PBS-value on the 10-point scale will be, given the prosodic cues. Conversely, when we know the PBS we know what the prosodic cues will be. This outcome was taken as the basis for developing 8 different sets of boundary marking rules, four of which are described here.

First of all, there was a version without boundary rules. This means that all the boundaries were realised with neither a pause, a declination reset nor melodic marker.

Secondly, a set of boundary rules distinguishing three levels of boundary strength was implemented. These three levels of boundary strength do not map onto the PBS, which are expressed on a 10-point scale. Therefore, the observed PBS values were clustered into three classes, which resulted in the recipes given in Table 1. The rule applying to parenthetical clauses and to 30 % of the nonrestrictive clauses is not mentioned in the table. It is as follows: there is a downward reset, with a pause shorter than 200 ms and mostly a pitch contour of the type rise-fall-rise.

Table 1: Recipes for three levels of boundary strength (1 is the weakest boundary and 3 the strongest). The probability of occurrence of each option is given between brackets.

	Contour type	Pause (ms)	Reset
1	no marker (1.0)	0 (1.0)	no(1.0)
2	no marker (.05) (.07) (.12) (.32) (.44)	0 (.45) 50 (.55)	no(1.0)
3	(.09) (.10) (.16) (.65)	150 (.16) 250 (.44) 350 (.27) 450 (.13)	no(.23) reset (.77)

Thirdly, a set of boundary rules with five levels was developed. To determine the prosodic realisation it was again necessary to cluster the possibilities of the 10-point scale, this time into 5 classes. This resulted in the recipes given in Table 2. The rule for parenthetical clauses and nonrestrictive clauses is the same as in the set of rules with three levels.

Finally, to assess the relative success of the different sets of boundary rules, a natural version was included: the pause durations and pitch patterns used by a professional speaker were copied onto the synthesized speech.

Table 2: Recipes for five levels of boundary strength (1 is the weakest and 5 is the strongest). The probability of occurrence of each option is given between brackets.

	Contour type	Pause (ms)	Reset
1	no marker (1.0)	0 (1.0)	no (1.0)
2	(.12) (.19) (.34) (.35)	0 (1.0)	no (1.0)
3	no marker (.09) (.04) (.06) (.29) (.52)	50 (1.0)	no (1.0)
4	(.10) (.10) (.18) (.68)	150 (.18) 250 (.50) 350 (.32)	no (.24) reset (.76)
5	(.07) (.07) (.86)	450 (1.0)	no (.20) reset (.80)

Material and Procedure

12 Dutch sentences, ranging in length between 6 and 26 words and varying in syntactic structure, were synthesized [7]. To determine the place of accents and the position and strength of boundaries, a professional speaker was asked to read out the sentences and listeners were asked to assign boundary strengths between all pairs of words. Then, the four sets of prosodic rules were applied. It can be seen from Table 1 and 2 that there are several possibilities to realize a boundary of a certain strength. In those cases a weighted random choice was made from the set of possibilities. Therefore, the rules for three and five levels were implemented three times each on every sentence to see if different random choices from the possible prosodic realizations would affect the results (in fact, they did not). This resulted in a total of 12 x 8 = 96 stimuli. The total set contained actually 156 stimuli, since there were more sets of rules, not described here.

The 96 sentences were synthesized by means of diphone concatenation [7]. The phonetic realizations of the various contours, such as the size and slope of pitch movements, slope of the declination line, start and end frequencies, were based on rules given in Terken [6].

In a perceptual experiment, the 156 stimuli were presented to 18 untrained Dutch listeners, who scored the acceptability on a 10-point scale, in 3 sessions. Each session contained 2 blocks and each block contained 2 sentences with its different realisations. The combinations of sentences, the combination of blocks and the sessions were randomized.

RESULTS

Figure 1 shows the mean acceptability scores of the 12 sentences for the 4 sets of boundary rules out of 8. The mean scores for the 8 rule sets ranged from 4.4 to 7.3 on the 10-point scale and differed significantly from each other; $F_{(7,2789)} = 103.38$ ($p<.0001$).

As can be seen from this figure, the natural version scored highest (7.3) followed by the rule set with 5 levels of boundary strength (6.8). These two versions did not differ significantly from each other. This means that the rule set distinguishing 5 levels of boundary strength is very acceptable.

The version with three levels of boundary strength scored 6.1 and differed significantly from the other versions plotted in Figure 1. In turn, this version scored much higher than that without boundary rules (4.4). This is in agreement with the pilot experiment [5].

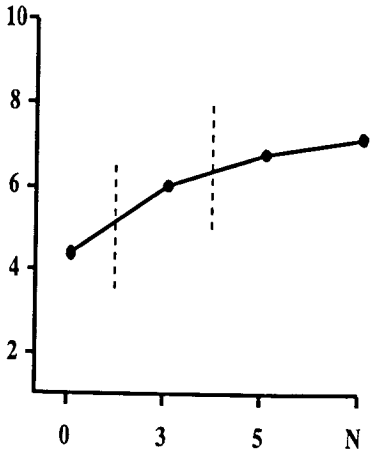


Figure 1: The mean score of acceptability for the four sets of boundary rules: 0 (version without boundary rules), 3 (version with 3 levels), 5 (version with 5 levels) and N (natural version).

DISCUSSION

From the results we can conclude that the rule set distinguishing 5 levels of boundary strength produced speech as acceptable as the natural version. These two versions did not differ significantly.

The rule set with 5 levels is more acceptable than that with three, but both improved the synthetic speech quality in comparison to the version without boundary rules. Clearly, listeners find a well-phrased utterance much more acceptable than a poorly-phrased one. The results also confirm that the prosodic cues pause, melodic marker and declination reset are appropriate to mark boundaries.

The question remains whether listeners comprehend these well-phrased utterances more easily compared to poorly-phrased ones. Follow-up research is underway to explore this question.

ACKNOWLEDGMENTS

Many thanks are due to R. Collier, J-R de Pijper and M. Swerts for commenting upon an earlier version of this paper.

REFERENCES

[1] Dirksen, A. (1992). Accenting and deaccenting: a declaritive approach, in Proceedings of the 15th International Conference on Computational Linguistics, COLING '92, 3., pp 865-869.

[2] Pijper, J.R. de & Sanderman, A.A. (1994). On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues. *Journal of the Acoustical Society of America*, 96 (4), 2037-2047.

[3] Pisoni, D.B, Manous, L.M. & Dedina, M.J. (1987). Comprehension of natural and synthetic speech: effects of predictability on the verification of sentences controlled for intelligibility. *Computer Speech and Language*, 2, 303-320.

[4] Sanderman, A.A. & Collier, R. (1994). Prosodic phrasing at the sentence level. *Festschrift for K. Harris of Physics, Modern Acoustics and Signal Processing Series*. American Institute of Physics.

[5] Sanderman, A.A. (1994). How can prosody segment the flow of (synthetic) speech? *Conference Proceedings of the second ESCA/IEEE Workshop on Speech Synthesis*, pp 147-150.

[6] Terken, J.M.B. (1993). Synthesizing natural sounding intonation for Dutch: rules and perceptual evaluation. *Computer Speech and Language*, 7, 27-48.

[7] Van Rijnsoever, P.A. (1988). A multilingual text-to-speech system. *Annual Progress Report No.23*. Institute for Perception Research, Eindhoven.

LACS: LABEL ASSISTED COPY SYNTHESIS

M. Scheffers and A. Simpson
IPDS, Kiel, Germany

ABSTRACT

We describe a knowledge-based method of deriving the control signals for the Klatt synthesizer. This method uses a combination of a rich acoustic analysis and an intelligent post-processing system which employs labels from the segmentation of the original signals to modify and augment the analysis data.

INTRODUCTION

The automatic generation of control signals to a drive a formant synthesizer offers an excellent method of validating phonological models by observing their phonetic output. This is made all the more challenging by the high quality of the speech which formant synthesis can produce when provided with appropriate control signals.

A synthesizer such as the early Klatt model [1] offers a large number of control parameters allowing adequate modelling of the acoustic products of the vocal tract.

However, obtaining parametric values which are to serve as the phonetic correlates of the phonological systems and structures of a language is a laborious task. One of the most interesting and enlightening methods of arriving at these numbers is undoubtedly copy synthesis [2], i.e. driving a synthesizer with the results of the analysis of a natural utterance.

There are, however, two serious problems involved in mapping the results of an acoustic analysis onto the control parameters of the Klatt formant synthesizer.

First, there is a discrepancy between the information delivered by the acoustic analysis of an utterance and the rich variety of synthesizer parameters which can be used to model the acoustic signal. Most acoustic analyses, for example, only allow a decision to be made as to whether the resonators should be excited with a periodic (Fo found) or an aperiodic source (no Fo found). The Klatt synthesizer, on the other hand, offers four dynamic parameters which can be used to model glottal and supraglottal sources, two parameters to model the aperiodic (glottal and supraglottal friction) and two for the periodic source (voicing and a low-pass filter to model the voicing in voiced stops and fricatives).

Second, parametric information about more complex products of the vocal tract is usually not available in the analysis. Voiced fricatives are an example of this. A voiced fricative such as [z] leaves an analysis either as a voiceless fricative (no Fo found) or as a frictionless approximant (Fo found). Although the former may be the most appropriate analytical outcome for a synthetic utterance, neither allows the original fricative to be modelled. Breathy voice presents a similar problem.

In this paper we would like to describe a method which attempts to overcome these problems by subjecting analysis data to an intelligent post-processing based on the manual segmentation and labelling of the original signals.

An acoustic data base of read speech such as that constructed at the IPDS Kiel [3] provides an excellent source of segmented and labelled signals. This data base contains 31374 segmented and labelled words of German spoken prose. Labels are primarily phonological in nature. Each phonological label is time-aligned with the start of a signal portion representing the chief phonetic correlates of the phonological item in question. Phonological labels are supplemented by quasi-phonetic labels to indicate aspects such as creaky voice, plosive release phases and vowel nasalization when other correlates of a nasal are absent.

ANALYSIS

We begin by describing the analysis method.

The short-term energy (RMS), Fo and formant analysis facilities of the ASSP signal processing package developed at Kiel [4] are used to obtain initial estimates of the parameters for the Klatt synthesizer. Fo values are analysed using an extremely fast and highly accurate periodicity detector [5,6]. Formant frequencies and bandwidths are determined by root-solving of the LPC polynomial [7,8]. Subsequently, formant amplitudes are estimated from the LPC spectrum. The speech signals being sampled at 16 kHz, 8 formants are analysed for male voices.

Conditioning and Sorting

Because the formant analysis always provides the number of formants specified, it must introduce pseudo-formants when there are fewer resonances. One class of pseudo-formants results from real roots. These receive a fixed, very high or low frequency and/or an extremely large bandwidth. The other pseudo-formants are also characterized by a very large bandwidth and occur either about midway between two "true" formants or very near to one.

Whereas these pseudo-formants are accommodated for in the synthesis model applied in ASSP, the discontinuities they cause in the formant tracks are disruptive in the Klatt model and hamper (semi-)automatic processing and interpretation of the data. The raw data are therefore sent through a conditioning stage: *ksort*.

First, the pseudo-formants resulting from real roots are removed. The bandwidth of the other formants is checked against a threshold, currently set at 1000 Hz. If it is above threshold, a set of heuristic rules is invoked to decide whether the formant is to be deleted or to be merged with a nearby formant (weighted mean).

Next, each formant is assigned a best fitting formant number by comparing its frequency with a list of average formant frequency values. When two formants receive the same number a more global best match is searched in which as few formants as possible obtain numbers different from the ones originally assigned to them.

Finally, gaps in the resulting formant tracks are filled with dummy values which can easily be identified by the next processing stage.

Analysis Results

The Fo analysis lives up to its reputation: in the 100 sentences currently under study, no gross errors were found. The few errors made mainly consist of:

- delayed voicing detection due to irregularities in the initial glottal pulses,
- failure to detect creaky voice,
- failure to detect stretches of weak and noisy voicing often found in utterance final syllables.

If these Fo errors are found to detract seriously from the quality of the synthetic utterance, as can happen at voiced-voiceless boundaries, they can easily be manually corrected. Failure to detect creaky voice is an exception to this and is one of the areas where label information can be successfully used for automatic correction (see below).

For voiced sounds, the lower formants are generally consistently found and numbered correctly. Keeping in mind that we only need the lower four formants for these sounds, these data can directly be used in the synthesis. Exceptions are typically nasals and nasalized vowels, where an additional nasal formant at about 2.5 kHz is detected. For nasals, this presents few problems since the discontinuities are aligned with the nasal closure and release. In fact, nasals come out quite nicely in the synthesis. Nasalized vowels pose a bigger problem because the formant sorting goes awry.

For unvoiced sounds, there are more diverse problems. First of all F1 data are rarely found and in many cases F2 data are also absent. Second, the scatter often found in the formant data that are present makes it difficult to properly number the formants. Although the absence of lower

formants may seem to pose no problems because the corresponding resonators are not excited in the synthesis, the Klatt model does use their frequency values to adjust the amplitudes for the higher ones.

We have recently started experimenting with a different kind of analysis, the so-called 'Robust Formant Analysis' [9]. As with root-solving, it delivers the number of formants specified, but the formant tracks are virtually continuous. Some other properties of the data obtained by this analysis are:

- F1 is continuously present and has a reasonable course.
- F2 corresponds quite closely to the values found by root-solving or peak-picking.
- In closures the data tend towards those of an open tube rather than scatter as in the other analyses.

However, since formants are defined purely operationally in this analysis and need not correspond to resonances in the spectrum, we observed that especially in the mid-frequency region (roughly 2 to 4 kHz) resonances are often represented by two "formants". Since their frequencies are rarely close, it is nearly impossible to detect this and merge the data. Presently, we are looking for ways to combine the results of the two analyses using the strength of each to compensate for their respective weaknesses.

POST-PROCESSING

The first pass through the data delivered by *ksort* ensures that any gaps in the formant tracks are filled in as harmless a fashion as possible. In general, this entails nothing more than carrying out a simple interpolation between two formant values. Furthermore, normalization of the RMS values is carried out and formant amplitudes are modified to compensate for the corrections made in the Klatt model.

Next, the analysis data are combined with labels from the manual segmentation such that each analysis frame is associated with one label. In certain cases the labels which are in sequence in the data base are collapsed into one. So, for instance, creaky vowels are represented in the data base as a sequence of two labels, the first of which indicates the presence of creak over the following vowel. So that this information does not get lost the vowel label is suffixed with a creak marker.

The second pass through the data uses the information provided by the labels to map the analysis data onto synthesizer control parameters. Below, we present three examples of the way in which label and analysis information can be successfully combined to exploit to the full the control parameters made available in the synthesis model. All the examples deal with different aspects of mapping analysis data onto control parameters for the source signals.

Creaky Voice

Portions of creaky voice are generally not found in the Fo analysis. If the label information indicates that a vowel is creaked, but frames have been declared unvoiced in the Fo analysis, creak is modelled by inserting random low Fo values from the vowel onset until the first voiced frame is found. Although it is not possible to model many aspects of creaky voice in the Klatt model we are using, such Fo values together with the fluctuating amplitude of voicing derived from the RMS values produce perceptually acceptable creaky voice.

h and its Correlates

Voiceless signal portions annotated with **h** are assumed to represent periods of turbulent airflow originating at the glottis. These are modelled by mapping the RMS value onto the amplitude of aspiration. Frames labelled with **h** and returned as voiced from the Fo analysis are considered to be periods of breathy voice. The RMS value is used to set the amplitude of voicing. Following [1,10], values for the parameters for the amplitude of aspiration and the amplitude of sinusoidal voicing are derived by subtracting 3 dB and 6 dB, respectively, from the voicing value.

Plosive release and aspiration

The label **-h** annotates a signal portion from the plosive release to the end of any aspiration. The place of articulation of the release is derived from the preceding plosive label and the following vowel. The release is modelled by using RMS values to set amplitude values for the supraglottal fricative source. Once the burst and initial release phases have passed, the RMS values are mapped onto the aspiration source. The length of the supraglottal and glottal friction are varied with the place of articulation of the plosive, the local friction being maintained longest for dorsal plosives.

DISCUSSION

The main aim of the copy synthesis method described here is to derive parameters for rule-driven synthesis. Using phonological/phonetic information allows us to be very selective in the way in which we modify the analysis data to arrive at synthesis parameters.

The advantages of this approach are manifold. Analysing, processing and synthesizing a large number of utterances is fast. Auditory inspection quickly identifies naturally sounding stretches of synthetic utterance. These are places where we can assume that the parameter courses can be used to derive the correlates for the rule-driven synthesis.

The modifications we can carry out on the basis of label information are wide-ranging. They can reflect the findings of others, e.g. the RMS mappings in breathy voice which are based on numbers taken directly from the literature. Other modifications can represent a step-by-step idealization of the analysis data. This is especially desirable when working towards easily definable parameter courses in synthesis-by-rule. The process of idealization can be gradual, allowing the consequences of each step on the naturalness and acceptability of the resulting signals to be assessed. The ultimate modification is to completely discard analysis data for difficult portions, such as voiceless fricatives, and insert numbers obtained elsewhere[1].

The next step in our work will be to investigate the advantages of using label information already at the analysis stage.

ACKNOWLEDGEMENT

We would like to thank Lei Willems for providing us with the software of the robust formant analysis.

REFERENCES

[1] Klatt, D.H. (1980), "Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Am.*, vol. 67, pp. 971-995.

[2] Holmes, W.J. (1989), "Copy synthesis of female speech using the JSRU parallel formant synthesizer", *Proc. EUROSPEECH*, vol. 2, pp. 513-516.

[3] IPDS (1994), *CD-ROM#1: The Kiel Corpus of Read Speech*, vol. I, Kiel: IPDS.

[4] Scheffers, M., Thon, W. (1991), "Workstation and signal processing software for experimental phonetics", *Proc. XIIth ICPhS*, vol. 2, pp. 486-489.

[5] Schaefer-Vincent, K. (1982), "Significant points: Pitch period detection as a problem of segmentation", *Phonetica*, vol. 39, pp. 241-253.

[6] Schaefer-Vincent, K. (1983), "Pitch period detection and chaining: Method and evaluation", *Phonetica*, vol. 40, pp. 177-202.

[7] Vogten, L.M. (1983), *Analyse, zuinige codering en resynthese van spraakgeluid*, doctoral thesis, Eindhoven University of Technology.

[8] Saito, S., Nakata, K. (1985), *Fundamentals of speech signal processing*, Tokyo/Orlando/London: Academic Press.

[9] Willems, L.F. (1987), "Robust formant analysis for speech synthesis", *Proc. Eur. Conf. Speech Technology*, vol.1, pp. 250-253.

[10] Allen, J., Hunnicutt, M.S., Klatt, D. (1987), *From text to speech: The MITalk system*, Cambridge: CUP.

---

[1] an idea suggested to us by John Local.

A TOOL FOR THE COMPLETE PRODUCTION OF COPY SYNTHESES FROM NATURAL TOKENS

A.M.Simpson
Department of Phonetics and Linguistics, University College London

ABSTRACT

An X Windows graphical user interface for the Klatt Cascade-Parallel Formant Synthesiser [1] is described. It includes facilities for initial synthesiser parameter estimation, for editing time-varying parameter values, and for aural, spectral, and spectrographic comparison of target and copy-synthesised stimuli.

COPY SYNTHESIS

High-quality copy synthesis allows the creation of speech-like stimuli which are sufficiently natural to ensure listeners are listening in the speech mode. The stimuli can be altered to investigate the relative contribution of different acoustic cues in encoding phonetic contrasts.

Formant synthesis produces speech-like stimuli by passing a source waveform through a complex filter whose resonances model those of the vocal tract. The precise nature of the source can be varied (e.g. periodicity, voice fundamental frequency, open quotient, jitter) as can the characteristics of the filter (e.g. formant frequencies, bandwidths, and amplitudes, and the presence of anti-resonances). This flexibility allows the complex spectro-temporal variation that occurs in natural speech to be closely modelled.

However, the specification of formant synthesiser parameter values is a complex and laborious process. Even with good initial estimates of the voice fundamental frequency contour and the formant trajectories, much work still has to be undertaken to model the variation in formant amplitudes and bandwidths before a synthetic copy will sound natural.

The task of refining synthesiser parameter values is made difficult by the inability to visualise how synthesiser parameters co-vary, to edit them easily, and to assess easily the effects of any manipulations both aurally and by using more objective analysis methods.

This tool addresses such difficulties by providing all such facilities within an integrated package which includes a version of the Klatt Cascade/Parallel Formant Synthesiser. It enables users to calculate initial parameter values from the target stimulus, to edit such values easily, to see and hear both natural target and its synthetic copy, and to perform spectral and spectrographic analysis of both to ensure closeness of match.

INITIAL PARAMETER ESTIMATION

The tool includes the facility to calculate the voice fundamental contour from the target natural token. In addition, it is possible to specify formant frequency trajectories by tracing each formant's path onto a spectrogram

PARAMETER EDITING

Time-varying parameters are edited using a 'canvas' onto which each parameter's trajectory can be drawn using a mouse-controlled cursor. Any number of parameters can be simultaneously displayed to allow users to co-ordinate the value of parameters which vary time-synchronously, for example, the formant amplitudes at the onset of voicing after plosive release, or at a plosive's release burst. Each parameter is displayed using a different colour to ensure it is distinguishable from others; parameters can be cycled through three states: being edited, displayed only, and not displayed, allowing arbitrary groups of parameters to be simultaneously displayed and edited. Figure 1 shows the amplitude of the first formant being edited, whilst its frequency trajectory is also displayed.

It is possible to specify parameter values with great accuracy as the precise parameter value at the cursor is displayed. The parameter canvas is time-aligned with both target and copy-synthesised waveforms and vertical cursors indicating the point in time being considered are displayed in all windows, allowing alignment between the two waveforms, and allowing the monitoring

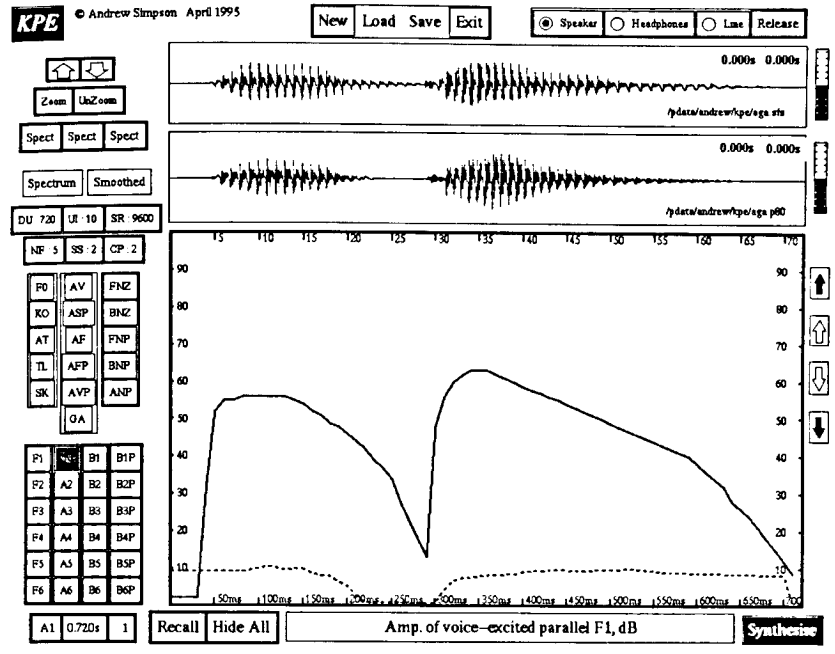


Figure 1. Natural (upper) and synthetic waveforms of /aga/ together with the values of A1 (amplitude of voice-excited parallel first formant, solid line) and F1 (first formant frequency, dotted line). Any number of the time-varying parameters can be displayed.

of the effect parameter changes have on the synthetic waveform. A parameter's trajectory can either be specified by drawing a line between start and end points, or can be drawn free-hand. To facilitate inspection and editing of parameter values over very brief regions of the stimuli a zoom facility allows users to display a region in greater detail, allowing very brief acoustic cues such as release bursts to be inspected and specified with great accuracy. Figure 2 shows how the zoom facility has been used to display the burst and first few cycles after release of an intervocalic voice plosive.

COMPARISONS

Aural comparison of complete or partial target and copy-synthesised waveforms is possible simply by marking the extent of the desired region using mouse-controlled cursors and then clicking on the appropriate waveform. It is also possible to calculate the amplitude spectrum of corresponding regions of both waveforms thus providing more detailed information about their degree of similarity. This is illustrated in Figure 2 where the release burst spectra of a natural and synthetic intervocalic voiced velar plosive are compared. Although the strong burst at around 2 KHz has been modelled well, there are discrepancies around the secondary peak at 4 KHz and in the region below about 200 Hz. Such comparisons are useful for comparing short or relatively unchanging regions of the signal, or for making gross spectral comparisons

To assess how closely the complex spectro-temporal variation seen in speech has been modelled a spectrogram calculation facility is provided which can

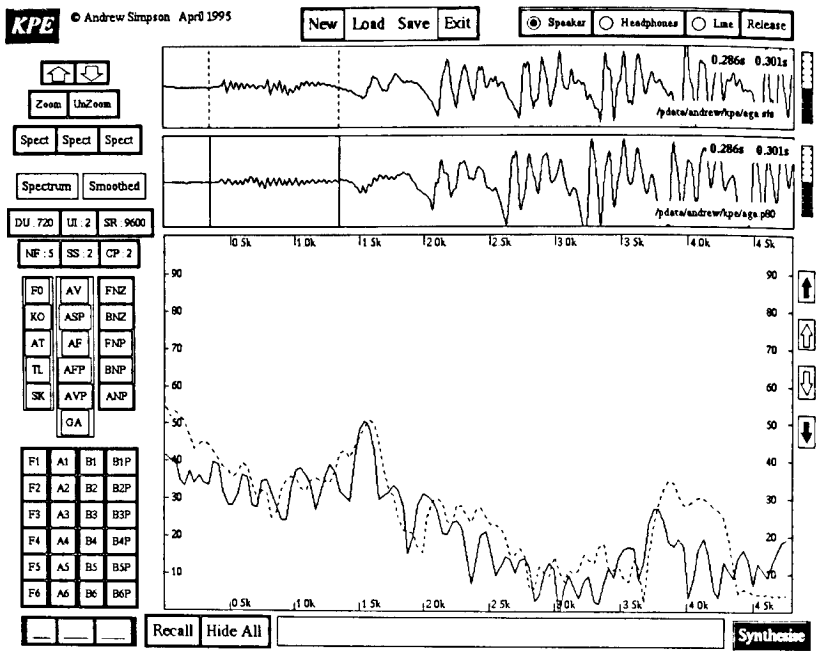


Figure 2. Detail of the burst and few cycles after release for natural (upper) and synthetic /aga/ stimuli. The amplitude spectra for the marked burst regions are compared below.

display either natural or synthetic waveforms, or both simultaneously, aligned in time to facilitate comparison.

FILE FORMATS

The tool supports a range of audio data file formats including Sun Audio format (.au), Microsoft audio format (.wav), and the SFS [3] format (.sfs). The synthesiser parameters are stored in the form of an ASCII text file with each parameter's value for each successive frame of the stimulus stored as a number.

PORTABILITY

The tool has been written in C using SUIT [2] and requires a UNIX platform with X Windows. Versions currently exist for Sun Sparc and Linux/XFree86 architectures. Plans exist to port the tool to other platforms. Contact the author (*andrew@phon.ucl.ac.uk*) for more details and for information about how to obtain the tool.

ACKNOWLEDGEMENTS

The formant synthesiser is an implementation of the Klatt Cascade-Parallel Formant Speech Synthesiser by Jon Iles (*j.p.iles@cs.bham.ac.uk*) and Nick Ing-Simmons (*nicki@lobby.ti.com*) The spectral, spectrographic, and pitch-extraction are from SFS [3].

REFERENCES

[1] Klatt, D.H. (1980), "Software for a cascade/parallel formant synthesiser", *Journal of the Acoustical Society of America*, vol. 67(3), pp. 971-995.

[2] SUIT, The Simple User Interface Toolkit, University of Virginia, *(suit@uvacs.cs.Virginia.EDU).*

[3] The Speech Filing System, Dept of Phonetics and Linguistics, University College London. (*sfs@phon.ucl.ac.uk*)

ON THE DEVELOPMENT OF TEXT TO SPEECH SYSTEM FOR HINDI

Rajesh Verma, A.Sada Siva Sarma, Nisheeth Shrotriya, Anil Kumar Sharma and S.S.Agrawal

Speech Technology Group, Central Electronics Engineering Research Institute Centre, CSIR Complex, Hill Side Road, New Delhi - 110 012.

ABSTRACT

Recently, the feasibility of using KLSYN88 has been shown to synthesize Hindi speech sounds including voiced/unvoiced and aspirated/ unaspirated stop consonants and trills with good quality [1-2]. The synthesizer was subsequently implemented on a PC(AT) and slightly modified to suit synthesis of Hindi aspirated consonants. Syllables were used to generate words after framing the joining rules and sentences were made using these words. This paper describes the feasibility and approach to develop a PC based Text To Speech (TTS) system for Hindi.

INTRODUCTION

Synthesizing high quality speech or natural sounding speech by machines/ computers has been a frontier research area for several decades. Efforts have been made to develop source coding techniques and modelling of the vocal tract system to achieve the above goal. During the past few years, major advances have been made with two terminal analogue formant synthesizers, the Klatt synthesizer [3] and the Holmes synthesizer, and both of them have been adopted for commercial use.

SPECIFIC FEATURES OF HINDI SOUNDS

The Hindi consonants possess certain special features which are not so common to European languages and American English [2]. The most significant differences are in stops and affricates which use both voicing and aspiration, to distinguish them from other languages. For this reason, the aspiration source of KLSYN88 has been modified to pull down the energy to 300 Hz onwards, to generate more natural sounding voiced/unvoiced aspirated sounds. The trills /r/ and /l/ have large allophonic variations in different contexts. Consonant clusters like CCV, CCCV, CCVC etc. also occur frequently in Hindi speech. It is therefore necessary to study these specific characterstics of clusters as a separate category.

TEXT TO SPEECH CONVERSION SYSTEM

Syllables have been chosen as the basic units of Hindi speech to generate an unlimited vocabulary in the proposed TTS system. The most frequently occurring 29 consonants and 10 vowels (5 long and 5 short) in Hindi give rise to 290 CV and 290 VC syllables in all, that form major part of the **database**. Experiments have been conducted to generate short vowels out of corresponding long vowels, using some durational and frequency rules. Therefore number of syllables to be generated have been limited to 290 only. In addition, a special class of around 150 clusters have been included as part of the **database**. Therefore about 500 basic units would be adequate to generate unlimited vocabulary.

The basic building blocks of a text to speech conversion system are shown in *figure 1*. Text Input from the keyboard is fed to the **word parser** . Therefore

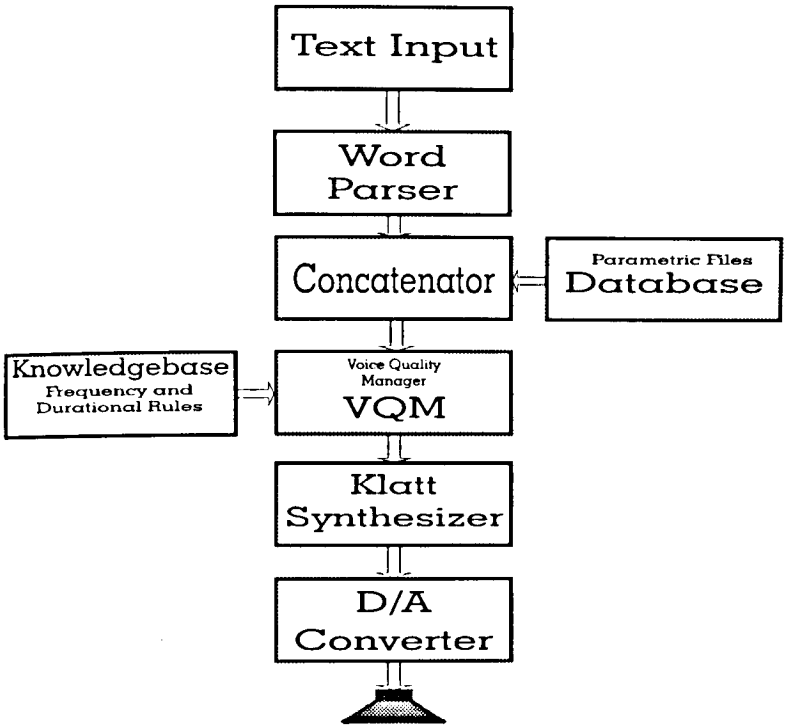


Figure 1. Block diagram of TTS system

word parser identifies the basic syllables, available in **database**, using which the words are generated. These file names are given to the **concatenator**, which picks up files from the **database** and merges them properly to make a new parametric file for the given input word.

Once the doc files are merged and a sound file is generated, the quality of sound is not very good because of discontinuities in sounds at the boundaries of the syllables. Therefore, a **knowlegebase** of durational and frequency rules is provided to voice quality manager **VQM** which applies these rules to the parametric file given by the **concatenator** to smooth out the discontinuities at the syllable boundaries. Finally this parametric file is sent to the **Klatt Synthesizer** to generate the sound file. This sound file is then fed to a loud speaker through a **D/A converter** card. A blank character or a punctuation mark acts as a delimiter for a word. Hence sound output is given word by word.

PROCEDURE FOR ANALYSIS AND SYNTHESIS

The 29 consonants have been recorded by single male speaker having a standard Hindi speaking background and mother tongue, directly in to a PC/AT 386 computer which is equipped with a DSP56000 based SENSIMETRICS speech

station hardware and software.

The syllables were analysed using KAY Sonograph, Sensimetrics speech station and CD_SPEC[4] analysis program having facilities of displaying LPC/FFT Spectrum of windowed segment, pitch and energy etc.. All the important source and vocal tract parameters were extracted and tabulated for use in synthesis.

Based on the analysed data, a preliminary parametric (DOC) file was created and synthesis was done to obtain a starting synthetic file. A Set of 60 parameters [2] have been used for creating a synthesized document file. Then the spectrograms of natural and synthetic syllables were compared. A number of source and tract parameters were adjusted iteratively, in order to achieve a close imitation to natural CV/VC syllables.

RESULTS

All the twenty-nine frequently occurring consonants in combination with five long vowels have been synthesized to obtain 145 CV and 145 VC combinations. Spectrograms showing the comparison of original and synthetic sounds for syllable /t ʃ a/ are shown in *figure 2.*

Experiments have been done to generate words made of simple CVCV type syllables. The words were fed through keyboard using equivalent ASCII codes of the syllables. Durational and frequency rules have been applied at the syllable boundaries to smooth the discontinuities. *Figure 3* shows three spectrograms of the word /t ʃ at ʃ a/. In this figure the first spectrogram does not include any rule, the middle spectrogram shows the application of durational rule

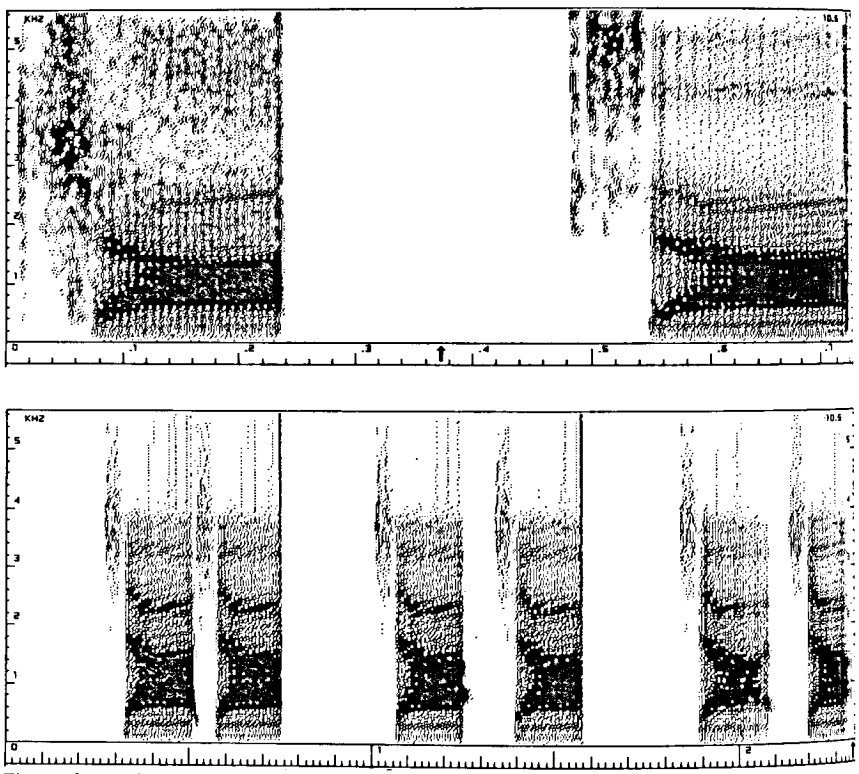


Figure 3. Application of durational and frequency rules to improve sound quality

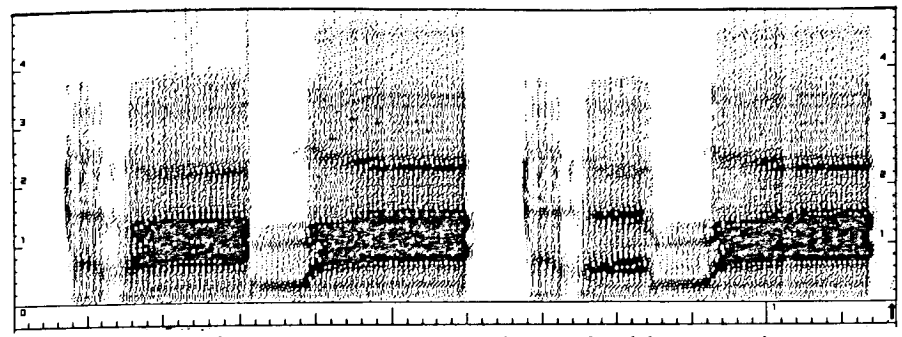


Figure 4. Coversion from /rama/ to /rʌma/ using durational and frequency rules

and the last spectrogram shows the effect of both durational as well as frequency rule to get a close imitation of the original sound.

Figure 4 shows the application of the rule for creating short vowel /ʌ/ from long vowel /a/. It shows the CVC type of word /rama/ and it is intended to make it /rʌma/. To achieve this two rules have to be applied. First, the vowel duration has to be reduced from 200 ms to 100 ms. Second, the first formant frequency is to be pulled down by 200 Hz and second formant frequency is to be pushed up by 200 Hz, in the pure vowel region. These rules lead to creation of vowel /ʌ/ from /a/. Different set of rules are being formed for different vowel contexts.

CONCLUSIONS

Feasibility of using KLATT synthesizer for close imitation of natural Hindi sounds have been shown. All the CV and VC syllables required for the **database** have been generated. Most frequently occurring clusters in Hindi have been selected. Work is under progress to synthesize those clusters as part of the **database. Word Parser** is currently working for words made of CV syllables only (maximum up to ten syllables) and is being updated. **Concatenator** block has been fully developed. **VQM** is currently working for few durational rules only. Some of the durational and frequency rules have been studied and further study is in progress for creation of **Knowledgebase.**

The study of the suprasegmental features, rules for intonation and stress patterns is in progress, for this system to act as a reading machine for blinds.

ACKNOWLEDGEMENTS

The authors are grateful to Prof R N Biswas, Director, CEERI, Pilani and DoE/UNDP for their encouragement and support. Award of SRF by CSIR to one of the author (NS) is highly acknowledged.

REFERENCES

[1] Agrawal S. S., and Stevens K., "Towards synthesis of Hindi consonants using KLSYN88," Proc. ICSLP-92, 177-180, 1992.

[2] Neal Pinto, et.al.,"Synthesis of Hindi CV Syllables in Three Vowel Contexts using the PC-KLATT Cascade/ Parallel Formant Synthesizer", Proc. 3rd ICAPRDT-93, ISI, Calcutta, India, Dec 28-31, 1993, p. 354.

[3] Klatt D H and Klatt L, "Analysis, synthesis and perception of voice quality variations among female and male talkers", JASA,87(2), 820-857,1990

[4] P K Dhanarajani, et.al.,"A PC based Graphic Tool for Analysis, Segmentation and Labelling of Speech Signals", Proc. 3rd ICAPRDT-93, ISI, Calcutta, India, Dec 28-31, 1993, p. 326.

PHONETICALLY SUFFICIENT ALLOPHONIC DATABASE FOR CONCATENATION SYNTHESIS OF RUSSIAN SPEECH

Nina V. Zinovieva
Moscow State University, Moscow, Russia

ABSTRACT

The paper describes a phonetically sufficient database of Russian speech samples corresponding to phoneme-size units (or so called allophones) derived from different phonetic contexts. The database was designed for the purpose of automatic speech synthesis and was implemented in a concatenation text-to-speech Russian synthesis system

INTRODUCTION

The goal of the research is to create an optimal database (or inventory) of speech units for a Russian text-to-speech synthesizer, based on waveform concatenation. During the work we had to solve several problems. One of them is the appropriate choice of basic concatenation units. They may be diphones, triphones, syllables, demisyllables, and even words. However, all of them usually modify their acoustic quality in speech string (due to the coarticulation influence of adjacent elements). It seemed reasonable to choose the allophone (which we understand as *acoustically and perceptually distinguished context dependent realizations of phoneme)* as the basic unit, so that we could model the coarticulation phenomena when these units are being spliced together to form the synthesized speech.

The choice of linguistically motivated units (allophones) enables (a) to cluster phonemes in classes according to their flexibility in continuous speech, and (b) to cluster different context environments in groups according to their influential power and type of influence. Using these two features of allophonic approach we created an optimal set of speech units (a total of 667, using about 1 Mb memory to store them) which covers all coarticulation effects and thus provides the natural sounding of the synthesized speech.

The basic method for elaborating an exhaustive inventory of Russian allophones consisted of (a) an expert estimation of extracted sound wave segments corresponding to allophones and (b) an expert estimation of allophones extracted from one context and implanted into another, with similar or different coarticulation effect.

As a rule, the units for the concatenation are phoneme-size segments of the speech wave, although there are some exceptions. For example, to create stops, affricates and trills [r] and [r'], usually more than one acoustic segment are used; while to synthesize some two-phoneme sequences, such as post-stressed endings, a single acoustic segment can be used.

The main result of the work is a full description of the Russian allophonic system and elaboration of the optimal allophonic inventory in the form of acoustic speech signal (wave-form).

PHONETIC INVENTORY

The necessary condition for the work of an allophonic processor is the availability of phonemic transcription of the text to be synthesized. It is provided by an automatic transcriber specially designed for the Russian speech synthesis system. We used the following inventory of the Russian phonemes.

1. Stressed vowels: [á], [ó], [ú], [é], [í], [ý];
2. Unstressed vowels of the first degree of reduction: [a], [u], [i], [y] [o], [e]. The last two unstressed vowels are not regularly used in standard Russian, but sometimes they are pronounced in borrowed words.
3. Unstressed vowels of the second degree of reduction: [ax], [ix], [ux]
4. Non-palatalized consonants: [p], [t], [k], [b], [d], [g], [s], [sh], [z], [zh], [f], [v], [x], [c], [dz], [γ], [m], [n], [r].
5. Palatalized consonants: [p'], [t'], [k'], [b'], [d'], [g'], [s'], [z'], [sh'], [zh'], [f'], [v'], [x'], [ch'], [dzh'], [m'], [n'], [r'], [j']

One can see that the phonemic inventory used in our work slightly differs from that prevalent in Russian phonetic descriptions. This is because, for the purpose of synthesis, we had to choose such units that not only represent the phonemic relationships but also have acoustic and perceptual identity. It means that we have different units in our phonemic transcription even for those pairs that are in no meaningful contrast, but nevertheless, have different acoustic patterns which cannot be derived from one another: [x-γ], [c-dz], [ch'-dzh'], unstressed vs. stressed vowels, etc.

BASIC CONCATENATION UNITS

In arranging our database we proceed from the following three assumptions:

1. the amount of context-dependent variants is significantly larger for vowels than for consonants;
2. different consonants are affected by context influence to different degrees;
3. because of the prevalent CV-type of the Russian syllable, the left context is more important for vowels while the right context is more important for consonants.

According to these assumptions and a vast amount of preliminary expert estimations of the phoneme-size wave segments taken from different contextual environments, we divided the set of phonemes into the following classes.

Classes of vowels

1. Stressed vowels: [á], [ó], [ú], [é], [i], [ý];
2. Unstressed vowels: [a], [u], [i], [y], [o], [e], [ax], [ix], [ux];

Classes of consonants

1. Non-palatalized stops: [p], [t], [k], [b], [d], [g];
2. Palatalized stops: [p'], [t'], [k'], [b'], [d'], [g'];
3. Non-palatalized non-velar fricatives and affricates: [s], [sh], [z], [zh], [f], [c], [dz];
4. Palatalized non-velar fricatives and affricates: [s'], [z'], [sh'], [zh'], [f'], [v'], [ch'], [dzh'];
5. Nasals: [m], [n], [m'], [n'];
6. Liquids and velar fricatives: [l], [l'], [v], [v'], [x], [γ], [x'];
7. Trills: [r], [r']
8. Glide [j']

For each class the following relevant contextual environments were determined which affect the phonemes of the class, creating their allophonic modifications.

Relevant contexts for vowels

A. Left contexts for vowels

1. Beginning of syntagm-initial word.
2. Dental and alveolar non-nasal non-palatalized consonants, and central vowels: [d], [t], [s], [z], [c], [dz], [sh], [zh], and [á], [é], [a], [e], [ax].
3. Labial non-nasal non-palatalized consonants, and labialized vowels: [b], [p], [f], [v], [m], and [ú], [ó], [u], [o], [ux].
4. Velar non-palatalized consonants: [k], [g], [x], [γ].
5. Dental nasal non-palatalized consonant: [n].
5. Labial nasal non-palatalized consonant: [m].
7. Non-palatalized trill [r].

8. Non-nasal palatalized consonants, and front vowels: all consonants marked with the palatalization symbol (except [n'] and [m']), and vowels [i], [ý], [i], [y], [ix].

9. Dental nasal palatalized consonant [n'].

10. Labial nasal palatalized consonant [m'].

B. Right contexts for vowels

According to the third assumption, only five types of right contexts were considered for vowels:

1. End of syntagm-final word.

2. Non-labial non-palatalized consonants, and central vowels: [d], [t], [s], [z], [c], [n], [dz], [sh], [zh], [k], [g], [x], [γ] (the last four, only when not followed by labialized vowels [ú], [ó], [u], [o], [ux]), and [á], [é], [a], [e], [ax], [ý], [y].

3. Labial non-palatalized consonants, labialized vowels, and velar consonants followed by labialized vowels: [b], [p], [f], [v], [m], [ú], [ó], [u], [o], [ux], and [k], [g], [x], [γ] followed by [ú], [ó], [u], [o], [ux].

4. Non-palatalized thrill [r].

5. All palatalized consonants, and front vowels [í], [i], [ix].

Relevant contexts for consonants

For different groups of consonants different sets of relevant contexts were determined, to minimize the allophone inventory and, respectively, the database. This was achieved due to the fact that different consonants are differently sensitive to the environment, so they have different numbers of allophones. This difference will be shown in Table 1.

In compliance with the above, let us divide the relevant contexts for consonants into the following groups

A. Groups of left contexts for consonants

1. Beginning of syntagm-initial word.

2. Labialized non-reduced vowels [ú], [ó], [u].

3. Labial consonants, and labialized non-reduced vowels [b], [p], [f], [v], [m], and [ú], [ó], [u].

4. Front vowels: [í], [ý], [i], [y], [ix].

5. Front vowels [í], [ý], [i], [y], [ix] and all palatalized consonants.

6. Other left contexts

B. Groups of right contexts for consonants

1. End of syntagm-final word.

2. Labialized non-reduced vowels [ú], [ó], [u].

3. Labial consonants, and labialized non-reduced vowels: [b], [p], [f], [v], [m], and [ú], [ó], [u].

4. Any non-final context for palatalized consonants.

5. All vowels.

6. Other right contexts

Table 1 shows what groups of left and right contexts are relevant for each class of consonants.

The table does not show that there are special intervocalic allophones for consonants of the seventh group (trills [r] and [r']). Each of these intervocalic allophones consists of a single element. When the trills occur not intervocalically, the corresponding context-dependent allophone is added to the intervocalic element from the side of the consonant neighbor. If both neighbors are consonants, trills are synthesized from two context-dependent segments and intervocalic allophone between them.

DATABASE PREPARATION

To prepare the allophonic database, a special vocabulary was compiled of words containing all the necessary allophones in positions convenient for cutting off. When selecting the environments, only one typical representative of each group of contexts

Table 1. Left and right contexts relevant to different groups of consonants

Classes of consonants	Groups of left contexts						Groups of right contexts					
	1	2	3	4	5	6	1	2	3	4	5	6
1						+	+	+				+
2						+	+	+		+		+
3	+					+	+	+				+
4	+					+	+			+		+
5	+					+	+					+
6	+	+		+		+	+	+		+		+
7	+		+		+	+	+		+	+		+
8	+	+			+	+	+				+	+

was taken. The selected words were pronounced and recorded, one by one, by a professional TV-announcer. Then, the recordings were digitized and, with the help of a special sound processor, the allophones were cut off and stored in the database as separate files. Each file was given a name containing information about the group to which the phoneme belongs, its individual identity, and both right and left contexts from which it was taken and to which it should be implanted during the synthesis procedure.

SYNTHESIS PROCEDURE

During the synthesis, the input text is automatically transcribed, then the chain of transcription symbols is converted into a sequence of allophones corresponding to the context in which they occur The allophone are labeled by the above mentioned file names, and by these names they are extracted from the database and spliced together. Specia rules for pitch and duration patterns are implemented to make the synthesized speech sound as natural as possible.

ACKNOWLEDGMENTS

The author thanks I.Frolova and L.Zacharov for their assistance in preparing database, and Dr. Vladimi Segal for his helpful suggestions.

MULTILINEAR MODEL OF FRENCH PROSODY IMPLEMENTED ON A TEXT-TO-SPEECH SYSTEM

S. Barber, D. J. Hirst**, J. House***, P. Nicolas**, P. Roméas** and B. Waernulf**
(alphabetical order)

**Telia Promotor Infovox AB, Sweden.*
***Laboratoire CNRS URA 261 "Parole & Langage", Université de Provence, France.*
****Department of Phonetics and Linguistics, University College London, England.*

ABSTRACT

This model provides a TTS system with a multilinear hierarchical approach to prosody. Text is processed through two levels of constituents parsing : Intonation Units and Accent Groups. The acoustic realisation level is reached after rhythmic and tone sequence adjustments.

INTRODUCTION

This model for the French language was implemented on the Infovox TTS system. The RULSYS [1] rules format allows a formalism which is quite close to that of generative grammar. The model provides a stepwise derivation from underlying abstract levels of prosody to acoustic realisation level, which makes it quite different to what previous versions of the system did [2]. Most of the representations presented here rely on recent developments of multilinear phonology, which are the general framework of theoretical and experimental works on prosody at the Aix Institute of Phonetics [3,4].

PHONOLOGICAL REPRESENTATION

It is assumed that prosody has an autonomous level of representation, which means that it can to a certain extent be predicted independently from the syntactic and lexical material of the text. General prosodic pattern to be generated for any utterance is known prior to the syntactic analysis of the text. The identification of prosodic constituents is carried out through an algorithm that looks for syntactic markers and labelled parts of speech in the text. This provisional parsing may later be called into question according to rhythmic constraints or to tonal phonotactic constraints.

LEVELS AND CONSTITUENTS

It is assumed that all French utterances consist of a sequence of Intonation Units (IU), the deepest constituent level. A second level is represented by Accent Groups (AG). The extent of AGs is subordinated to the extent of IU : no AG is allowed to spread on both sides of a IU boundary. No AG can exist out of an IU constituent, whereas an IU may (in some extreme cases) not be parsed into AGs.

IUs can be terminal or non-terminal. All sentences have a single terminal IU. The number of non-terminal IUs in a sentence can be 0 and is theoretically unlimited. The assignment of IUs and AGs relies on the assumption that two types of syllables may potentially be assigned stress in French : word initial and word final ones. Assuming that the syllable is the phonological unit that carries stress, the IU must be considered a prosodic constituent made of a sequence of unstressed syllables followed by one stressed syllable (i.e. the IU head). Thus French is described as having right headed major prosodic constituents. IU stress is always on word final syllables, regardless to word class (+ or - OPEN). The system's feature <PSTRESS>, which can be attached to vowels, is used to identify this stressed syllable.

There is no lexical prosody in French: no distinctive stress feature belongs to the definition of lexical units. Yet the lexical unit is the domain of secondary stress in French [5,6], the primary stress being represented by IU heads. Our assumption of secondary stress refers to both lexical initial pitch accent and lexical final lengthening (except IU boundaries). Such a grouping defines what we call Accent Group (AG). Lexical units are identified by the feature <+OPEN> in the system lexicon. AG stress is exclusively assigned to *initial* and *final* syllable of the lexical unit. So at this deep level of representation, any lexical unit boundary syllable is potentially stressable, but no AG head is specified. The system recognises these syllables as their vowels carry the feature <ASTRESS>. At a later stage of processing, deletion rules, lengthening rules, and context sensitive rules provide appropriate tone and/or duration interpretation of <ASTRESS>.

IU-PARSING

IU-parsing consists of finding the end of every IU in the sentence. It is carried out through two successive levels of boundary assignment. The first level of IU boundary assignment deals with no longer deletable boundaries. The second level deals with IU boundaries that may be deleted according to rhythmic constraints. If not deleted, these second level boundaries are interpreted as tones. The way rhythmic constraints apply is explained below.

At the first level, boundaries are exclusively determined by punctuation marks, lexicon labels, or by some markers given as output of the system syntax module, without any rhythmic determination. For example, a comma always generates the same set of four features of the system, which itself is always interpreted as a given tone, a given lengthening, and a 25 frame pause. It is important to point out that this is an exception to the general idea of this model, since all other levels of prosodic categories assignment do not rely on term to term relations between some module output and one prosodic category. All other prosodic categories are subject to later contextual changes and deletions.

Boundaries of terminal IUs are generated at the first level. Punctuation marks ".", "?" and "!" indicate them. Depending on terminal punctuation mark and on the presence of <WH> question feature in the sentence, these boundaries will be interpreted by four possible tones at the realisation rules level. The last vowel in these IU are assigned features <PSTRESS, TERM> which lengthening rules will recognise as to be interpreted as the maximal lengthening (i.e. L5).

As for questions, a distinction is made between : wh-, yes/no, and "A ou B?" questions (consisting of two sections separated by the conjunction "ou"). This distinction relies on text and part of speech labels carried by the words.

Other punctuation marks "," ";" ":") as well as "(" and ")" also generate a first level boundary at their left. Some of them convey features that are responsible for further tone reduction. Lengthening in these cases is L4. Tones are defined in the system by sets of binary features. Lengthening is defined as a percentage of the default durations given in the definition module of the system.

A few connective words are IU initiators in French, and were labelled as such in the lexicon. They always generate a first level boundary at their left. They are a small list of previously <-OPEN> words, mostly conjunctions, now turned to <+OPEN> in the present system. Some of them get <+OPEN> in some restrictive morpho-syntactic contexts. Their label creates a specific feature at the appropriate processing stage.

The end of a relative clause <ENDREL> also generates a first level IU boundary. <ENDREL> can coincide with any punctuation mark, or with the beginning of the main clause verb. The main problem here was to identify this verb.

As to the second level of IU boundaries, syntactic markers initiate parsing but rhythmic balance rules may delete boundaries even though they were assigned at major syntactic constituents ends. Nevertheless, rhythmic rules do not absolutely ignore the boundary depth.

At a first stage of assignment, many IU boundaries are generated, each being assigned a rank which depends on syntax markers, part of speech labels and syllable count. Undeletable IU boundaries (i.e. first level) are assigned rank 0. Then processing is as follows :

1) Assign a rank 1 IU boundary in front of a relative clause marker.

2) Assign a rank 2 IU boundary in front of a prepositional phrase marker.

3) Assign a rank 3 IU boundary after a verb (identified by its PS-label).

4) Assign a rank 4 IU boundary in front of a verb phrase marker.

5) Assign a rank 5 IU boundary in front of any remaining phrase marker.

6) Assign a rank 6 IU boundary after the final syllable of a lexical unit (<+OPEN> word) if there are at least 4 syllables between it and the preceding boundary.

This of course gives far too many boundaries and a selective deletion process is needed. Rank by rank, starting from rank 6, rules delete as many boundaries as possible, namely as long as the number of syllables of the new formed IU is not over 10. The probability for a rank 1 boundary to be deleted is thus lower than the probability for a rank 6 boundary. A short sub-algorithm can be used in order to prevent leaving very short IUs either at the end or at the beginning of a sequence of words bounded by rank 0 boundaries after the deletion process.(See line (2) below).

In the following representation of a sentence, "x" stands for a syllable, a digit between two "x" represents a boundary with its rank, and "D" stands for "deleted". Each line represents a step in the derivation.

(1) xx0xxx1xx0xxx3xx2xxxx5xx2xx1xx0xxx0
(2) xx0xxx1xx0xxx3xx2xxxx5xx2xxDxx0xxx0
(3) xx0xxx1xx0xxx3xx2xxxxDxx2xxDxx0xxx0
(4) xx0xxx1xx0xxxDxx2xxxxDxx2xxDxx0xxx0
(5) xx0xxxDxx0xxxDxx2xxxxDxx2xxDxx0xxx0
(Result) xx0xxxxxx0xxxxx2xxxxxxx2xxx0xxx0

After deletions, remaining boundaries receive a tonal interpretation. At this level, tones are opposed to each other using a set of binary features (terminal or not, rising or falling, expanded range or not, reduced range or not) which are coded in the system. Reduction of range can occur depending on a syllable count threshold. Later on, in AG processing rules, another series of tones, of a different type, is needed. All tones are interpreted as acoustic Fo patterns at the end of all prosodic rules.

Relative clause, interrogative or exclamative contexts impose boundary transformations. As a summary :

- Any /+FoRISE/ boundary tone that is located inside a relative clause is turned to /-FoRISE/. Connected relative clauses are also specifically treated.

- All 3 types of questions require specific tone adjustments or assignments in non-terminal locations, since it is assumed that specific tonal marks of question occur in 3 locations in the sentence, not just at the end. These were called the "secondary" (sentence internal) and the "tertiary" (sentence initial) question marks. First a type-specific transformation of one of the non-terminal IU boundaries is made. Whatever the type of the question is, all non-terminal boundaries that follow the secondary question mark change /+FoRISE/ feature to /-FoRISE/. Eventually, another particularity of all questions is that all boundaries that are located between the secondary question mark and the preceding rank 0 boundary (if any) also reverses this feature polarity. Exclamative sentences are processed as questions in a first step. Then at the end of all parsing algorithms, some transformations are made. The tertiary question mark is a matter of AG stress assignment and is treated thereafter.

AG-PARSING

After IU-parsing is carried out, no IU boundary tone can be deleted any longer. Although AG stresses can only be assigned to initial and final syllables of <+OPEN> words, an AG stress should never be assigned to a syllable that has already been assigned an IU boundary. If the sentence is a yes/no question, an "A ou B?" question or an exclamative sentence with no <WH> word, then an AG stress is assigned to the first syllable of the first word in the IU ending with the secondary question mark. This is the only exception where an AG stress may be assigned on a <-OPEN> word. This initial AG stress will be interpreted by a specific higher tone later on, representing a tertiary question mark. As said earlier, all vowels that belong to a syllable that is assigned an AG stress take the feature <ASTRESS>. The first AG stress in any IU (AG1) is undeletable, except by a non-terminal /+FoRISE/ IU boundary tone located on its right adjacent syllable. If the sentence consists of just one terminal IU, this undeletable AG1 shifts to the end of the next word-final syllable located on a polysyllabic item. The undeletable peculiarity of this AG stress is coded by a feature on the vowel.

IU boundary tones never delete an adjacent AG stress that belongs to the next IU to the right (No cross-IU deletions). Non-terminal /+FoRISE/ boundary tones delete any AG stress on an adjacent syllable to its left. So do /-FoRISE/ boundary tones, terminal or not, except with AG1.

Then comes the AG deletion stage : proceeding from left to right, each AG stress deletes the next deletable AG stress

if it is located on the immediately following syllable. Example :
"*le CHAPEAU du MAIRE de MARSEILLE(...)*" becomes :
"*le CHApeau du MAIRE de MarSEILLE(...)*", in which "peau" is deaccented by AG stress on "cha", and "Mar" is deaccented by non-terminal rising IU boundary tone on "seille".

A couple of somewhat ad hoc but very useful tones were created, namely <ds> (for DownStep) and <rs> (for ReSet). <ds> is assigned under some conditions in declarative sentences on the antepenultimate syllable of the sentence, in order to avoid the effect of undesired smooth decaying of Fo towards the final low tone. It can be considered a variant of the declarative terminal boundary tone. <ds> cannot be assigned either on a syllable that already carries an undeletable AG stress, nor on any of its adjacent syllables. Tone <ds> is assigned in all other declarative contexts, regardless of the word class and of the position of the syllable in the word. It deletes any deletable AG stress on either the same syllable (i.e. antepenultimate) or one of its neighbours. Thus <ds> is stronger than AG stresses in phonotactic adjustments. <ds> provides a significant improvement of the pitch pattern's perceptual quality. <rs> also avoids unconvenient Fo transitions : assigning <rs> is a way to maintain a flat low Fo pattern throughout long sequences of <-OPEN> words located between the IU boundary tone and the following AG stress. Yet some IU tones do not allow it.

The AG level adds one more binary distinctive opposition among tones : AG tones, as opposed to IU tones, have specific Fo pattern and location. Thus, including <ds> and <rs>, the complete set of available tones in the system provides 15 possibilities. Every undeleted AG stress is interpreted as a specific tone, where all variants of reduced, expanded range, rising or falling realisations may occur. Tone phonotactic rules determine the polarity of tone features responsible for this specification. At this stage, all remaining word initial AG stresses have been assigned a tone, but word final AG stresses get a tone (namely a downstep called <ab>) *only if* they belong to a question or to an exclamative sentence. Each tone assignment makes the vowel lose its <ASTRESS> feature. Thus after AG-parsing, only word-final AG stresses that have not been interpreted as a tone are still identifiable as AG stress, since the vowel still carries <ASTRESS>.

Retained <ASTRESS> feature is exploited in lengthening assignment. Vowels that belong to syllables that carry an IU boundary tone have feature <PSTRESS>, which is exploited in lengthening assignment too. In addition, the feature <STRESS> is assigned to all <+OPEN> word initial and final syllables (and also to <-OPEN> word initial and final syllables if they carry tone <ds>). Thus : vowels carrying an IU tone become <PSTRESS, STRESS>; <ASTRESS> becomes <ASTRESS, STRESS>; vowels carrying an AG tone (or a <ds> at word boundaries) become <-ASTRESS, STRESS>; other syllables are <-STRESS>. Lengthening rules are :
L1 if <-STRESS> (shortening rule)
L2 if <+STRESS,-ASTRESS>
L3 if <+STRESS,+ASTRESS>
L4 if <STRESS,PSTRESS>
L5 if sentence final <STRESS, PSTRESS>.

CONCLUSION:

This multilinear hierarchical model introduced significant improvements in the processing of linguistic knowledge. The autonomy of prosodic representations, of interpretative rules between levels, and of adjustments, avoids unsatisfactory prosodic patterns obtained directly from morpho-syntactic analysis.

REFERENCES :

[1] Carlson, R.; Granström, B. (1990), "An environment for multilingual text-to-speech development", Proc. of the ETRW on speech synthesis, Autrans, France, Vol.2, 73-82.
[2] Barber, S.; Granström, B.; Touati, P. (1988), "French prosody in a rule-based text-to-speech system", Proc. of 7th FASE Symp., 3, 967-974.
[3] Hirst, D. J. (1994), "The symbolic coding of fundamental frequency curves: from acoustics to phonology", International Symposium on Prosody, Yokohama, 1-5.
[4] Hirst, D. J. ; Di Cristo, A. (1984), "French intonation: a parametric approach", *Die Neueren Sprachen* , 83 (5), 554-569.
[5] Di Cristo, A.; Hirst, D. J. (forthcoming), "L'accentuation non-emphatique en français : stratégies et paramètres", *Hommage à I. Fonagy.*
[6] Pasdeloup, V. (1993), "A prosodic model for French TTS (...)", *Talking Machines : theories, models and designs*, Bailly et al. (Eds), Elsevier.

TONAL CORRELATES OF DISCOURSE STRUCTURE

Antonis Botinis
Phonetics Laboratory, Department of Linguistics
University of Athens, Greece

ABSTRACT

The results of the present investigation indicate that discourse boundaries and tonal boundaries coincide and each topic constituent is aligned with tonal boundaries as a rule. Topic onset boundaries have a rising tonal pattern onto the first stressed syllable and topic offset boundaries have a falling tonal pattern off the last stressed syllable. Major topic constituents are reflected on tonal structure but the hierarchical structure of discourse does not correspond to tonal structure.

INTRODUCTION

This paper is an investigation of discourse tonal correlates in Greek. Tonal segmentation is an acoustic correlate of discourse structure as it defines tonal phrase boundaries which may be aligned with variable discourse structure units. Furthermore, tonal (i.e. intonation) phrases may be combined into larger tonal structures with distinctive binding patterns and coherence relations. Both segmentation and binding are related to information structure as the former defines the speech units as information units and the latter combines information units in larger information structures. Tonal segmentation patterns and binding distinctions in Greek discourse are outlined in recent reports ([1], [2]). In the present contribution tonal segmentation correlates and tonal reflections of discourse structure are investigated in a short news corpus. Our analysis is concentrated to tonal segmentation strategies with reference to different discourse domains of tonal segmentation applications.

SPEECH ANALYSIS

Speech material. The speech material of the present investigation consists of four short news reports from a Greek radio station. Each report was about one minute long and referred to national and international news. The recordings were carried out in four consecutive days. The language style is quite compressed, and elliptical structures may occur, to the extend they do not produce ambiguous interpretations. Topic constituents are well-structured and topic changes are very abrupt and well defined.

Speaker characteristics. The speaker was a male professional news speaker, rather young, and spoke standard (Athenian) Greek. He spoke at a rather fast tempo and sounded as if he had memorised the text and uttered the news in a natural and fluent way.

Experimental analysis. The speech material was recorded at the Athens University Phonetics Laboratory and the acoustic analysis was carried out at the Lund University Phonetics Laboratory with the ESPS/Waves+ software package.

Speech analysis. The speech material was segmented into topic constituents and classified in three kinds of topics: simple topics, compound topics, and minor topics. Simple Topics (TS) refer to a single topic of discussion and may be composed of a simple or a compound sentence. Compound Topics (TC) refer to a topic of discussion which has several aspects and may be composed by several minor topics which correspond to different aspects of the topic. Minor Topics (TMr) may be composed by a simple or a compound sentence, or by an elliptical sentence with close attachment to another topic constituent.

F0 measurements covered all topic constituents and were taken at five points: (1) at Topic Onset (TON), (2) at topic First Stressed Syllable (FSS), (3) at Topic Maximum F0 (TM), (4) at topic Last Stressed Syllable (LSS), and (5) at Topic Offset (TOF). The TM F0 is an independent measurement, regardless of the TON and the FSS F0 values. Furthermore, topic pause durations (not shown at figures) were measured. Two TC TOFs showed an alternative tonal realisation (tonal rise) from the rest of the material and were not included in the presentation.

TONAL SEGMENTATION

Intratopic (syntagmatic) and intertopic (paradigmatic) tonal relations are referred to as Tonal Conditions and Topic Conditions respectively.

Figure 1 shows tonal correlates of the first news report. This material is composed of 12 topics from which 7 are TSs and 5 are TCs. The TCs are composed of 13 TMrs (from 2 to 4 TMrs for each TC). Thus, there were 20 topic condition measurements.

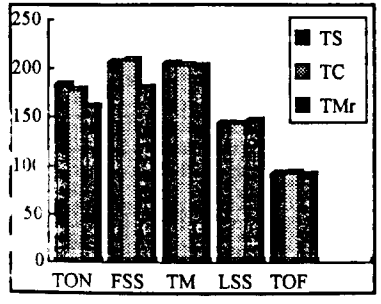


Figure 1. Tonal distribution of Simple Topics (TS), Compound Topics (TC) and Minor Topics (TMr) (see text).

The tonal condition indicates the following: First, there is a tonal anathesis from TON to the FSS. Second, there is hardly any difference between FSS and TM except for the TMr tonal condition. Third there is a LSS tonal catathesis and, fourth, a tonal fall from LSS to TOF which reaches the speaker's tonal floor.

The topic condition indicates the following: First, there is a slightly descending order from TS to TC to TMr for TON. Second, there is an equal tonal distribution to TS and TC for the FSS but a lower one for the TMr topic condition. Third, the TM, the LSS, and the TOF for all topic conditions are the same.

Figure 2 shows tonal correlates of the second news report. This material is composed of 9 topics from which 6 are TSs and 3 are TCs. The TCs are composed of 12 TMrs (from 2 to 6 TMrs for each TC). Thus, there were 18 topic condition measurements.

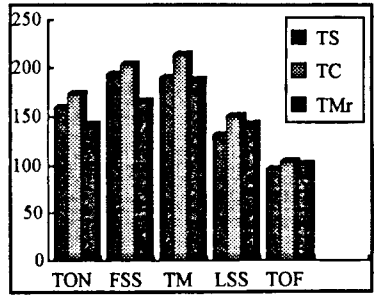


Figure 2. Tonal distribution of Simple Topics (TS), Compound Topics (TC) and Minor Topics (TMr) (see text).

The tonal condition indicates the following: First, the FSS and TM conditions do not have major F0 differences except for the TMr tonal condition. Second, both FSS and TM have higher F0 than the TON condition for all topic conditions. Third, the LSS has lower F0 than the TON and FSS conditions for all topic conditions and, forth, the TOF condition has the lowest F0 than all tonal conditions for all topic conditions.

The topic condition shows the following regularities: First, the TC condition has the highest F0 for all tonal conditions, most for the TON, FSS and TM conditions between the TC and TMr and least for the LSS and TOF tonal conditions. Second, the TS condition has higher F0 than the TMr condition for the TON and FSS tonal conditions, equal F0 for the TM condition but lower F0 for the LSS and TOF tonal conditions.

Figure 3 shows tonal correlates of the third news report. This material is composed of 11 topics from which 5 are TSs and 6 are TCs. The TCs are composed of 15 TMrs (from 2 to 4 TMrs for each TC). Thus, there were 20 topic constituency measurements.

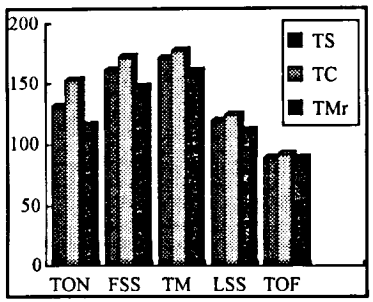


Figure 3. Tonal distribution of Simple Topics (TS), Compound Topics (TC) and Minor Topics (TMr) (see text).

The tonal condition indicates the following: First, the FSS and TM conditions do not have major F0 differences except for the TMr topic condition. Second, both FSS and TM have higher F0 than the TON tonal condition for all topic conditions. Third, the LSS has lower F0 than the TON and FSS conditions for all topic conditions and, forth, the TOF condition has the lowest F0 than all tonal conditions for all topic conditions.

The topic condition shows the following regularities: First, the TC condition has the highest F0 than the TS and TMr conditions, most for the TON, FSS and TM tonal conditions between the TC and TMr and least for the LSS and TOF conditions. Second, the TS topic condition has higher F0 than the TMr one for all but the TOF tonal conditions.

Figure 4 shows tonal correlates of the forth news report. This material is composed of 9 topics from which 2 are TSs and 7 are TCs. The TCs are composed of 21 TMrs (from 2 to 5 TMr for each TC). Thus, there were 23 topic condition measurements.

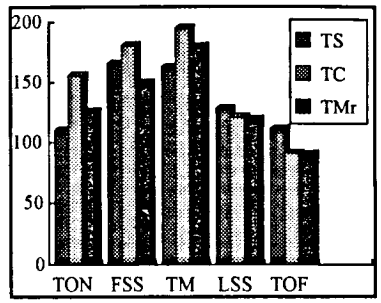


Figure 4. Tonal distribution of Simple Topics (TS), Compound Topics (TC) and Minor Topics (TMr) (see text).

The tonal condition indicates the following regularities: First, there is an anathetic structure from TON to FSS and from FSS to TM for all topic conditions except for the TS condition between FSS and TM. Second, there is a catathetic structure onto TOF for all topic conditions. Third, The LSS has lower F0 than the TOF for TC and TMr and, fourth, the TOF condition has the lowest F0 for all topic conditions.

The topic condition shows the following: First, the TC condition has the highest F0 for the TON, FSS and TM tonal conditions whereas the TS condition has the highest F0 for the LSS and the TOF tonal conditions. Second, TMr has higher F0 than TS for TON and TM but not FSS. Third, TC and TMr have the same F0 for LSS and TOF tonal conditions.

Figure 5 shows an average of the tonal conditions and topic conditions of all four news reports.

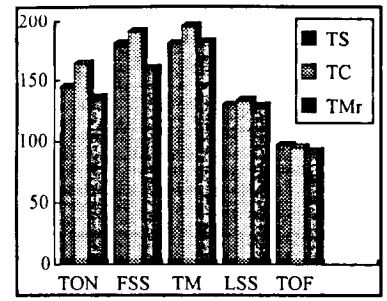


Figure 5. Tonal distribution of Simple Topics (TS), Compound Topics (TC) and Minor Topics (TMr) (see text).

In summary, the tonal condition shows an anathetic structure from the Topic Onset onto the First Stressed Syllable and no major difference between the First Stressed Syllable and the Topic Maximum. On the other hand, there is a catathetic structure onto the Last Stressed Syllable and a final fall to the Topic Offset. The topic condition shows a hierarchic structure for the Topic Onset and the First stressed Syllable with Topic Compound highest, Topic Simple next, and Topic Minor last. The Topic Simple and Topic Minor conditions are neutralised for Topic Maximum. The topic condition is neutralised for the Last Stressed Syllable and the Topic Offset conditions. The most constant tonal segmentation correlate in the present analysis has been the Topic Offet tonal fall which reaches the speaker's tonal floor as a rule (about 90 Hz). Pause distribution at topic boundaries vary considerably (from 2.5 to 6.0 cs) with no regular relation between Topic Condition and pause durations.

DISCUSSION AND CONCLUSIONS

The results of the present investigation indicate that topic segmentation may be correlated with tonal segmentation in terms of lower tonal distribution at topic last stressed syllable and topic offset boundary. Topic segmentation and topic structure may be reflected on tonal structure in terms of higher tonal distribution at topic onsett boundary and first stressed syllable. Tonal structure does not however have a one to one hierarchical correspondance to discourse structure. This implies that discourse units (e.g. Grosz and Sidner [3]) and textual features (cf. Hirshberg [4]) may have variable effects on tonal structures in accordance with the communicative requirements of the message in which every topic constituent should be one information unit. Other types of Greek material (e.g. [1], [2]) may show different types of tonal segmentation and binding patterns with complex information structures.

REFERENCES

[1] Botinis, A. (1992), "Accentual distribution in Greek discourse", *Travaux de l'Institut de Phonetique d'Aix*, Vol. 14, pp. 13-52.

[2] Botinis, A. (1994), "Intonation and prosodic coherence in Greek discourse", ESCA Workshop on Prosody, *Working Papers* 41, pp. 100-103, Dept. of Linguistics and Phonetics, Lund University.

[3] Grosz, B.A. and Sidner, C.L. (1986), "Attensions, intensions, and the structure of discourse", *Computational Linguistics* 12, pp. 175-204.

[4] Wang, M.Q. and Hirschberg, J. (1992), "Automatic classification of intonational phrase boundaries", *Computer Speech and Language* 6, pp. 175-196.

PROSODIC MARKERS AT SYNTACTIC BOUNDARIES IN SPANISH

Juan M. Garrido, Joaquim Llisterri, Rafael Marín, Carme de la Mota and Antonio Ríos

Departament de Filologia Espanyola, Facultat de Filosofia i Lletres, Edifici B, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain.
Fax: +34.3.581.16.86; E-mail: joaquim.llisterri@cc.uab.es

ABSTRACT

This paper presents a preliminary examination of the use of vowel duration and F_0 movements as markers of different types of syntactic boundaries in Spanish. The study is based on a corpus of sentences extracted from read newspaper articles. Results reveal that certain prosodic cues might convey information about internal syntactic boundaries and that other cues may be related to a syntactic cohesion.

INTRODUCTION

Studies devoted to the use of prosodic markers to signal syntactic boundaries have been carried out for different languages. For example, Klatt [1] reports on the use of vowel duration to signal syntactic boundaries in English, and the use of F_0 cues is discussed, among others, by Cooper and Sorensen [2]. As far as Spanish is concerned, Signorini *et al.* [3] consider the relationship between local F_0 movements and juncture, but the question of syntactic boundaries is not directly addressed.

In the present paper the relationship between prosody and syntactic boundaries in Spanish is explored in two ways. First of all, it is intended to establish the type of phonetic phenomena which may signal the presence of a syntactic boundary. Secondly, the prosodic characteristics of different types of boundaries are compared.

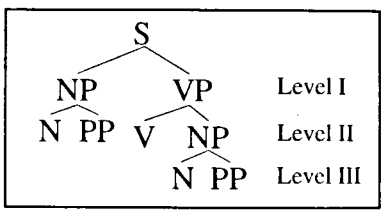
Vowel duration and F_0 movements have been taken into account here, but other cues such as pauses should be explored in future studies.

EXPERIMENTAL PROCEDURE

Corpus

A large corpus of paragraphs extracted from Spanish newspapers has been collected and labelled according to syntactic and prosodic criteria. One hundred nineteen sentences read by three non professional speakers have been chosen for the present experiment. Recordings were made in semi-anechoic conditions using a Tascam 112 cassette recorder and a Sennheiser MKH2 microphone.

Four types of syntactic boundaries are considered in order to study its effects on prosodic parameters: NP(Subject)-VP, VP-NP(Object), N-PP where N is the head of the NP (Subject) and N-PP where N is the head of the NP(Object). The hierarchical structure in three different levels can be seen in the syntactic tree:



The following variables have been taken into account in choosing the sentences from the corpus: (1) presence or absence of a pause at the NP-VP boundary; (2) length of the phrase preceding the boundary; two types of phrases have been included: long ones - more than 8 syllables - and short ones - between 2 and 8 syllables -; (3) stress placement, in order to have an equal number of stressed and unstressed syllables before the boundary; (4) syllabic structure and number of allophones in the syllables analyzed: CV syllables and CVC syllables have been included in equal numbers.

Measurements

The signal has been low-pass filtered and digitized at 10 kHz sampling rate using a MacAdios II™ card. Waveform displays, broad-band spectrograms and F_0 contours were plotted for each sentence using the Mac Speech Lab II™ software running on an Apple Macintosh IIvx™. F_0 contours were obtained with a pitch tracking algorithm using an auto-correlation technique.

The duration of the vowel preceding the boundary and also in certain word-internal positions has been measured. In order to characterize F_0 movements, maximum and minimum values in the F_0 contour observed before and after boundaries have been measured at vowel centers. Maxima were located in vowels showing an F_0 value higher than the value at the preceding and following syllables. Accordingly, minima were located in vowels showing an F0 value lower than the value at the preceding and following syllables.

RESULTS

Vowel duration

In order to compare the effect of different types of boundaries in vowel duration, only syllables containing vowels [e] and [o] have been taken into account, since it is known that these vowels have very similar intrinsic durations [4]. All vowels considered in the comparisons are non-prepausal and unstressed.

It can be seen from data in figure 1 that no significant differences have been found between the duration of vowels preceding the four types of syntactic boundaries.

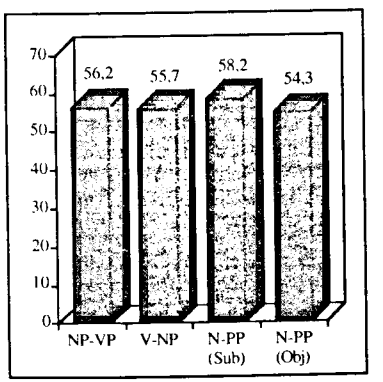


Figure 1: Vowel duration (in ms.) at different types of syntactic boundaries.

In order to assess the differences in duration between vowels located before a syntactic boundary and vowels which are not in this position, a comparison was made with a different group of vowels. They were selected from the penultimate syllable of nouns heading an NP or selected by a PP; both phrases were adjacent to the verb and subcategorized by it.

In this case, only non-prepausal vowels were analyzed and an equal number of stressed and unstressed vowels were compared; the comparison includes the vowels [e], [o] and [a]. Results are shown in figure 2:

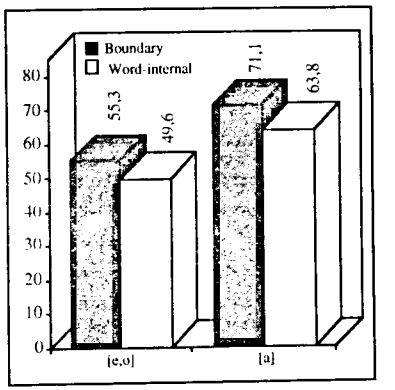


Figure 2: Vowel duration (in ms.) at syntactic boundaries and in word-internal position

Although no statistically significant differences appear, a tendency can be observed towards longer vowel durations when they occur in a syntactic boundary. It should be noted that vowels were non-prepausal in both cases, thus, the effects of pre-pausal lengthening have been avoided.

F_0 movements

Two parameters have been studied for each type of boundary: contour shape and reset.

F_0 movements have been characterized in terms of the local shape of the F_0 contour. Peaks and valleys adjacent to the boundary are used to define this contour.

An initial distinction will be made between boundaries signaled by a pause and boundaries which are realized without pause. Results for boundaries which are not coincident with a pause are summarized in table 1:

Table 1: Relationship between F_0 shape and type of syntactic boundary in boundaries without pause (B: type of boundary; S: stressed / unstressed syllable; % P: percentage of occurrence of a peak before the boundary; % PFS: percentage of cases in which, when there is a peak, it appears in the syllable just before the boundary)

B	S	% P	% PFS
NP-VP	stressed	73	75
	unstressed	95	72
V-NP	unstressed	68	100
N-PP(S)	unstressed	79	73
N-PP(O)	unstressed	92	64

Table 1 shows that in most cases boundaries are preceded by F_0 peaks; it also shows that when peaks appear, they are mostly located in the syllable that immediately precedes the boundary, irrespective of the stressed or non-stressed character of the syllable. It seems that no influence of the type of syntactic boundary can be found in the occurrence and location of peaks.

The results for boundaries occurring simultaneously with a pause are shown in table 2:

Table 2: Relationship between F_0 shape and type of syntactic boundary in boundaries with pause (see table 1 for legend)

B	S	% P
NP-VP	stressed	59
	unstressed	74

Resets have been examined in all types of boundaries. A reset has been considered to occur if the F_0 peak following the syntactic boundary is higher than the preceding peak, or if the following valley is higher than the preceding one. Three categories of reset can then be defined: (a) only peak-to-peak reset, (b) only valley-to-valley reset, and (c) simultaneous peak-to-peak and valley-to-valley reset.

Results are presented in tables 3 and 4 below.

Table 3: Percentage of resets (%R) and most frequent reset type (RT) at the NP-VP boundary (P: peak-to-peak reset; V: valley-to-valley reset)

	S	% R	RT
With pause	stressed	82	P+V 64%
	unstressed	74	P+V 50%
Without pause	stressed	58	V 57%
	unstressed	42	V 75%

Resets appear more frequently when the NP-VP boundary is coincident with a pause. It is also worth noting that in those cases, simultaneous peak and valley resets are found.

Table 4: Percentage of resets (%R) and most frequent reset type (RT) at syntactic boundaries without pause preceded by an unstressed syllable (P: peak-to-peak reset; V: valley-to-valley reset)

B	% R	RT
NP-VP	42	V 75%
V-NP	42	V 38%
N-PP (S)	26	V 80%
N-PP (O)	34	P 75%

It appears from these results that the frequency of occurrence of resets can be related to the degree of syntactic cohesion.

DISCUSSION

As far as temporal markers of syntactic boundaries are concerned, only vowels in non-prepausal position have been examined, since the effects of pre-pausal lengthening in Spanish are already well established [4]. A tendency towards a lengthening of vowels in syntactic boundaries seems to appear, but the results are not conclusive from a statistical point of view (see figure 2).

An examination of the shape of the F_0 contour shows that an F_0 peak is consistently present in the syllable preceding the boundary when a pause is not produced. The presence of this peak could be also explained as a function of the lexical accent. It is interesting to note that in the case of unstressed syllables, the peak could be the result of a shift from the lexically stressed syllable, as discussed in [5]. Nevertheless, the presence of a peak simultaneously to the stressed syllable preceding the boundary suggest that NP-VP boundaries may, at least in some cases, inhibit peak displacement, so that F_0 peaks could mark the boundary. It seems also that pauses tend to reduce the occurrence of F_0 peaks at NP-VP boundaries (see table 3).

Resets also tend to appear at syntactic boundaries. It is worth noting that they seem to be related to the occurrence of a pause in the case of NP-VP boundaries.

The discrimination of the type of syntactic boundary by means of prosodic cues has also been studied. It seems clear that vowel duration does not distinguish between the types of syntactic boundaries observed. The F_0 shape at boundaries without pause does not seem to contribute either to this differentiation.

F_0 resets seem to be the only prosodic cue that behaves in a different way according to the type of boundary, since NP-VP and V-NP boundaries show a higher percentage of resets than N-PP boundaries (see table 4). Resets could be then related to major boundaries.

Further research is needed to establish whether the degree of reset can cue differences in the syntactic hierarchy.

The relationship between concomitant cues can also be explored. On the one hand, the presence of a pause seems to trigger a reset (see table 3). On the other hand, it could be also possible that F_0 peaks were related to pauses, since peaks at boundaries seem to be more frequent when pauses are not present (see tables 1 and 2).

CONCLUSIONS

A first attempt to describe the behaviour of prosodic cues at syntactic boundaries in Spanish has been presented. The results of the study are not conclusive about the use of prosodic cues in marking different types of boundary. Moreover, vowel lengthening and F_0 movements adjacent to boundaries which do not coincide with a pause can not be clearly related to syntactic phenomena. However, some results tend to favor the hypothesis that resets could be indicators of the degree of syntactic cohesion.

Acknowledgment

R. Marín's work has been partially supported by a grant from the *Generalitat de Catalunya (Programa de Formació de Personal Investigador, Àrees Prioritàries)*

REFERENCES

[1] KLATT, D.H. (1975) "Vowel lengthening is syntactically determined in a connected discourse", *Journal of Phonetics* 3: 129-140.

[2] COOPER, W.E.- SORENSEN, J.M. (1971) "Fundamental frequency contours at syntactic boundaries", *Journal of the Acoustical Society of America* 62,3: 683-692.

[3] SIGNORINI, A., BORZONE, A. M. - MASSONE, M. Y. (1989) "Los movimientos de F0 como correlatos de Juntura y acento", *Revue de Phonétique Appliquée 91-92-93:* 376-388.

[4] MARÍN, R. (in press) "La duración vocálica en español", *Estudios de Lingüística* (Alicante).

[5] GARRIDO, J.M.- LLISTERRI, J.- de la MOTA, C.- RÍOS, A. " Prosodic differences in reading style: Isolated vs. Contextualized Sentences" in *Eurospeech'93. 3rd European Conference on Speech Communication and Technology.* Berlin, Germany, 21-23 September 1993. Vol 1. pp. 569-572.

DECLINATION IN DUTCH AND DANISH: GLOBAL VERSUS LOCAL PITCH MOVEMENTS IN THE PERCEPTUAL CHARACTERISATION OF SENTENCE TYPES

Charlotte Gooskens (1) and Vincent J. van Heuven
Department of Linguistics/Phonetics, Leiden University, The Netherlands
(1) Now at Department of Linguistics, Nijmegen University, The Netherlands

ABSTRACT

This research deals with the question whether Standard Dutch and Standard Copenhagen Danish have systematically different speech melodies for interrogative, non-final, and declarative utterances. Rate of declination and local pitch movements were investigated. The Danish model developed by Thorsen [1] could be confirmed in its use of different rates of declination for different sentence types. The Dutch model [2] proved too simple in so far as it uses a uniform declination for all utterances of the same length: in Dutch, too, there is an effect of sentence type on declination rate.

According to the models, the two languages differ in their use of local pitch movements as cues to sentence type. The Danish model recognizes only one kind of local pitch movement (an accent, not a boundary tone), whereas several local pitch movements are distinguished by the Dutch model. The results of our experiments perceptually confirm that in Dutch, but not in Danish, a local final pitch rise is used in combination with declination rate to signal sentence type.

INTRODUCTION

The Dutch intonation grammar [2] claimed that the Dutch sentence melody, whether in statements or in questions, is characterized by a general downtrend of fundamental frequency (F0) over the course of the utterance (declination). The difference between sentence types such as statement and question is coded in the final portion of the sentence melody only, where F0 remains high after the last accent-lending rise or where a non-accent-lending terminal rise is executed when the last accent is a fall. These configurations may occur at the end of a sentence as well as at the edge of a non-terminal clause, in which case they signal continuation.

According to Thorsen [1], Danish intonation in terminal statements is comparable to Dutch: in both languages the same type of declination applies. However, non-terminal statements in Danish are characterised by a more gradual declination. Moreover, declination is claimed to be absent from Danish (echo) questions. The Danish model of intonation does not distinguish between different kinds of pitch movements.

Our first aim is to establish whether Danish differs from Dutch by using various degrees of declination (full, half and no declination for 'statement', 'continuation' and 'question' respectively), whereas Dutch does not employ declination as a parameter, or whether the intonation research for Dutch has overlooked functional contrasts in the declination parameter.

Our second aim is to establish whether indeed final rise cannot be used in Danish, as opposed to Dutch, to signal a difference between statement and continuation/question.

EXPERIMENT I

Method

The aim of this experiment was to investigate the importance of declination and pitch movements for the perceptual discrimination of sentence type in Danish and Dutch. Declination and pitch movements of the Dutch and the Danish models of intonation were systematically combined, imposed onto a Danish and a Dutch utterance and presented to Dutch and Danish listeners.

The two test utterances used in this experiment were the Danish and the Dutch translations of the sentence "(There are) mány bússes out of Tíflis". Only the last four words were used in order to deprive the respondents of syntactic clues to sentence type.

Dutch accents were implemented as an early rise followed by a gradual fall, i.e. configuration 1D in [2]. Additionally the interrogative and non-final utterances had a rise 2 on the last word. Declination slopes were generated on the basis of the formula in [2].

Danish stress groups were generated on the basis of Thorsen's model [1]. Declination slopes were estimated from both the model and sample utterances provided (for details see [3]).

Since there is no difference between 'question' and 'continuation' in the Dutch intonation model, we end up with nine pitch contours, which were applied to the Dutch and the Danish sentences:

declin.	rise	stat.	cont.	quest.
DK	DK	#1	#2	#3
NL	NL	#4	#5	
DK	NL	#6	#7	#8
NL	DK		#9	

Experiment I consisted of three parts. In the first two parts 20 Dutch and 17 Danish respondents listened to the nine versions of the sentence in their own language presented randomly. Their task was to decide for each utterance whether it expressed question, continuation or final statement. In part II they judged the naturalness of each utterance on a scale from 1 (least natural) 10 (most natural). In part III pairs of versions were presented to the subjects, who selected the member of each pair which they thought sounded more natural.

Results

Part I. The respondents had to decide whether each of the utterances expressed question, non-final or final statement. The results are presented in the left half of Table 1.

Table 1. Results of part I and II for the Dutch respondents (NL) and the Danish respondents (DK). Horizontally the possible answers: question (?), continuation (,) and final statement (.). Vertically the nine different versions of the utterance. The results of part I are in %. The results of part II are mean judgments on a scale from 1 to 10. The 'correct' answers are underlined.

	responses in % part I						responses part II	
	NL listeners			DK listeners				
stimuli	?	,	.	?	,	.	NL	DK
#1	69	20	10	85	10	5	5.1	6.0
#2	21	55	24	26	68	6	5.2	4.9
#3	13	33	55	9	68	24	5.1	5.3
#4	5	70	25	6	38	56	7.2	4.8
#5	0	30	70	12	0	88	7.0	5.9
#6	73	28	0	76	18	6	7.0	5.1
#7	51	46	3	47	44	9	6.6	3.8
#8	3	40	58	6	18	76	6.2	4.3
#9	8	28	64	3	26	71	5.7	5.8

Clearly, the Thorsen model is confirmed: in Danish, the declination slope is very important for the recognition of sentence type. The question type has no declination (stim. #1, 85%) and the final statement has a rather steep declination slope (stim. #5, 88%). The continuation type should have a slope in between that of the question and the statement type (stim. #2, 68%).

Dutch listeners also need the declination slope to recognize sentence type. For the Dutch listeners to recognize a

question there should be no declination and a rise 2 should be present (stim. #6, 73%). The continuation type is again recognized best if there is a final rise, but the declination slope has to be steep (stim. #4, 70%). The Dutch statement requires a steep declination and F0 has to end at the baseline (stim. #5, 70%).

Part II. In this part the respondents judged the naturalness of the utterances of part I. The conclusions of part I are confirmed in part II both for the Dutch model and for the Danish model (Table 1, right half). The utterances which were the best recognized instances of the three sentence types (#4, #5 and #6 for Dutch and #1, #2 and #5 for Danish) also obtained the highest naturalness scores.

Part III. In this part the respondents had to decide which of two sentences sounded more natural. The information which one gets from pairwise comparisons is generally more precise than information from categorical estimation. For lack of space no results will be presented, and we will proceed immediately to the conclusions.

For *Danish* the results from the first two parts are confirmed and thus the model of Thorsen has been confirmed in general. The *Dutch* results of part I and II are confirmed for the continuation and the final statement types, but no confirmation was obtained that the question intonation has to be without declination. The steep declination of the Dutch model is found to be more natural than no declination at all. In order to be able to draw stronger conclusions as to the importance of declination slope for the recognition of sentence type *in Dutch*, experiment II was carried out.

EXPERIMENT II

Method

The set-up of this experiment is very similar to that of the first. Different utterances were made for Dutch only, this time with five different declination slopes of -2, 0, 2, 4, and 6 semitones over the course of the 235 ms utterance (cf. table 2, leftmost column). These slopes were combined (non-orthogonally) with final pitch rises of 0, 2, 4, and 6 st/150 ms. The rest of the pitch contour was generated as in experiment I. The resulting 18 utterances were presented to 15 Dutch students as in experiment I, except that the third part was subdivided into three smaller sections (3a, 3b and 3c). In 3a the subjects were instructed to consider the stimuli as questions, in 3b as continuations, and in 3c as declarative statements.

Results

For the declarative and the continuation types the conclusions that can be drawn on the basis of the results are very similar to those of the first experiment. Again, only the results of the first two parts are presented.

As for questions, the first two parts confirm the findings of the first two parts of experiment I: questions should have an F0 that does not decline and a final rise 2 (table 2). In section 3a the subjects, however, show a (slight) preference for the utterance with both declination and a final rise, just as is prescribed in the Dutch model. The versions without declination, however, were judged almost as adequate, as long as they have the final rise. This means that Dutch listeners might use declination as a cue in the recognition of sentence type. It is also confirmed that the final rise is crucial for the recognition of a Dutch question.

SUMMARY AND DISCUSSION

The results clearly constitute a perceptual confirmation of the Danish model of intonation: Danish indeed distinguishes different slopes of declination in order to express sentence type; one local final pitch movement cannot be used for this purpose. Interrogative utterances (echo questions) have no declination, non-final utterances have a moderate rate of declination, and declination has to be steep for the declarative type.

Table 2. Percent responses (part I) and mean naturalness rating (1 to 10 scale, part II) broken down for 18 stimulus types differing in global declination slope and excursion size of local end rise.

stimuli		responses in % part I			responses part II
decl. st/234 ms	exc. st. end-rise	?	,	.	
6	0	0	7	93	7.5
6	2	0	37	63	6.4
6	4	10	73	17	6.0
6	6	13	63	23	6.1
6	8	33	40	27	6.3
4	0	0	7	93	7.0
4	2	57	43	0	6.0
4	4	3	73	23	5.9
4	6	33	63	3	6.0
2	0	10	47	43	5.7
2	2	20	60	20	5.6
2	4	47	43	10	5.7
2	6	43	50	7	6.1
0	0	13	50	37	4.8
0	2	37	57	7	5.9
0	6	60	37	3	6.4
-2	6	43	50	7	6.1
-2	6	93	7	0	6.1

The Dutch model of intonation, on the other hand, needs revision. According to this model, no distinction has to be made in declination slopes of different sentence types. However, the most important conclusion of this investigation is that (at least in some situations) questions should not decline in Dutch, while the declination in non-final and declarative sentences is equally steep. In contrast to Danish, however, final pitch movements are needed as well in order to distinguish between the different sentence types. Interrogative and non-final utterances should end on a high (i.e. non-low) pitch whilst declarative sentences should end on a low pitch.

On the basis of the results of the present investigation it might be possible to draw the conclusion that the use of different rates of declination (i.e. global intonational features) as opposed to local pitch movements to distinguish sentence types is a typological parameter differentiating between e.g. Danish and Dutch. If this is indeed the case it is likely that rate of declination is used in similar ways in other languages. It would therefore be very interesting to investigate the use of declination versus local pitch movements to express sentence type in a wider variety of languages, both Indo-European and non Indo-European.

REFERENCES

[1] Thorsen, N. (1980). A study of the perception of sentence intonation - Evidence from Danish. *Journal of the Acoustical Society of America*, 67, 1014-1030.

[2] Hart, J. 't; Collier, R.; Cohen, A. (1990). *A perceptual study of intonation.* Cambridge University Press.

[3] Grønnum, N. (1990). Prosodic Parameters in a Variety of Regional Danish Standard Languages, with a View towards Swedish and German. *Phonetica*, 47, 182-214.

PARAMETRIC DESCRIPTION OF F_0-CONTOURS IN A PROSODIC DATABASE

B. Heuft; T. Portele; F. Höfer; J. Krämer; H. Meyer; M. Rauth; G. Sonntag
Institut für Kommunikationsforschung und Phonetik, University of Bonn, Germany
Poppelsdorfer Allee 47, 53115 Bonn

ABSTRACT

A maximum-based model for parametrization of F_0-contours was developed. A prosodic database is described, which is used for a rule-driven as well as a data-driven approach to synthetic prosody. For each syllable, it contains perceptive, acoustic, and linguistic information.

MOTIVATION

For an investigation of intonation it is necessary to describe a given contour with as few parameters as possible, i.e. only the relevant information should be kept. The parameters must be automatically computable. On the other hand, it should be no problem to generate any possible F_0-contour with these parameters from an adequate input. So there is a double need: First, a model for parametrization has to be developed, and second, a database has to be constructed that contains all information that is supposed to have an influence on the model parameters.

THE MODEL

When developing this model, we assumed the F_0-maxima to be the perceptually most important points of the F_0-contour. This hypothesis was motivated by the experiences made in Bonn using the parametrization with Fujisaki's model [1], which does not predict the position of the F_0-peaks exactly. Further problems exist for the modeling of the utterance-final F_0-decrease. German stress can be signalled by a falling F_0, so quite a lot of unintended stress-shift occured.

Our new method of parametrization is called maximum based description. Each F_0-contour is parametrised describing only its maxima: for each maximum, four parameters are given.

First, the maximum is located precisely in time, relative to the onset of the accented vowel assigned to it. This distance is called *delay*, it is negative when the peak preceeds the vowel onset, positive when it follows.

The second parameter, the height of the maximum (*amplitude*), is described as a percentage value between a top- and a baseline. To simplify the description, these lines are currently kept constant for a given speaker. At a later stage, these lines may be used to modify easily the F_0-range.

The third and the fourth parameter describe the steepness of the contours preceding (*left slope*) and following (*right slope*) the maximum. These slopes are interpreted as sinoidal slopes with different degrees of damping. Fig. 1 explains the four model parameters with a stylized F_0-contour.

Minima are not described explicitly, but they are the crossing points of the contours preceeding and following the maxima. This method of description does not make any presumption about the functions of F_0-maxima; for example, no boundary tones are labelled, maxima at the end of questions are not treated differently to maxima caused by focus.

TEXT CORPUS

First, the corpus consists of three short texts, each about 400 words long. Then, there are 50 wh-questions and yes/no-questions each, with their pertinent answers. Some of the questions are segmentally identical but focus different words. There is further a couple of instructions and categorical questions, as

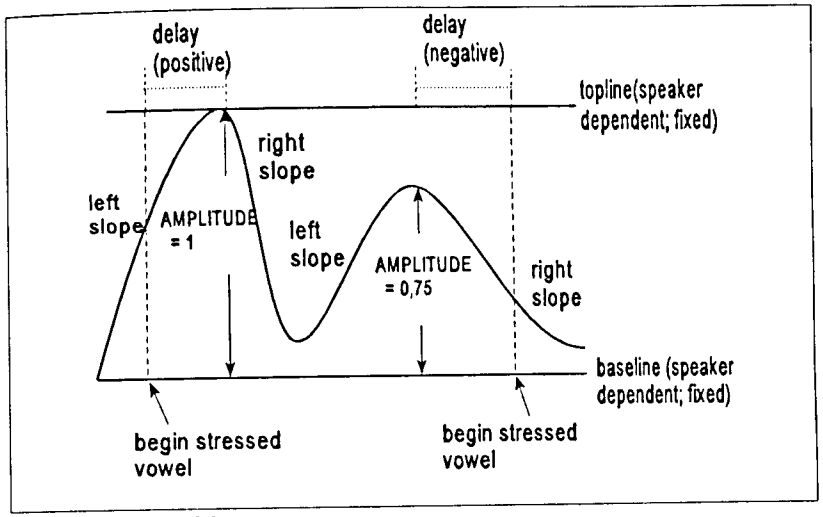


Figure1: The model parameters

well as hundred simple structured, isolated one-phrase utterances. The corpus was read by three speakers, two female and one male, among them the two speakers of the unit inventories for the speech synthesis system developed in Bonn [8]. The recording was done in the anechoic chamber of the IKP.

ACOUSTIC INFORMATION

Model parameters: An algorithm was developed, that automatically extracts the model parameters mentioned above from a raw F_0-slope [2]. As additional input it needs information about phrase boundaries and maxima locations and the onset time of the accented vowel. The algorithm first determines the *delay* from the maximum position, second the *amplitude* relative to top- and baseline. Finally, the F_0-slope left and right of each maxima are approximated as $\cos^2$ functions. The optimization uses the least square error, which is increased next to the maxima, thus ascribing them a bigger importance.

In most cases, no auditive differences appeared between the original and the parametrized and resynthesized contours; the functional aspects of a contour certainly remained unchanged. A quantitative analysis showed a high correlation between original and parametric contours (see fig.2).

In the database, the model parameters are labelled only for syllables associated with an F_0-maximum. A very important and also difficult task in the future will be to determine these syllables using only the text-input into a synthesis-system.

Duration: The database was segmented automatically [3]. Syllable duration was determined using the segmentation output. Moreover, the duration of the coda (i.e. the sounds following the syllable nucleus), the syllable onset (i.e. the sounds preceding the syllable nucleus) and the duration of the nucleus itself are given.

Boundary-types: Two types of prosodic boundaries are distinguished: progredient and non-progredient. The distinction was made using exclusively the slope of the F_0 at the previously perceptively determined boundaries (see below). Rising contours are labelled as progredient, falling contours are labelled as non-progredient.

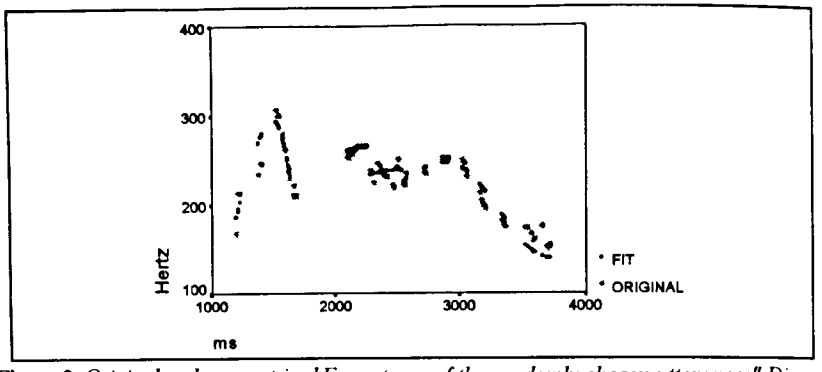


Figure 2: Original and parametrized F_0-contours of the randomly chosen utterance:" Die Schüssel mit Äpfeln haben wir auf dem Küchentisch gedeckt"; female speaker

In the database, all information is given in respect to syllables: each syllable contains information about its distance to the following boundary and about type and strength of this boundary.

PERCEPTIVE INFORMATION

Syllable-prominence: The perceptive prominence of each syllable was rated by three subjects. The rating was done on a graphical scale ranging from 0 to 31. The adequacy of this method has been proved by Fant & Kruckenberg [4]. The subjects could listen to an utterance as often as they needed to, and they could determine the size of the presentation frame. We found a high inter-subject correlation ($\rho = 7,9$).

Place of boundaries: The boundaries were set by two labellers experienced in prosody work. They consulted each other in problematic cases, but most of the labelling was done by one person alone. Both the auditive and the instrumental analysis (graphical presentation of F_0-slope and waveform) were used. In cases of doubt, the auditive impression was of prime importance.

LINGUISTIC INFORMATION

Linguistic accentability: The syllables with a nucleus consisting of a schwa-sound or a vocalic /r/ as well as the German suffix /IC/ were labelled as unaccentable, all other syllables were labelled accentable.

Syllable structure: The number of sounds in each syllable as well as the number of sounds in syllable onset and coda is determined. Additionally, information about the type of articulation of the sounds in onset, nucleus and coda is given.

Boundary prominence: For both the rising and the falling boundary type, four degrees of boundary prominence are distinguished. The distinction is made after the perceptual determination of the boundary place. The lowest level was labelled when a prosodic phrase occured within a sentence. The second level describes boundaries after rising or falling wh-questions. The third level was labelled at the end of a declaration or an decision question. The highest level of prominence is labelled at the end of a text or an isolated sentence or an echo question.

CONTEXT INFORMATION

As mentioned earlier, the information about each syllables distance to a following boundary is given. Furthermore, the distance to the preceding and the following F_0-maxima are labelled (in syllables and in milliseconds) as well as the number of voiced sounds following the syllable-nucleus up to the next nucleus.

CURRENT APPLICATIONS

The database is used in two ways: in a data-driven system and to develop rules for duration and intonation. The data-driven system (a neural network) generates the model parameters and the syllable duration in one step [5]. As input it uses most of the parameters represented in the database.

In the rule-based prosody systems, duration and F_0 are generated in two steps. The duration is generated with a syllable-based model. A model to predict syllable durations was worked out using the new database [6]. It uses information about utterance-finality, number of sounds within a syllable, perceptive prominence and syllable structure. This duration control is already implemented in the Bonn speech synthesis system.

As for F_0, a preliminary model has been implemented as well. It predicts mean values for the model parameters *delay*, *amplitude* and *slope* using information about boundary distance, the type of boundary, the place of the syllable within a given utterance and three degrees of syllable prominence. Although this F_0-generation is very simple, its output is acceptable. This is at least a hint to the apropriateness of our description.

The database is further used to investigate more extensively the relations between perceptive and acoustic prominence [7].

CONCLUSION

A database wich includes perceptual acoustic and linguistic information proved to be a valuable research tool. A new method of parametrization shows good results for resynthesis of natural speech. If it is really adequate for the generation of prosody will have to be proved by future research.

REFERENCES

[1]Möbius, B. (1993): *Ein quantitatives Modell der deutschen Intonation - Analyse und Synthese von Grundfrequenzverläufen.* Tübingen: Niemeyer

[2]Portele, Th.; Krämer, J.; Heuft, B.; Sonntag, G.(1995): Parametrisierung von Grundfrequenzkonturen. *Fortschritte der Akustik-DAGA'95*, Bad Honnef

[3] Wesenick, M.B.; Schiel, F. (1994): Applying Speech Verification to a Large Data Base of German to obtain a Statistical Survey about Rules of Pronunciation. *Proc. ICSLP'94* pp 279-282

[4] Fant, G. & Kruckenberg, A. (1989): Preliminaries to the study of Swedish prose reading and reading style. STL-QPSR 2/1989, pp. 42-45

[5] Portele, T.; Reuter, A.; Heuft, B. (1995): Generating synthetic prosody with a neural network. Submitted to: Eurospeech'95

[6] Meyer, H.; Portele, T.; Heuft, B. (1995): Ein Silbendauermodell für die Sprachsynthese. Fortschritte der Akustik, DAGA'95, Bad Honnef.

[7] Heuft, B.; Portele, T.; Höfer, F.; Meyer, H.; Rauth, M. (1995): Betonungsstufen von Silben und ihre Beziehung zum Sprachsignal. *Fortschritte der Akustik-DAGA'95*, Bad Honnef.

[8] Portele, T.; Heuft, B.; Höfer, F.; Meyer, H.; Horst, W.(1994): A New High Quality Speech Synthesis System for German. *Progress and Prospects of Speech Research and Technology*, München

COMPARISON OF PROSODIC CHARACTERISTICS IN ENGLISH, FINNISH AND GERMAN RADIO AND TV NEWSCASTS

A. Iivonen, T. Niemi and M. Paananen
Department of Phonetics, Helsinki, Finland

ABSTRACT

Some prosodic features in English, Finnish, and German radio and TV newscasts are compared with each other. These features include pause lengths, speech rate, articulation rate, F0 contours, and F0-distribution.

NEWS STYLE CHARACTERISTICS

There are several features common to newscasts all over the world such as monology, reading aloud (speech execution with previous verbal planning), economical time consumption, standard pronunciation, very clear articulation, correct grammar, quite complex syntax, objective ("neutral") but convincing presentation, and the absence of immediate listeners. Besides, certain speech acts, eg. questions or exclamations, cannot occur in newscasts, which limits the prosodical patterning. This does not, however, mean that there would exist a homogene *genre* of newsreading. We have discovered differences between languages, media, channels, and speakers. The speech style in newscasts is influenced especially by two factors: 1) the tendency to create and maintain a homogene style consisting of recurrent prosodic patterns characteristic of just one programme type (representing a channel or speaker), and 2) the maintaining of a certain homogene attitude within the programme type.

When news in several languages are compared with each other, it is not easy to distinguish stylistic and language specific features from each other. Our aim is not to make a clear distinction between style and language.

MATERIALS

The following radio and TV channels were recorded between March 29 and April 4 1994 (number of sentences):

- English (British = BrE)(49): SKY, ITN, BBC
- English (American = AmE)(46): NBC, CBS, TODAY, WORLDNET, VOA
- Finnish (Fin)(96): STT, YLE (radio), TV1, TV2, MTV (commercial)
- German (North = Ger)(44): RTL, ZDF, Deutsche Welle

MEASUREMENTS AND METHODS

In order to describe the mannerisms in newsreading we have investigated several macro and micro prosodic parameters.

- the speech rate (duration of the sentences and internal pauses divided by the number of syllables) and articulation rate (without internal pauses);
- pause lengths (within and between sentences);
- global prosodic features such as the use of F0-contours; F0 was measured mainly syllable by syllable; if special tones occurred within a syllable, more F0 points were measured; attention was paid to the F0 grid in sentences;
- the local use of F0-contours to express stress-groups, constituent structures and the thematic and rhematic parts of the sentences;
- F0-distribution (mean, standard deviation, minimum and maximum values of all observed glottal periods).

All metasentences were excluded.

The ISA speech processing system (*Intelligent Speech Analyser*; software designed by Raimo Toivonen) was used for analysis and documentation. A sophisticated multi-checking procedure based on zooming the low part of the frequency domain was applied to F0 measurements.

The intensity of the syllables was also measured, but the results are not reported here, because it is technically controlled during the broadcasts. The sentence final unvoiced and laryngealized portions have some effect on the statistics. 100% of the observed periods were included.

The way the material is grouped for calculations affects the statistics, too.

RESULTS

Pausing

The shortest pauses within the sentences (272.8 ms) and between the sentences (442.3) were found in the AmE newscasts (Table 1). The other language variants had in average longer pauses: BrE (295.8, 487.5), Ger (388.4, 549.9), and Fin (481.5, 752.1), in this order.

Table 1. Contrastive information on pausing.

Language	pauses within	pauses between
Finnish	481,5	752,1
German	388,4	549,9
Br.Engl.	295,8	487,5
Am.Engl.	272,8	442,3

Averages: pauses within sentences and pauses between sentences

Speech and articulation rate

Table 2 shows the statistics. The number of speakers (18) and sentences (96) is greatest in Finnish material. The number of speakers varied between 5 and 9 in the other materials. The longest average sentence duration was found in AmE (mean 6028.8 ms). The other language variants showed the following values: 5987.9 (Fin), 5381.1 (Ger), 4838.9 (BrE).

The shortest mean syllable duration was found in Finnish (160.6 ms). German had longer values (175.5). AmE and BrE had the longest syllables (194.1 and 192.4). The shorter syllable duration in Finnish might depend on longer words in Finnish: bisyllabic word is the most frequent structure, but words with four and five syllables are not rare. Therefore the stress groups get longer and the mean syllable duration becomes shorter according to a rhythmic principle (isochrony).

Consequently, the average number of syllables per second varies in different languages. The order is English (BrE 5.4 and AmE 5.4), German (5.9), and Finnish (6.5). The individual speech rates can, however, vary considerably. This concerns especially the American readers, who had the following personal values: 5.6, 5.6, 5.1, 5.4, 4.9, 4.8, 5.9, 5.6. 4.7. The British readers showed less variation: 5.1, 5.2, 5.1, 5.5, 5.3, 5.3, 5.4, 5.1.

Articulation rate seems to follow practically the above patterns (mean syllable duration and number of syllables per second): Finnish (155.8 ms; 6.5 syllables), German (170.6; 5.9), AmE (187.6; 5.4), and BrE (188.2; 5.4).

Global and local prosodic features

We have paid attention to the global features of the F0 contour of a single newscast as a textual unit (cf. Fig. 1), sentences as textual units and in relation to their contexts, and clauses as units within sentences. Our analysis is not yet completed, but the following qualitative and functional features can be reported. All four language variants use extra high F0 peaks at the beginnings of newscasts (cf. [1], [2]). Sentence initial F0 level depends very much on the thematic connection to the previous sentence. Low beginnings imply close semantic connections; higher beginnings start more independent topics. The newscast final syllables are most typically the lowest ones. In Fig. 1 the sentence final

Table 2. Contrastive information on speech and articulation rate.

Lang.	N speaks	N sent.	average dur	speech rate ms/syllable	syll/ second	artic. rate ms/syllable	syll/ second
Finnish	18	96	5987,9	160,6	6,3	155,8	6,5
German	5	44	5381,1	175,5	5,8	170,6	5,9
Br.Engl.	8	49	4838,9	192,4	5,3	188,2	5,4
Am.Engl.	9	46	6028,8	194,1	5,2	187,6	5,4

Sums (N): number of speakers, number of sentences
Averages: duration/sentence (ms), ms/syllable, syllables/second

syllables are equally low. The F0 declination happens to be most obvious in the first sentence in Fig. 1, but this feature is not a regular one.

The peaks and valleys are connected with the word stresses and the syntactical relationships between words. The higher peaks within sentences indicate special emphasis, usually contrastive stress. The intervals between peaks and valleys (accented maxima and unaccented minima which define the *grid*) can be kept more or less constant. This feature seems to have a strong stylistic effect. It is most typically used in Finnish radio newscasts (STT)(Fig. 2). The relative effect of the grid can be shown on a semitone scale. Fig. 2 contrasts an example of Finnish radio news (STT) with the commercial TV news (MTV). The intervals comprise a larger variation span in the latter case.

The empirical application of the grid notion often causes great problems (cf. Fig 1, sentences 2 and 3).

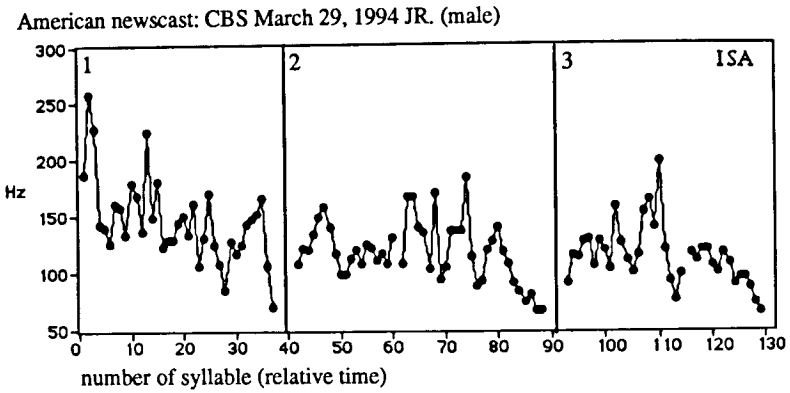


Figure 1. F0 contour (with relative time) of an American newscast consisting of three sentences. One F0 point per syllable was measured, in syllables with special tones of the F0, more points were measured.

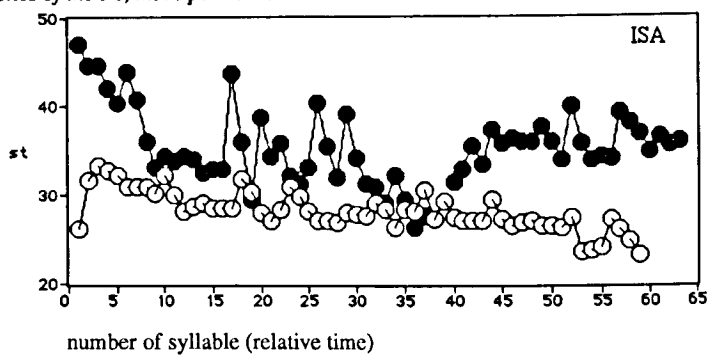


Figure 2. Comparison of one sentence by two male speakers representing Finnish radio (STT) news (= white dots) and commercial MTV news (= black dots). Comparison on a semitone scale shows the relative monotone F0 contour (= a more narrow grid) of the STT speaker in relation to a more variable MTV speaker.

Table 3. Contrastive information on fundamental frequency distribution.

Lang.-sex	N speaks	N sent.	N periods	F0 (st)	sd (st)	F0 (Hz)	range (st)	min (st)	max (st)
Fi-F	7	26	11056	39,8	2,9	164,6	16,2	30,1	46,3
Ge-F	2	17	9887	42,7	3,0	193,3	15,1	35,4	50,5
Br-F	4	23	11237	41,8	2,7	183,2	12,5	35,5	48,0
Am-F	3	14	6400	42,4	3,3	193,2	16,5	33,8	50,3
Fi-M	12	70	23307	32,0	2,6	105,4	14,6	24,3	38,7
Ge-M	3	27	6053	32,8	3,3	109,8	15,1	25,5	40,6
Br-M	4	26	9139	36,9	3,3	139,2	15,9	28,7	44,5
Am-M	6	32	15766	35,2	3,7	126,1	19,1	25,1	44,1

F=females, M=males
Sums (N): Number of speakers, number of sentences, number of periods
Averages: F0 semitones, standard deviation, F0 Hz, range, minimum and maximum values in semitones

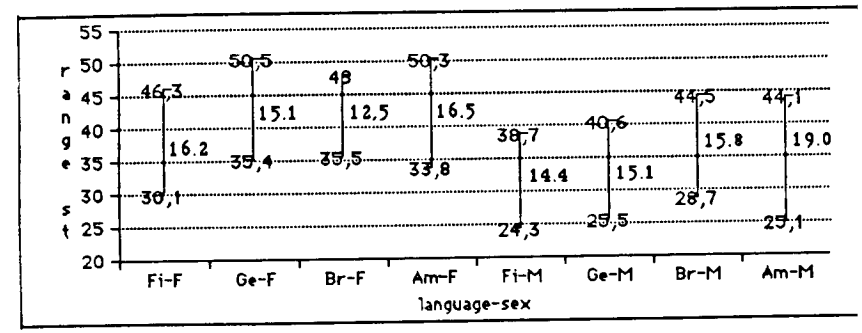


Figure 3. Average ranges of F0 in semitones among female and male speakers in Finnish, German, British English, and American English newscasts. The minima, maxima and the total range are indicated.

F0 distribution

Table 3 shows the statistics of the F0 distribution. The min. and max. values and the variation range of F0 are mainly affected by the intervals between F0 peaks and valleys as well as high sentence starts and low ends.

In the female data, the highest F0 mean values were observed in Ger (193.3 Hz; 42.7 st) and in AmE (193.3; 42.4). BrE had medium values (183.2; 41.8). Seven Finnish speakers had considerably lower values (164.6; 39.8).

In the male data, the highest mean F0 values were found in BrE (139.2 Hz; 36.9 st). American men had the medium values (126.1; 35.2). The lowest values were observed in Ger (109.8; 32.8) and in Fin (105.4; 32.0) males.

Fig. 3 shows among other things that the Fin female F0 ranges are only about 2 st higher than those of the BrE males.

REFRENCES

[1] Enkvist, N.E. & Nordström, H. (1978) On textual aspects of intonation in Finland-Swedish newscasts. *Speech and Lan*guage, vol. 32, pp. 63–108.

[2] Lehiste, I. (1975) The phonetic structure of paragraphs. A. Cohen & S. G. Nooteboom (eds.) (1975), *Structure and Process in Speech Perception*. Berlin ect.: Springer, pp. 195–203.

MODELLING INTRA- AND INTER-SPEAKER PITCH RANGE VARIATION

D. Robert Ladd and Jacques Terken***
**University of Edinburgh, Scotland, **Institute for Perception Research, Eindhoven, The Netherlands*

ABSTRACT

This paper reports preliminary findings from a large-scale study of pitch-range variation within and across speakers. The purpose of the study as a whole is to understand better the ways in which pitch range may vary with different speaking modes and in relation to linguistic structure, within and across speakers. An important empirical goal of the study is the collection of a large amount of quantitative data. We report results of comparisons between normal speaking mode, raised voice and local emphasis, and discuss some preliminary conclusions.

INTRODUCTION

Beginning with Bruce [1], numerous instrumental studies (e.g. [2,3]) have pointed to the existence of relatively invariant pitch targets in intonation contours, normally local maxima or minima. This finding is the basis of autosegmental descriptions of intonation in terms of H(igh) and L(ow) tones, such as the influential description of English intonation proposed by Pierrehumbert [4]. Likewise, in descriptions of intonation based on pitch movements rather than pitch targets (e.g. [5]), it is assumed that the beginning and ending levels of movements can be accounted for by a quantitative model. We assume without further comment the validity of the idea that targets have a reality of some sort, and an important role to play in the quantitative modelling of intonation.

At the same time, however, it is clear that the phonological description of tonal targets is still controversial, in part because there is no general agreement about the nature of tonal targets even among proponents of autosegmental analyses. A significant source of disagreement in these debates is the treatment of pitch range variation. It is obvious that a given speaker's voice may be higher or lower than another's, and it is similarly obvious that individual speakers may raise or lower their voices for a variety of reasons. What is not well understood, often due to the lack of empirical data, are the relations between such global within- and across-speaker differences, and their relation to other, more linguistic sources of variation in the "scaling" of individual pitch targets.

For example, it is known that F0 is generally higher at the beginning of paragraphs or other large discourse chunks, and that the F0 on individual accented words can be locally raised to convey emphasis. Is the raising of F0 in these types of cases quantitatively identical to "raising the voice"? Or again, if one speaker has a low monotonous voice and another a high animated voice, are these voice types quantitatively comparable to the differences within a single speaker between speaking monotonously and speaking animatedly? For that matter, what is the relation between "low" and "monotonous" or "high" and "animated": what would be the quantitative characterisation of a "high monotonous voice"?

This paper reports preliminary findings from a large-scale study focusing on these and related questions about pitch range variation. The purpose of the study as a whole is to understand better (a) the ways in which pitch range may vary within speakers; (b) the extent to which target scaling is invariant both within and across speakers; (c) what sort of scale (linear, logarithmic, or other) is most appropriate for characterising any invariance involved. An important empirical goal of the study is the collection of a large amount of quantitative data. This paper reports conclusions based on the analysis of some of the data from speech materials read aloud. Further analysis is in progress and fuller reports will be published in due course.

METHOD

Speech Materials

Our basic approach was to measure F0 at specific pre-selected points (e.g. utterance-initial unstressed syllable, first accent peak, utterance-final low in statements, low accent valleys in questions, etc.), in multiple repetitions of utterances with comparable contours. By doing this we hoped to establish stable mean pitch values for certain putative targets, creating a kind of "map" of the relative pitch of these targets which could then be compared between speakers or between different pitch-range settings of the same speaker.

For the study as a whole, we designed several sets of sentences intended to elicit specific intonation patterns which we expected would have consistent and identifiable peaks and valleys at well-defined points. These included ordinary statements of varying lengths, short questions, statements with explicit contrasts (of the sort "X not Y"and "Not X but Y"), and a news bulletin containing 18 short paragraphs. The recording session also included a short section of spontaneous speech (a description of the speaker's route to work). Only a small portion of the data is discussed and analysed here.

The materials discussed here fall into three main groups. Group 1 sentences contain two noun phrases with a total of four accented words; there are two subtypes, one in which both noun phrases have two accented words ("2-2") and one in which the first has three and the second only one ("3-1"). For example (accented words are written in capitals):

2-2: *Je moet de MOOIE ROZEN in een GELE VAAS doen* (You should put the pretty roses in a yellow vase)
3-1: *Je moet de MOOIE GELE ROZEN in een VAAS doen* (You should put the pretty yellow roses in a vase)

Altogether there were 8 such pairs, and each sentence was read twice, for a total of 32 sentences in group 1. Targets to be studied in this group are initial pitch, peak and valley on auxiliary verb, the four accentual peaks, medial valley between the two noun phrases, and the final low.

Group 2 sentences were all of the form

We zouden wel eens naar [X] kunnen gaan (We really ought to be able to go to [X] sometime)

in which X was one of four place names. Each of the 4 versions was read 4 times, for a total of 16 utterances. We intended that these should be accented only on the place name, though in the event many speakers put a weak accent on the auxiliary *zouden* as well. Targets to be studied in this group are initial pitch, weak accent on auxiliary, valley immediately preceding accent, accent peak, and final low.

Group 3 sentences were all of the form

Ik zei niet [X], maar [Y] (I didn't say [X], but [Y])

where X and Y were similar-sounding words that might plausibly be confused in a real situation, e.g. *mannetjes/lammetjes* (`little men / little lambs'). There were 4 pairs of words, presented in both possible orders, with 2 repetitions of each sentence, for a total of 16 utterances. Targets to be studied in this group are initial pitch, valleys preceding accents, accent peaks, final low, and both valley and peak of medial continuation rise.

In constructing the sentence materials we balanced prosodic, pragmatic, and segmental phonetic considerations.In particular, we avoided words with high vowels (to minimise intrinsic F0 effects) and obstruents (to minimise segmental perturbations of F0). Further details are beyond the scope of this report.

Recording and analysis procedures

For the recording sessions, all the materials were organised into 8 blocks, and the session lasted typically 75 minutes with a short break in the middle. The sentences discussed here were included in three of the blocks: one (Block 2) in which speakers read normally without any special instructions, another (Block 4) in which the speakers were told to raise their voice as if talking on a bad overseas telephone connection

(the situation was made more realistic by exposing speakers to rather loud (over 90 dB) non-steady noise over headphones), and a third (Block 7) in which individual words were capitalised and the speakers were told to emphasise those words. Of the materials discussed here, Block 7 included only the group 3 sentences ('not X but Y').

Speakers were 16 native speakers of Standard Dutch, 8 males and 8 females, all students or employees at the Institute for Perception Research (IPO), Eindhoven. This report is based on results from only 8 speakers, 4 males and 4 females.

The recordings were made in a quiet recording studio at IPO, using professional equipment. The sentences to be read were presented one at a time on a computer screen placed on a table in front of the speaker. The experimenter controlled the presentation from a neighbouring control room.

The recordings (on DAT tape) were transferred to the computer system at IPO and separate speech files were made for each sentence. F0 extraction was done by means of an algorithm based on subharmonic summation ([6]) with tracking. F0 values for each target were determined on the basis of time-aligned displays of the waveform and the F0 trace, obtained by means of an interactive wave form processing package developed at IPO. Details of the measurement criteria are beyond the scope of this limited presentation.

The maximum number of utterances per speaker in the portion of the study reported here was as follows:

	Normal	Raised voice	Local_ Emph.
Group 1 2-2	16	16	--
Group 1 3-1	16	16	--
Group 2	16	16	--
Group 3	16	16	16

In many cases one or more utterances had to be discarded because of disfluencies, etc.

RESULTS AND DISCUSSION

Sample data (for Speaker RS's Group 3 and Speaker RW's Group 1 sentences) are shown in Figs. 1 and 2. For all the speakers whose data we have analysed so far, the contours for each sentence group, and the patterns of modification for the different conditions, are strikingly similar. Quantitative modelling of the similarities is still only at a very preliminary stage. However, several findings are common to most or all of the speakers and must presumably be incorporated into any quantitative model. These are summarised here:

(1) There is a clear distinction between overall raising and local emphasis. The former raises both peaks and valleys, whereas the latter affects only peaks. This is seen in Fig. 1. This could be incorporated into a quantitative model by distinguishing two aspects of what is often loosely called "pitch range", namely the overall *level* and the *width* of the space in which tonal targets (or tonal movements) are realised. In overall raising of the voice, it is primarily level that is affected. In local emphasis, level is unaffected, but the width of the tonal space is expanded.

(2) Whereas many earlier reports (e.g. [2]) suggest that final F0 low is very stable for individual speakers, it appears that overall rasing also slightly raises the speaker's final F0 low. This is seen in both Figs. 1 and 2.

(3) For all targets, the effect of raising overall pitch range is extremely constant. For all speakers the correlation between targets in normal range and corresponding targets in raised range is extremely high (on the order of r = .90).

(4) It appears that the most invariant characterisation of the F0 relations in our data is achieved using an ERB scale [7], which at typical speech F0 levels is intermediate between the linear Hz scale and a logarithmic scale. This is reflected in the fact that range modifications look most similar across speakers when expressed in ERB: on a Hz scale overall raising is generally greater for men than for women, while on a log scale it is generally greater for women than for men. On the ERB scale the amounts by which males and females raise their voices are most comparable. Further discussion of quantitative details is beyond the scope of this paper.

REFERENCES

[1] Bruce, G. (1977), *Swedish word accents in sentence perspective*, Lund: CWK Gleerup.

[2] Liberman, M. and Pierrehumbert, J. (1984), "Intonational invariance under changes in pitch range and length", In Aronoff, M. and Oehrle, R., editors, *Language Sound Structure*, Cambridge: MIT Press,.

[3] Van den Berg, R., Gussenhoven, C., and Rietveld, A. (1992), "Downstep in Dutch: implications for a model", In Docherty, G. and Ladd, D., editors, *Papers in Laboratory Phonology II: Gesture, Segment, Prosody*, pp. 335-359. Cambridge, U.K: Cambridge University Press,.

[4] Pierrehumbert, J.B. (1980), *The Phonology and Phonetics of English Intonation*, Ph.D. thesis, Massachussets Institute of Technology, Cambridge MA (Bloomington, IN: Indiana University Linguistics Club).

[5] 't Hart, J., Collier, R. and Cohen, A. (1990), *A perceptual study of intonation*, Cambridge: Cambridge University Press.

[6] Hermes, D.J. (1988), "Measurement of pitch by subharmonic summation", *J. Acoust. Soc. Am.* Vol. 83, pp. 257-264.

[7] Hermes, D.J. and Van Gestel, J. (1991), "The frequency scale of speech intonation", *J. Acoust. Soc. Am.*, Vol. 90, pp. 97-102.

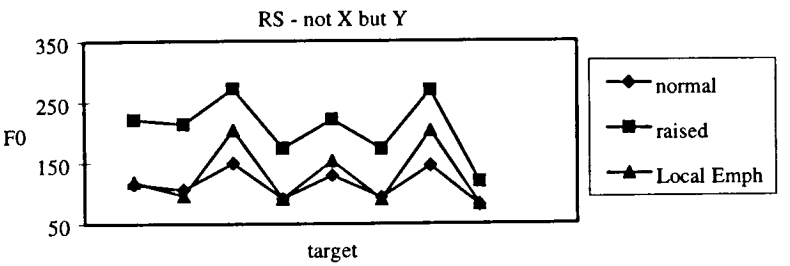


Figure 1. Average target values for successive targets (see text) in "not [X] but [Y]" utterances for speaker RS, in normal speaking mode, with raised voice, and with local emphasis on accented words.

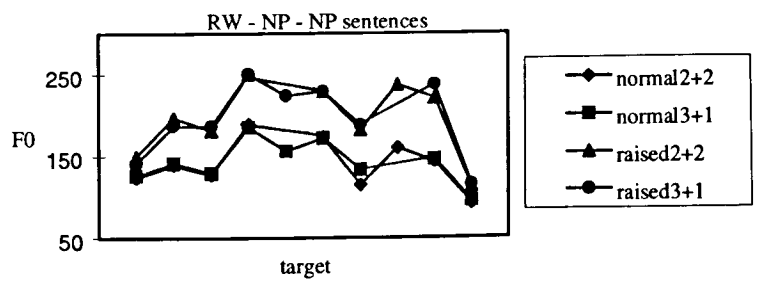


Figure 2. Average target values for successive targets in NP-NP sentences with 2+2 and 3+1 structure (see text) for speaker RW, in normal speaking mode and with raised voice.

IMPLEMENTING A FLOATING TONE

Yetunde O. Laniran, Northeastern University, Boston

ABSTRACT

This paper discusses the phonetic implementation of the floating Low tone in Yorùbá. It provides phonetic evidence that the floating L tone has the same effect as the lexical L and LH contour tones . It also shows that a LH (rising contour) tone differs from a L tone in resisting H tone spread from a preceding syllable.

1. INTRODUCTION

This paper examines the phonetic realization of both the lexical and floating Low (L) tones in Yorùbá. Yorùbá is a tone language with three tones, H(igh), M(id) and L. Two Tone Spread rules apply in Yorùbá, H tones spread to the following L tone syllable, H-Spread and L tones spread on to the following H tone syllable, L-Spread[1, 2]. A floating L tone is created in the phonology by the process of vowel deletion at V + V juncture. This L tone is either relinked or it remains floating [1, 2, 3]. The following examples illustrate the process. All the examples are verb + noun combinations except (1a), which is a noun + noun sequence: (H (´), M (unmarked) and L (`).

1a. ọmọ ọ̀kan → ọmọ!kan
child one "one child"

b. gbé ọ̀bẹ → gbọ̀!bẹ "carry knife"

c. gbé ọdẹ → gbọ́dẹ "carry hunter"

d. gbé àdá → gbádǎ "carry machete"

The tonal output is the same regardless of the deleted vowel [1, 2]. In examples (1a) and (b), the L tone remains floating before a M tone after vowel deletion. This creates a "downstepped" or "lowered" M tone, represented as ! before the syllable [3], [4], [5]). But the L tone links to a following H tone as illustrated in (1d). Ward [4] claims that if a H tone follows the lowered M tone, it is in turn pulled down compared to an earlier H tone in the sentence, that is, the downstep effect of the L tone on the M tone persists throughout the rest of the utterance. Similarly, Bamgbose [5] characterizes the floating tone as the "assimilated L tone" whose effect is to lower the fundamental frequency (f0) values of the following tones.

The experiment discussed below examines:

- the effects of the floating L tone on M tone, that is, to see the difference(s) in the realization of the M tone and the !M;
- the realization of the rising tone (LH) in the same context as M and !M tones;
- the effects of the floating L, the rising tone, LH and the lexical L on preceding and following tones;
- the realization of a sequence of M!M M tones, that is to see if the lowering of M tone persists to the next syllable (and beyond).

2. EXPERIMENT

The sentences below were recorded by four native speakers of Yorùbá as part of a larger study reported in Laniran [3]. Two factors were systematically varied in these sentences: the target tones (T) which are underlined, and the following tones (F). In each set of sentences, the following tone is H in (a), M in (b) and L in (c). The target tone is M in Set I, !M in Set II, LH in Set III. In Sets IV and V the target tones are (T_1=M and L, and T_2 =!M and M respectively.

Set I
a. Mo gbọ́dẹ̀ lálẹ́
'I carried the hunter at night'
b. Mo gbọ́dẹ̀ lọ sílé
'I carried the hunter home'
c. Mo gbọ́dẹ̀ bọ̀ lálẹ́
'I carried the hunter back at night'

Set II
a. Mo gbọ́!bẹ̀ lálẹ́
'I carried the knife at night'
b. Mo gbọ́!bẹ̀ lọ sílé
'I carried the knife home'
c. Mo gbọ́!bẹ̀ bọ̀ lálẹ́
'I carried the knife back at night'

Set III
a. Mo gbádǎ̀ lálẹ́
'I carried machete at night'
b. Mo gbádǎ̀ lọ sílé
'I carried machete home'
c. Mo gbádǎ̀ bọ̀ lálẹ́
'I carried machete back at night'

Set IV
a. Mo gbọmọ̀ !mi lálẹ́
'I carried my child at night'
b. Mo gbọmọ̀ !mi lọ sílé
'I carried my child home'
c. Mo gbọmọ̀ !mi bọ̀ lálẹ́
'I brought back my child at night'

Set V
a. Mo gbalẹ̀ lálẹ́
'I got land at night'
b. Mo gbalẹ̀ lọ sílé
'I got land and took it home'
c. Mo gbalẹ̀ bọ̀ lálẹ́
'I got land back at night'

3. RESULTS

An average of 6-8 tokens of each sentence was analyzed each for all subjects but only three are discussed here. See Laniran [3] for a discussion of the other speaker, whose result is similar. The sentences were digitized at 10KHz using the Waves+ software by Entropics. On the graphs following, the f0 values in each syllable is represented by two measurement points (see [3] for a detailed discussion on methodology).

In Figure 1 (page 4), an overlay of the f0 contour for Sets I-III sentences where the target tones are M, !M and LH are shown. The legend for all the graphs is shown at the beginning. SB data was recorded last. To make all the sentences be of the same length the parenthesized segments were excluded from his data set.

The consistent difference in the f0 tracks for two of the three speakers reported here is that the H tone preceding the !M tone is realized higher at points *a* and *b* than those preceding the M tone (filled symbols dashed lines). For SB, the results are in the same direction but there is not a significant difference in the f0 values. Also for SB and YL, the f0 value for the !M is slightly lower than that of the M tone.

In Figure 2 (page 4), the f0 values for only the Set III sentences are presented. For BJ the rising tone (LH) is realized with a slightly rising f0 contour when the following tones are M and L. But in Figures 2b and 2c, the f0 contour in the LH syllable is falling (SB, YL). In Figure 2 graphs, the f0 contour at the beginning of the following L tone syllable is realized with a high f0 value, an f0 value higher than that in the corresponding H tone syllable.

In Figures 3 and 4 (page 4), the f0 contour for the sentences in Sets IV and V are presented. In Figure 3, the f0 contours from the second M tone syllable to the end of the !M tone in the fourth syllable are falling. The f0 values of the M tone syllable following the !M tone are about the same. Figure 4 shows the sentences in Set V. The f0 values in the M tone syllable following the L tone are higher. In addition, the L tone syllable has a falling slope. An examination of Figures 3 and 4 show that both the L tone and the !M tone have falling f0 contours. The distinction, as shown by an overlay of both figures for SB, is that the L tone has f0 values lower than that of the !M and M tone, Figure 5.

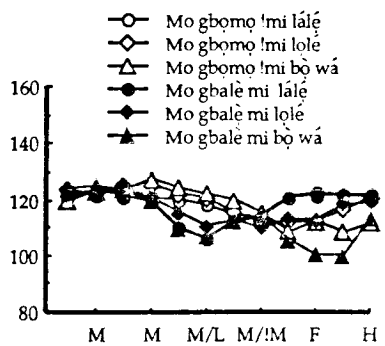


Figure 5: F0 Contour Comparing Target L, M and!M tones (SB)

4. DISCUSSION

The floating L tones survives from the phonology and is phonetically interpreted (rather than deleted) as shown in Figures 1a and 1c. Its main effect seems to be to raise the f0 value of the preceding H tone. The implication of this result is that a major cue to a !M tone seems to be the greater distance between a H tone and a following !M tone than that between a H and a following M tone for some speakers. SB's data is interesting in this regard since it is not clear from Figure 1b exactly what cues his listeners that the target syllable has a M as opposed to a !M tone.

If you recall, it was stated earlier that Ward [4] asserted that M and H tones occurring after a lowered M tone are realized lower than any preceding M and H tones in the utterance. This is the case as demonstrated in Figures 1 and 2, where the H tones occurring after the lowered M are not as high as the preceding H tone.

The f0 patterns for the rising tone (LH) were not all as expected in Figure 3. First, the application of the Tone Spread rule between the H and the following LH contour tone was suspended for all subjects, that is, no spreading takes place. For all subjects, point *c* on the graphs in Figure 2 shows no spreading effect of the preceding H tone. The expected pattern should have the f0 target for the H tone on the following L tone syllable as illustrated below for SB. Figure 6 shows an overlay of the Set IIIa sentence and a similar sentence where *gbádá* has been replaced with *gbóbè* (The sentences were recorded with those in Sets I-V.)

The H-spread rule is prevented from applying as shown by the difference in f0 values at point *c* in Figure 6, when the following tone is a contour tone (LH) as opposed to a single L tone (empty circles) which has a high f0 value. This is probably due to the fact that only two f0 targets are allowed in each syllable in Yorùbá. Since, the following syllable has two targets already, L and H, the Tone Spread rule does not apply.

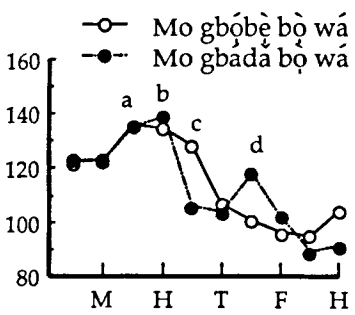


Figure 6: F0 Contour Comparing Target L and LH Tones (SB)

Point *d* in Figure 6 provides evidence that for the application of the H-Spread rule from the preceding LH contour tone to the following L tone. The f0 value at point *d* is higher when the preceding tone is LH as opposed to L supporting the hypothesis above that H-spread was blocked by a following contour tone.

5. CONCLUSION

The data presented in this paper has shown that there is no downstepped M tone in Yorùbá. Rather, the floating L tone preceding a M tone raises the f0 value of a preceding H tone for most speakers. The M tone preceded by the floating L tone itself does not consistently have an f0 value lower than that of other M tones in other environments. Therefore, the term "downstepped M tone "or "lowered M tone" is a misnomer.

It was also demonstrated that a LH (rising contour) tone differs from a L tone in resisting H tone spreading from a preceding syllable. The phonetic effect of the L tone, the floating L and the LH contour is the same on a preceding H tone. The F0 value in the H tone syllable is raised compared to when the following is non-L. The main difference seems to be that the H-Spread rule is prevented from applying to the LH syllable.

REFERENCES

[1] Pulleyblank, D. (1986). *Tone in Lexical Phonology*. (Reidel).
[2] Akinlabi, A. (1985). *Tonal Underspecification and Yorùbá Tones* unpublished Ph.D Dissertation University of Ibadan, Nigeria.
[3] Laniran, Y.O. (1992) Intonation in Tone Languages: the Phonetic Implementation of Tones in Yorùbá. Ph.D. dissertation, Cornell University.
[4] Ward, I.C. (1952) *An Introduction to the Yoruba Language*, CUP, Cambridge.
[5] Bamgbose, Ayo. (1967). "The Assimilated Low Tone in Yorùbá" *Lingua* 16: 1-13

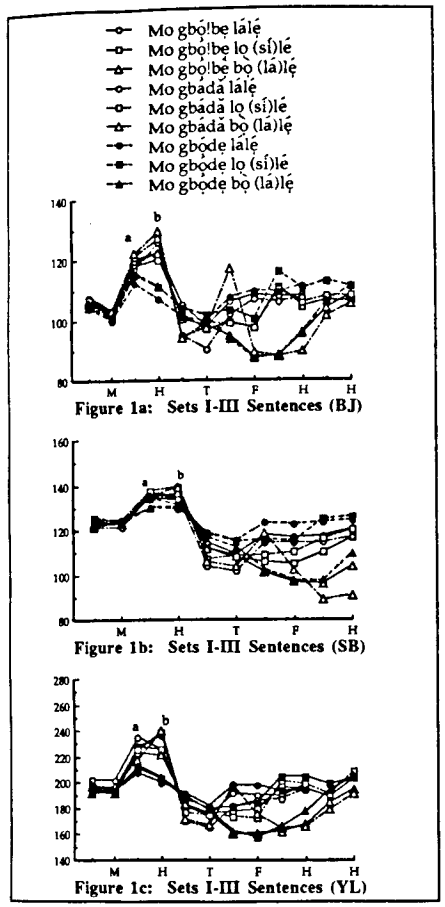


Figure 1a: Sets I-III Sentences (BJ)

Figure 1b: Sets I-III Sentences (SB)

Figure 1c: Sets I-III Sentences (YL)

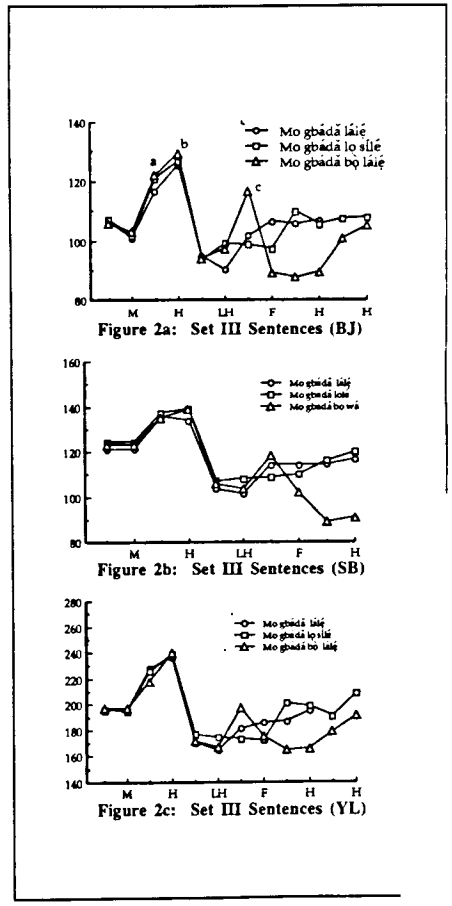


Figure 2a: Set III Sentences (BJ)

Figure 2b: Set III Sentences (SB)

Figure 2c: Set III Sentences (YL)

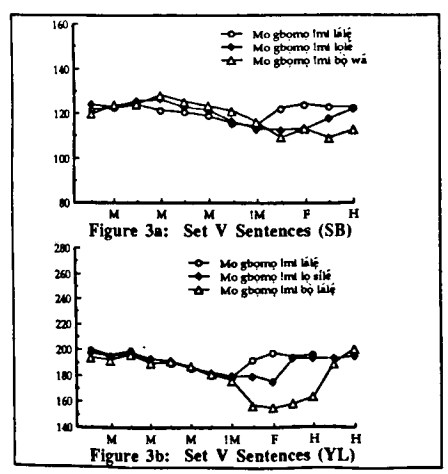


Figure 3a: Set V Sentences (SB)

Figure 3b: Set V Sentences (YL)

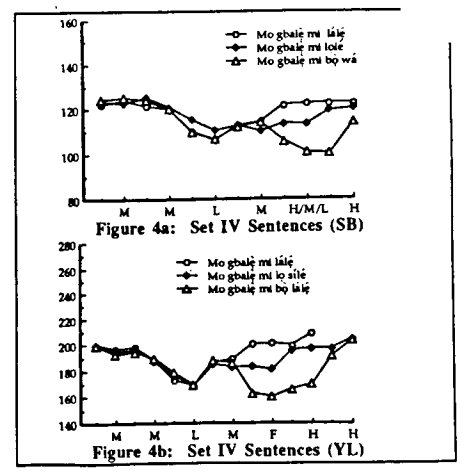


Figure 4a: Set IV Sentences (SB)

Figure 4b: Set IV Sentences (YL)

VOWEL INTRINSIC PITCH IN QUEBEC FRENCH: MEASURING IFO IN CONNECTED SPEECH

J. Lavoie and C. Ouellon
CIRAL, Université Laval, Québec, Canada

ABSTRACT

As it was established mostly on the basis of carrier sentence corpora built to that effect, the "intrinsic F0" phenomenon exists in all languages, including French. However, the values defined for France French by Di Cristo can hardly be applicable to Quebec French vowels. Being more interested to study IFO effects in connected speech, we decided to investigate a new procedure to calculate these on the basis of the interval between the stressed syllable and a reference line.

INTRODUCTION

In a first task, we have established IFO intervals for vowels in a carrier sentence corpus of Quebec French [1]. To make possible a comparison with France French values [2], our corpus was similar to the Di Cristo's one, that is to say carrier sentences containing a target vowel in a CVC stressed syllable like the following: *Le CVC de (NP) est (copula)*. We analyzed a minimum of 30 tokens of each vowel. An interval of more than 2 semitones was calculated between High and Low vowels (Figure 1). These IFO values are consistent with those calculated for most languages [3].

Meanwhile, our main objective was to find out if these results would be the same in connected speech and, ultimately, in spontaneous speech, that is to say the type of informal language used in a sociolinguistic interview. So we designed a new procedure to calculate IFO in connected speech, which could be considered as intermediate between carrier sentences and spontaneous language. If such a new procedure is functionning correctly, we assume it should work with spontaneous speech.

Some authors [4] [5] [6] [7] analyzed vowel IFO in connected speech. Umeda was the only one, as far as we know, who used a corpus similar to ours. She found there were no significant IFO intervals in connected speech, but her findings were afterwards rejected. Further works calculated IFO variations between High and Low vowels in connected speech. But the authors [6] [5] [4] did not use real connected speech corpora in the sense that most often their works were based on target vowels which commuted in preestablished stressed positions in a text. In such conditions, IFO variations were estimated to be smaller than in carrier sentence corpora.

METHODS

Second task corpus

A 600-syllable text was selected from a reading book; we neither modified the text nor considered its phonetic characteristics. It was recorded in excellent acoustic conditions. Subjects were four Quebec French speaking university students (2 males and 2 females) aged between 20 and 30, and living in Quebec City. They also read the carrier sentence corpora. The data were processed and analyzed with the CSL System (Kay Elemetrics).

Measurements

The main difference between connected speech and carrier sentence corpora for purposes of IFO analysis is that in the latter, it is possible to take direct FO measurements [2] of target vowels which always appear in the same position under controlled phonetic variables. In such conditions, mean IFO intervals can be quite easily calculated for each vowel. A connected speech corpus like the one chosen does not allow such a procedure; there is neither any target vowel, nor any real control of stressed vowels or phonetic environment. Moreover, the intonation variation is important enough to forbid any direct F0 calculations for IFO estimations purposes.

So, we believe it is necessary to consider IF0 values by taking into account, to a certain extent, the prosodic context of the retained vowel. A prominent vowel, most of the time uttered with an increase of F0 over a baseline, is consequently evaluated in comparison with this line [8] [9] [10] [11]. Although there is no consensus on the exact nature of such a baseline, we agree with Ladd's proposal [10]: "We need to acknowledge that the key to normalizing prominence is some sort of abstract reference value in a comprehensive model of pitch range".

Practically, we assume that unstressed vowels constitute points of the said reference line and consequently we decided to retain the interval between the stressed vowel and the preceding unstressed one to evaluate stressed vowel IFO.

Thus, stressed syllables in the four recordings were localized, without a hierarchy being established between primary and secondary stresses. We did not consider final position stress, which was always at the end of a decreasing intonation line in our corpus. We also neglected to appreciate the preceding unstressed vowel nature since IFO in this particular position would not be really conditionned by vowel aperture [12].

Finally, on the basis of the first task which showed us vowels of similar aperture having comparable IFO variations in a carrier-sentence corpus, and because there was not a significant number of occurrences for each vowel in the second task corpus, we grouped them in classes (High tense, High lax, Mid, Low and Nasal) for calculation of IFO intervals. Pitch values were measured at the mid point of vowel duration, where the influence from surrounding segments is seen as minimal.

RESULTS

Task 2 vs Task 1 results

Figure 1 gives a view of the comparison between vowels in both tasks, that is to say carrier sentence and connected speech corpora. A similar IFO classification can be seen in both situations, IFO diminishing regularly from High to Mid to Low and Nasal vowels. However, there is a

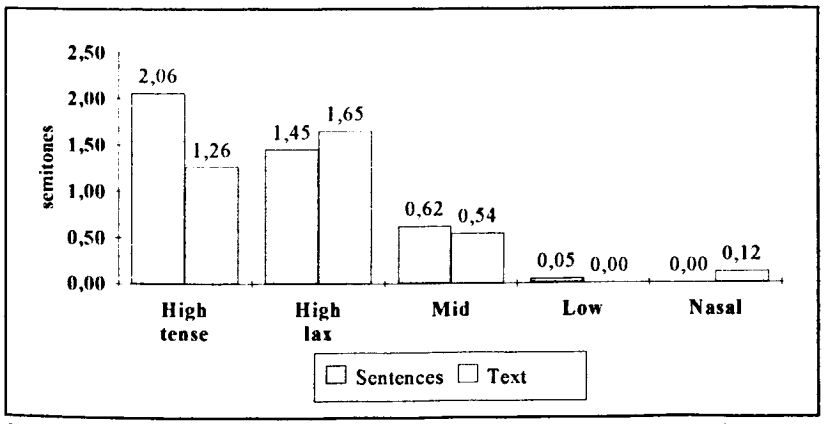


Figure 1. Intrinsic F0 of Quebec French vowels in sentences and in connected speech

noticeable difference: High-tense and High-lax vowel ranks are inversed in connected speech data, High-lax being 0.4 semitone higher than High-tense. This is consistent with some observations for German vowels [13].

At first sight, it seems that IFO intervals are less important in Task 2, High-Low interval being 1.65st compared to 2.06st in Task 1. Such results confirm Ladd & Silverman's [6] findings but they might not be meaningful in our corpus. As a matter of fact, it is important to recall that Task 2 vowels were indifferently under primary or secondary stress conditions, which theoretically should explain a reduction of the High-Low interval.

Detailed IFO calculations for Task 2 are shown in Table 1. High-tense and High-lax classes were grouped into High class. There is evidence for the IFO effects to exist for the four subjects. It seems that female subjects, F1 and F4, are characterized by more important IFO intervals. However, this has to be validated since an inverse tendency, even not significant, was noted by Whalen & Levitt [14].

Table 1. Interval values (in semitones) calculated between the stressed vowel and the preceding unstressed one in Task 2

Class	Interval				Mean
Speaker	*F1*	*M2*	*M3*	*F4*	
High	3.3	4.5	3	2.4	3.3
Mid	2.2	3.4	2.3	1.7	2.4
Nasal	1.5	3.3	2.1	0.9	2
Low	1.7	3	1.8	0.8	1.8
H-L Interval	1.8	1.5	1.2	1.6	1.5

Quebec French vs Di Cristo's French values

As far as we know, there are only two studies on French vowel IFO [2] [14]. As we said earlier, we designed Task 1 expressly to allow the comparison with Di Cristo's findings. Figure 2 illustrates results for both dialects. Two facts are of evidence. On the one hand, the Quebec French High-Low mean interval is twice as important in comparable phonetic contexts, that is in Task 1 and Di Cristo's corpus. On the other hand, more important is the relative position of Nasal vowels which have similar IFO values than High vowels in France French. On the contrary,

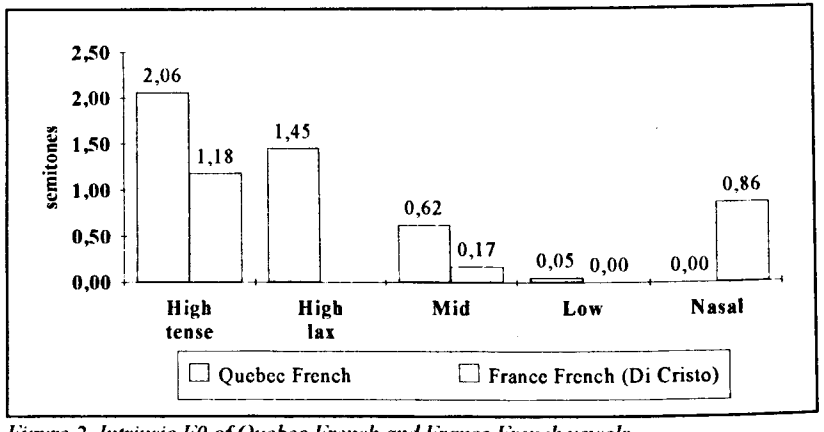


Figure 2. Intrinsic F0 of Quebec French and France French vowels

Nasal stand on Low vowel level in both Quebec French tasks. Di Cristo failed to find an appropriate explanation for the Nasal vowel particular position in his IFO scale; notwithstanding its flaws, the tongue pull theory accounts for Quebec French Nasal vowel classification.

DISCUSSION

For Task 1, standard deviations were calculated at 5% of measured F0 values. This is not the case in the connected speech corpus, for which s.d. go up to 75% of F0 measurements. Such an important variation was observed by Ladd & Silverman [6] in a connected speech corpus. If we also consider the IF0 intervals for each subject in Table 1, it must be admitted that IF0 calculation is characterized by an important variability, increasing with the formal style of the corpora. Such a variability exists between subjects, from a register to the other, from a dialect to the other, and so on. How could it be possible, in these conditions, to propose IF0 correction factors for intonation description purposes?

In future works, we will refine our IFO measurement procedure for connected speech data. In order to do so, we will look for a more satisfactory description of our reference line inspired by Ladd's proposal [10] of a model "whose reference values lie between the valleys and the peaks - which is where both Liberman and Pierrehumbert's reference line and the 'zero line' in the middle of Ladd's 'tonal space' are located". The last part of our research will be an application of this revised procedure to more informal corpora.

ACKNOWLEDGEMENTS

We would like to thank Jean Dolbec (Université du Québec à Chicoutimi) and Marise Ouellet (Université Laval) for their help in data collection and analysis. This study is a part of PROSO (dir. Claude Paradis) which was supported by SSHRC Grant 410-90-1410 and FCAR Grant 92-ER-1111.

REFERENCES

[1] Lavoie, J. (1994), "La fréquence intrinsèque des voyelles en français québécois", *Actes des 8èmes Journées de linguistique*, CIRAL B-197, Université Laval, Québec, 109-113.

[2] Di Cristo, A. (1985), *De la microprosodie à l'intonosyntaxe*, Publications de l'Université d'Aix-en-Provence, Aix-en-Provence.

[3] Whalen, D.H. and A. G. Levitt (to be published), "The universality of intrinsic F0 of vowels", *Journal of Phonetics.*

[4] Steele, S.A. (1986), "Interaction of vowel F0 and prosody", *Phonetica*, 43, 92-105.

[5] Shadle, C.H. (1985), "Intrinsic fundamental frequency of vowels in sentence context" *J.A.S.A.*, 78, 1562-1567.

[6] Ladd, D.R. and K.E.A. Silverman (1984), "Vowel intrinsic pitch in connected speech", *Phonetica*, 41, 31-40.

[7] Umeda, N. (1981), "Influence of segmental factors on fundamental frequency in fluent speech", *J.A.S.A.*, 70(2), 350-355.

[8] Terken, J. (1991), "Fundamental frequency and perceived prominence of accented syllables", *J.A.S.A.*, 89, 4, 1768-1776.

[9] Terken, J. (1993), "Baselines revisited: Reply to Ladd", *Language and Speech*, 36, 4, 453-459.

[10] Ladd, D.R. (1993), "On the theoretical status of "the baseline" in modelling intonation", *Language and Speech*, 36(4), 435-451.

[11] Ladd, D.R., J. Verhoeven and K. Jacobs (1994), "Influence of adjacent pitch accents on each other's perceived prominence: two contradictory effects", *Journal of Phonetics*, 22, 87-99.

[12] Reinholt Petersen, N. (1978), "Intrinsic fundamental frequency of Danish vowels", *Journal of Phonetics*, 6, 177-189.

[13] Fischer-Jørgensen, E. (1990), "Intrinsic F_0 in tense and lax vowels with special reference to German", *Phonetica*, 47, 99-140.

[14] Rossi M. and D. Autesserre (1981), "Movements of the hyoïd and the larynx and the intrinsic frequency of vowels", *Journal of Phonetics*, 9, 233-249.

TOWARDS AN INTONO-DISCURSIVE SEGMENTATION OF SPONTANEOUS UTTERANCE IN FRENCH

*C. Leleu *$, A. Lacheret-Dujour $@ and M-A. Morel **
**Paris III University, Paris, France*
$Limsi-CNRS, Orsay, France
@ELSAP, Caen, France

ABSTRACT

Our goal is to define a set of phonetic markers that can be used to segment an utterance and to predict discursive boundaries in spontaneous French speech. Our acoustic analysis of descriptive monologues is carried out on two prosodic parameters and the relationship between them: (1) melody, (2) pauses.

1 INTRODUCTION

With the hypothesis that prosody has a demarcative function in the structuring of a spoken text into discourse units, we propose a method of segmentation based on high and low points of melody and pauses. For the present, this method is not totally discursive because the prosodic status of functional words is rather ambiguous; nevertheless, we will show how the coupling of cues can be efficient.

The entity of our analysis is no longer the sentence but the paragraph which we divide into intonative and pausal constituents (refered to hereafter as **IC**s and **PC**s respectively).

Our methodology is based on the hypotheses of our research team [1, 2] on the intonative and enunciative structuring of spontaneous utterances which are given by the variation in the F0 level-points of the final syllable of prosodic constituents. We also rely on the work of some linguists, for example Mertens [3], Rossi [4], etc., on the structuring of French utterances by intonation. This linguistic analysis is however restricted in our work because we foresee formal constraints of implementation and also because our database is highly specific. Nonetheless, the scope of our analysis is made larger with pausal constituents added to intonative ones.

First, we will describe the database used to carry out our study. Next, we will describe the different stages of the segmentation allowing us to derive the discursive structure from acoustic cues. Then, we will show the formation of PCs and propose hypotheses on the relations between pausal and intonative boundaries. Finally, intono-discursive markers are interpreted on the enunciative level and corresponding boundaries are given a specific weight relative to their position in the hierarchy.

2 ANALYSIS

2.1 Description of the Database

Our data was collected from the ICY database developped in Limsi by V. Péan [5] to study inter- and intra-speaker variability. ICY was obtained from a strategy by which utterances elicited are both spontaneous, i.e. generated on the fly, and under the experimenter's control. The speakers were instructed to describe two identical pictures differing only in the colour or spatial positions of several objects chosen deliberately to make the speakers produce groups of words which contain a phonological context at word boundaries (for example: in the case of "pantalon orange", we can find a possible nasalization context). They did not know the real purpose of this experiment and were given a false pretext to persuade them to produce three different styles of speech of which we have studied the casual version.

As a result, such constraints give a corpus that consists of descriptive monologues with declarative sentences. They also explain why the resulting speech does not show as much variation as in real-life situations. However, this type of spontaneous discourse is more realistic than read-aloud, often isolated sentences generally used in automatic speech processing.

The pauses and F0 analysis were obtained using the UNICE software created in Limsi-CNRS (Orsay). In this article, we study 5 speakers among the 21 recorded (BP, AG, GM, GS, SR).

2.2 Data Analysis

Now, we will present briefly the different stages that allow the prosodic structure to be derived from acoustic data. Above all, correspondence rules between the phonetic and the structural prosodic levels are explained and illustrated with an example. As for pauses, they are used for the moment merely to assure the cohesion among certain ICs. We will therefore demonstrate that they form a structural system among pausal constituents, and that the relationship between pauses and melody is well worth taking into consideration.

2.2a Melodic Cues

Our analysis is not based on perceptual criteria, but acoustic data. The segmentation process that we propose consists of 5 stages (Cf. Figure 2) as explained below:

1. Lexical Filtering: between two different grammatical categories of words, that is functional and lexical ones, the former is not taken into account especially because of its lack of intonative autonomy. As a result, acoustic measurements concern only lexical words, i.e nouns, verbs, sentence adverbs like "sans doute" (not modifying adverbs, for example "beaucoup") and postpositional adjectives.

In agreement with Grosjean [6], we believe that lexical words have an independent status on the prosodic level because they bear most of the time high tones in our data. However, in spite of this relative autonomy, both words must be joined and belong to the same constituent.

2. Tone Labelling and Identification of Pauses: after recording, the speech was orthographically transcribed and divided into words and syllables. The value of F0 on each syllable of lexical words is then extracted manually.

Our two processes of tone labelling and prosodic regrouping are dependent on the 'relative height' notion which P. Mertens [3] used as well. In the same way, each tone label attributed to a syllable relies on the melodic interval with the height of the preceding syllable. As a result, each syllable is a potential point of reference for the syllable to its right, and for that reason we obtain only local informations (for instance, a syllable labelled L can be higher than H farther in the same paragraph).

We proceed therefore from left to right: a P_{n+1} point is labelled **'L'** (Low) when it is lower than the P_n point; otherwise, it is labelled **'H'** (High).

As for the determination of the initial tone height of each paragraph (noted **Pi** before identification), we proceed in the opposite direction.

On the other hand, we distinguish two different pauses: (1) silent pauses (silence and respiration) which seem to have a closing function (**Pc**), (2) non-silent pauses (glottal stop, buccal noises, vocal lengthening and typical French "euh") which are likely on the contrary to open a constituent (**Po**). We suppose that non-silent pauses have a cognitive function of lexical access or of lexical emphasis, except for false starts. As a result, pauses at the beginning of a paragraph are not taken into consideration. When several pauses follows one another, only the first one is kept. Furthermore, we need to extract the durations relative to such labels.

3. Delimitation of the Paragraph: we can determine the right hand boundary of the paragraph (noted L-) when the three conditions below are satisfied:

a. If $L_n < P_i$
b. If L_n is followed by a long (> 50 ms) and a silent pause (Pc category).
c. And if $L_n < L_{n+1}$

Then $L_n = \mathbf{L\text{-}}$

Let us add that we have noticed, at least in our data, that the declination line and the length of the paragraph coincide, which explains condition a.. As for c., it corresponds to the typical resetting at the beginning of a new utterance observed by many linguists.

4. Determination of the Prosodic Structure: in order to segment a paragraph into ICs, the notion of relative height is applied between labelled high tones. Some neighbouring units ending with H can thus be grouped. Therefore we can say that there is a conjunctive relationship between two units when the high tone of the first is lower than that of the second. On the contrary, when the high tone of the first unit is higher, there is a disjunction, in which case the highest

tone is relabelled **FH** (Final Height) and corresponds to the right hand boundary of an IC (noted #). The highest FH is noted **FH+**. We formalize this process as follows:

If $H_n > H_{n+1}$
Then $H_n = FH$

5. Grouping Function of Pauses: we have noticed that pauses can take over from melody by linking two ICs. We distinguish two kinds of pausal strategies assuring this cohesion: (1) the two ICs are separated by a Po, or (2) the second IC is between two pauses, i.e. Pc or Po In all these cases, the intermediate boundary is demoted and relabelled **H+** (Cf. Figure 2). Ex. GMS223: "un bouquet de fleurs[216] # **Po** rouges[211] # avec les tiges vertes[195] # **Pc**" (F0 in Hz).

Above all, this kind of cohesion seems very common inside nominal groups (whether its function is rather local is questionable) and allows morpho-syntactic rules to be dismissed such as: a noun cannot be separated from its post-positional adjective or complement.
In case of strong disjunction between a noun and its postpostional adjective, for instance the insertion of other words, the cohesion between them is restored using linking by both intonation and pauses. In this way, cues can be added to assure a strenghthened cohesive function.
Ex. GMS29: "y'a une petite lampe[186] en revanche **Po** en bas **Po** rose[250] **Pc**", GMS221: "il a un petit noeud papillon[138] **Po** à droite[145] **Po** orange[444] et **Pc**".

2.2b Pauses

Besides this latter partial use, pauses also form a structured system with minor and major contituents (Cf. Figure 1). The first stage of its formation concerns minor PCs all closed by a Pc. In other words, we proceed from left to right and note the boundary of a minor PC each time we encounter a Pc. If a Po is found, then we close the minor PC immediately after the nominal group which it contains. Next, we obtain major PCs by comparing the Pc duration: thus, if Pc1 < Pc2 then PC1 is included in PC2; otherwise, PC2 is included in PC1.

As for the relationship between cues, in our data, we noticed some cases of complementary distribution between FH and pauses; for example, in BPS225 we can observe that, independently of lexical and syntactic considerations quite similar here, two strategies are used by the same speaker to articulate his utterance: the first relies on pauses and the second on FH. In this case, as in our previous stage of grouping motivated by Po, we can note that pauses have a continuative function.

Ex. BPS225: "Sur le dessin[174] de gauche[163] **Pc** la table[178] basse[154] **Pc** en bas à droite[148] **Pc** est blanche[267] **FH+** sur le dessin[157] de droite[200] la table[167] basse[195] en bas[138] à droite[190] est bleue **L-**".

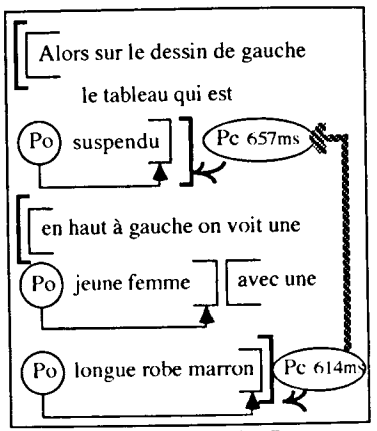


Figure 1. Organization of PCs.

However, FH and pausal effects can also be combined to mark a boundary more obviously; for instance, the major boundary in GMS21 is marked twice (FH+ and Pc). We already exploited in our rules the cohesive effect of this particular boundary on the preceding ICs, and saw that it could also restore a syntactic linking between two separated words .

2.2c Enunciative Interpretation of Prosodic Markers

According to enunciative criteria, the goal of this last part is to decide on a weight of previously determined boundaries. The paragraph, divided into blocks with indetermined statuses of utterance, frame, and rheme, is limited by Pi and L- at its edges. From an enunciative point of view, it is characterized by its thematic

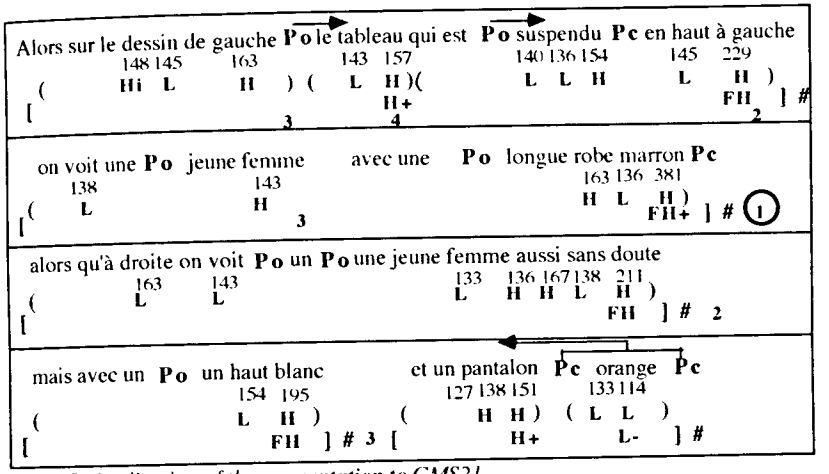


Figure 2. Application of the segmentation to GMS21.

unity due to the attention of the speaker drawn on a same object in pictures. The paragraph also contains at least two utterances, the major frame to the left and its rheme. These two words in turn are separated in the same way; in other words, utterance is recursive.

More precisely, we proceed as follows: (1) the blocks ending with FH+ and L- are considered, (2) starting from these final labels and from right to left, we divide the block into two parts at the highest point, i.e. frame and rheme, (3) the same operation is repeated up to H inside new blocks, (4) after this retroactive analysis, we attribute hierarchical values to the boundaries (1 to the strongest boundary) that correspond to the degree of inclusion of blocks (Cf. Figure 2).

ACKNOWLEDGEMENTS

This paper was supported by EA 1483 "Recherche sur le français contemporain", Paris III-Sorbonne Nouvelle.

3 CONCLUSION

Our findings, that we have recovered from several speakers, have helped to highlight the importance of the relationship between pauses and melody and to confirm that it deserves to be studied in order to improve the discursive segmentation of utterances Now, we need to give an equal weight to pauses and melody by comparing their structures more precisely. Furthermore, we have to perceptually validate the relevance of our boundaries and extend our study to spontaneous speech in different situations. As for the processing of functional words, there still exists an obstacle in this area. Further studies need to be carried out in the future in order to overcome the present problem of automatic implementation of spontaneous speech .

REFERENCES

[1] Morel, M.-A. (dir.) (1995), *Structuration linguistique du dialogue oral*, n° 1; *Langue orale et linguistique du discours*, n° 5, Documents de travail du centre de recherches sur le français contemporain.
[2] Rialland, A., Vaissière, J. (dir.) (1994), "Prosodie du français", *Rapport d'activité de l'U.R.A. 1027*, pp. 16-22.
[3] Mertens, P. (1987), "L'intonation du français. De la description linguistique à la reconnaissance automatique", *Doctorale dissertatie:*, Katholieke Universteit Leuven.
[4] Rossi, M., Di Cristo A. et alii (1981), *L'intonation, de l'Acoustique à la Sémantique*, Paris: Klincksieck.
[5] Péan, V. (1992), "Conception de la base de donnée ICY", *Notes et documents LIMSI*, n° 92-1: Orsay.
[6] Grosjean, F. & Monnin, P. (1993), "Les structures de performance en français: caractérisation et prédiction", *l'Année Psychologique*, fasc. 1, pp. 9-30.

A CONTRASTIVE STUDY OF THE INTONATION PATTERNS OF CHINESE, MALAY AND INDIAN SINGAPORE ENGLISH

Lisa Lim
*Department of Linguistic Science, University of Reading, U.K.**
Department of English Language and Literature, National University of Singapore

ABSTRACT

This study attempts to compare the intonation patterns of Chinese, Malay and Indian Singapore English. Analysis is conducted using the pitch extraction feature of the Kay CSL. Preliminary findings indicate a tendency for the intonation patterns of the three ethnic groups to be distinguished in terms of the alignment of the accent peaks with syllables.

INTRODUCTION

Singapore is a multi-ethnic society whose official languages are English, Mandarin, Malay and Tamil, with English serving as the primary working language. In such a situation, there has emerged a uniquely Singaporean English (SE), which is distinctive and varied, showing influences from the major speech varieties. Numerous studies have considered the diagnostic features of SE. While the segmental characteristics are widely described, however, the suprasegmental aspects are less well documented. Older studies have made observations on the basis of auditory impressions of recorded data -- not any less deserving of merit, of course -- usually making comparisons with British English [eg., 1]. Only in the last few years has research taken a more experimental and instrumental slant [2]. Work has also been done on phonetic features that distinguish the three main ethnic groups, focussing, again, largely on segmental features [3]. Some attitudinal and identification studies hold that it is not possible to distinguish between a Chinese, a Malay and an Indian Singaporean just by listening to them speaking English, particularly with the younger and more educated [4].

It is still felt, however, that Chinese, Malay and Indian varieties of SE may be distinguished [5]. My research attempts to identify distinctive patterns in the intonation of Chinese, Malay and Indian speakers of SE, an area felt not to have been addressed instrumentally or comprehensively enough.

METHOD

Subjects

The subjects consisted of undergraduates from the National University of Singapore, obtained using the "friend of a friend" network technique. Altogether, five each of Chinese, Malay and Indian males, three each of Chinese and Indian females, and four Malay females were recorded.

Material

The data collected included unprompted sentences comprising declaratives, WH- questions, Yes/No questions, and exclamations, sentences prompted by scenarios, a reading passage, a conversation between subject and researcher, and games of 'twenty questions'.

Analysis

Utterances were analysed using the pitch extraction feature of the Kay CSL (Computerised Speech Lab). Pitch synchronous pitch tracking was used by first employing the automatic peak picking capabilities of CSL which marks the division between each voicing impulse in the waveform. This process separates the voiced signal into its periodic components, the inverse of each period being the fundamental frequency (F_0) of the signal. Peak picking was done at a sampling rate of 10 kHz, with analysis range and display of 50 to 250 Hz, and frame size and advance of 20 ms.

The analysis concentrated on the unprompted utterances and free conversation and questioning.

* The author is presently a PhD research student at this institution.

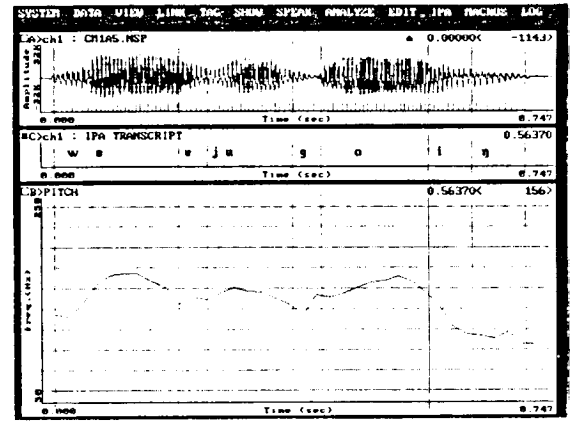


Figure 1. CSL printout of Where are you going uttered by Chinese subject (CM1).

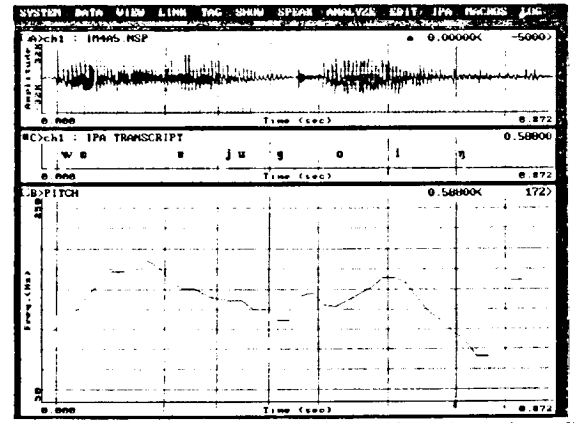


Figure 2. CSL printout of Where are you going uttered by Indian subject (IM4).

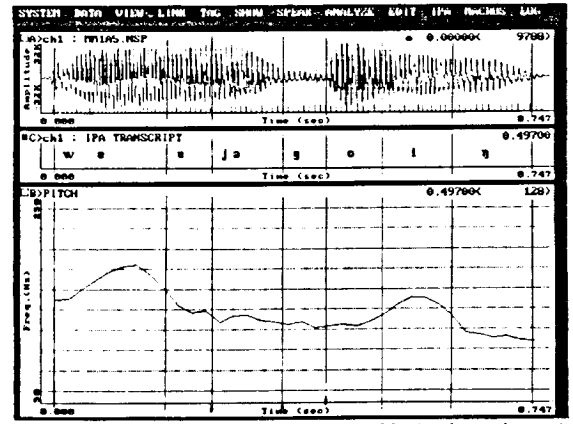


Figure 3. CSL printout of Where are you going uttered by Malay subject (MM1).

RESULTS

Due to the constraints of this paper, the results obtained from one of the utterances which shows interesting patterns, *Where are you going*, will be highlighted. We shall examine in greater detail the pitch movements of one male subject in each of the three ethnic groups, all of which have as utterance contour:

Where are you going
H L !HL L%

Figures 1,2 and 3 are printouts of pitch extraction conducted on the CSL for a Chinese (CM1), Indian (IM4) and Malay (MM1), respectively.

Attention is paid particularly to the final key word in the utterance, as this is where a fair amount of movement occurs. In all three subjects, we note that a similar rise-fall contour spans the final word in the utterance, *going*. The difference between the subjects seems to lie in the alignment of the pitch peak with the syllables [6].

In CM1, the pitch peak is located late within the penultimate syllable; this results in a rise in /go/ and the beginning of a fall, with the fall continuing in /iŋ/. Similarly, in the utterance *Where were you earlier*, while the pitch movement on the final word is reversed, i.e., a fall-rise occurs across *earlier*, the alignment of the pitch peak (trough, in this case) is the same: there is a fall on the penultimate syllable and a rise on the final, with the trough occurring late in the penultimate.

In IM4, the pitch peak is aligned almost exactly at the boundary between the two syllables, resulting in a clear rise in /go/ and a clear fall in /iŋ/. This pattern is also observed in the utterance *Where were you earlier*, where a clear rise-fall is located across the final word, with the pitch peak aligned with the end of the penultimate syllable.

In MM1, the pitch peak is found early in the final syllable of *going*, within the segment /i/. The directional change, as it were, thus occurs in the final syllable /iŋ/. This characteristic pitch peak alignment exhibited by MM1 here is also evident in other utterances. In *Do you like durians*, for example, the rise-fall pattern is found on *durians*, with the pitch peak located early in the final syllable. Similarly in the utterance *Where were you earlier*, the rise in the penultimate syllable continues into the final syllable, with an early-mid peak going into a fall.

Table 1. Ratio of duration of time from start of /o/ to pitch peak, to duration of time from start of /o/ to end of /i/, in the word going.

Subject	Ratio
CM1	0.5136
IM4	0.4807
MM1	0.6897

As can be seen in Table 1, when the duration of time from the start of the vocalic nucleus in /go/ to the pitch peak is calculated as a ratio to the duration of time from the start of the vocalic nucleus in /go/ to the end of that in /iŋ/ , the Malay subject has the largest ratio, followed by the Chinese, and then by the Indian. This corresponds to the visual observation made in the previous paragraphs about the alignment of the pitch peak with the syllables.

Tables 2a to c indicate the size, duration and slope of the pitch movements over the word *going*. As we wish to make comparisons between different speakers with different ranges of voice, the magnitudes of distances are better represented independently of the actual frequency in Hz. A conversion into logarithmic units is the usual solution, and the range of fundamental frequency values in Hz (f_1 and f_2) is converted to a range in semitones (ST), using the formula [7]:

$$\text{SIZE (ST)} = 12.\log_2 \frac{f_1}{f_2} = \frac{12}{\log_{10} 2}.\log_{10} \frac{f_1}{f_2}$$

The rise and fall exhibited over the word *going* are separated into two different phases, at the points where there appears to be a marked change in the value of the slope. As can be seen from the tables, as well as from a visual inspection of the CSL printouts, ignoring the slight perturbations occurring during the plosive, IM4 has clean rise and fall, while the fall in CM1 tapers off to a more gradual slope at the end. In MM1, a somewhat level movement rises to a steeper rise; conversely, his fairly steep fall tapers off, like CM1's, to a very gentle slope.

DISCUSSION

Pitch movement alignment refers to the location of a pitch movement with respect to the syllable boundaries of utterances. The importance of such distinctions to the characterisation of intonational phenomena has long been recognised, particularly with regard to so-called pitch accent languages like Swedish and Serbo-Croatian, as well as languages without an accentual system, like English, German and Dutch [8].

The preliminary results of the present research appear to indicate that pitch accent alignment may also serve to distinguish between ethnic sub-varieties of Singapore English, characterising, particularly, the Malays. This, along with modifications in steepness of contour, would appear to translate to a perceptual sensation of there being more instances of pitch movement within the final word, as uttered by the Malay subject, as compared to, for example, the Indian, which would contribute to the layman's impression that Malays sound more "musical". The contribution of other factors, both segmental and suprasegmental, to this phenomenon must certainly also be borne in mind.

The next logical step in such a study would be the synthesis of utterances, involving variations in the alignment of the accent peak in the pitch contours, as well as variations in steepness of slope, according to the results obtained for the three ethnic groups, and the subsequent obtaining of identification judgements from listeners.

Tables. Pitch movements over the word going, not in alignment with specific syllables.

2a. Of Chinese subject (CM1).

	Rise	Rise	Fall	Fall
size (ST)	-	4.1934	5.8050	2.5751
duration (sec)	-	0.140	0.080	0.100
slope (ST/sec)	-	29.9528	72.5625	25.7513

2b. Of Indian subject (IM4).

	Rise	Rise	Fall	Fall
size (ST)	-	3.1967	10.6453	-
duration (sec)	0.120	0.080	0.140	-
slope (ST/sec)	-	39.9590	76.0377	-

2c. Of Malay subject (MM1).

	Rise	Rise	Fall	Fall
size (ST)	Level	3.5248	4.5308	1.3857
duration (sec)	0.108	0.080	0.060	0.100
slope (ST/sec)	Level	44.0600	75.5126	13.8573

REFERENCES

[1] Tay, M. and A.F. Gupta (1981), *Towards a Description of Standard Singapore English*, RELC 16th Regional Seminar, Singapore.

[2] Deterding, D.H. (1993), *Intonation and Stress Placement in Singapore English*, Paper presented at the SAAL/STU seminar, Singapore.

[3] Sng, K.H. (1986), *Some phonetic properties which may distinguish Tamil, Chinese and Malay speakers of English in Singapore from each other*, Unpublished B.A. Hons. Academic Exercise, National University of Singapore.

[4] Platt, J. and H. Weber (1980), *English in Singapore and Malaysia: Status, Features, Functions*, Kuala Lumpur: Oxford University Press.

[5] Tay, M.W.J. (1982), "The Uses, Users and Features of English in Singapore", in: Pride, J.B. (ed.), *New Englishes*, Rowley, Mass.: Newbury House.

[6] Ladd, D.R. (1994), personal communication.

[7] t'Hart, J., R. Collier and A. Cohen, (1990), *A perceptual study of intonation: An experimental-phonetic approach to speech melody*. Cambridge: Cambridge University Press.

[8] Verhoeven, J. (1994), "The discrimination of pitch movement alignment in Dutch", *Journal of Phonetics*, vol. 22, pp. 65-85.

PITCH PATTERNS AND DURATION: ANALYSIS AND SYNTHESIS

Sandra Madureira, Cairo Humberto da Silva and Patricia A. de Aquino
Pontificia Universidade Católica de São Paulo, S.P., Brasil
Universidade Estadual de Campinas, Campinas, S.P., Brasil

ABSTRACT

The objectives of this paper are twofold: the description and the derivation of prosodic features. Based on the results of an acoustic-phonetic analysis of features of duration and fundamental frequency in neutral declarative sentences and using a simple parser for the generation of pauses, an approach is proposed to derive duration and pitch contours for declarative sentences in a text-to-speech system for Brazilian Portuguese.

INTRODUCTION

The present research stems from two purposes.

One of them "synthesizing reading intonation" pursuits the immediate goal of improving the quality of a concatenative synthesizer by implementing some pitch and duration rules.

The other is to be viewed as part of a more comprehensive project of research on speech signal processing systems in Brazilian Portuguese.

It presupposes a phonological theoretical model directed towards phonetic implementation (Albano, 1993) and requires a thorough description and analysis of phonetic events in Brazilian Portuguese which is under way by a team of researchers.

ANALYSIS

Material, Method and Discussion

This analysis takes into account simple, complex and compound neutral declarative sentences recorded by two male speakers under laboratory conditions.

Speakers were told not to give an emphatic reading so that "neutrality" could be maintained.

This instruction was meant to prevent speakers from using an over-emphatic mode of expression. However, it is not to be taken as reflecting the authors' rejection of an approach in which intonation and grammar are taken to be independent.

On the contrary, Bolinger's thesis that "in intonation there is no distinction between the grammatical and the ideophonic except as they represent extremes of a scale" appears supported by our data.

The sentences of the corpus contain varied number and types of syntactic constituents, syllables and phonemes. Several strategies have been used to build up the sentences: changing the number of word syllables by adding affixes; expanding the heads of phrases with several types of modifiers; replacing subordinate clauses in a complex sentence and changing coordinators in a compound sentence. All sentences were analyzed into their constituents in a top-down hierarchy.

A simple parser was developed to set up prosodic boundaries automatically and introduce silence as well as pitch, loudness and length variations.

Six categories of boundaries have been established. The parser assigns specific markers to each type of boundary, taking into account syntactic constituents and number of syllable diversity. Subject and predicate, for example, are separated depending on the number of syllables constituting the noun phrase subject.

As a result, prosodic domains which are roughly correspondent to syntactic constituents are introduced and this was considered satisfactory in dealing with reading intonation. Conversational discourse intonation would not allow so, since speakers manipulate the melodic and durational components of prosody more freely and introduce pauses in a quite different way.

The pauses introduced by the research subjects in their reading of the sentences occurred mainly between subject and predicate and before shifted syntactic constituents, intensifiers and numerals or at places where there were punctuation marks.

Acoustic measurement of F_0 values, duration of segments and pauses were taken.

For the acoustic analysis of the sentences, the sonograph Kay model 5001 was used. For the extraction of pitch and measurement of the duration of segments, besides the sonograph, a locally implemented software was used.

The acoustic analysis of the data has served as a reference basis for the building up of the duration and pitch models.

Results

The declarative sentences in Portuguese show a global declining pattern. The F_0 at the beginning of the sentences analyzed was about 20Hz higher than at the end of it.

Pitch accent peaks were 20 to 50Hz higher than F_0 utterance onset.

Pitch accented syllables were found to have a strong rising component and longer duration. When a unstressed syllable follows them and precedes an internal boundary, the stepward movement is continued.

Syllables occurring after the antepenultimate stressed syllable in the sentence exhibit falling tones.

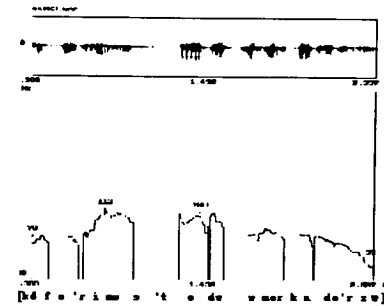


Figure 1. F_0 values: initial, final and at pitch accented syllables in the sentence "Conferimos toda a mercadoria"(We have checked all merchandise).

Pitch rises before non-terminal boundaries between main and subordinate nominal or adverbial clauses

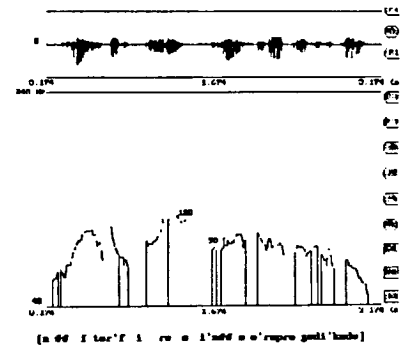


Figure 2. High tone followed by a 220ms pause between the clauses of the complex sentence "Não interfira, senão será prejudicado" (Do not interfere, otherwise you will be at a disadvantage)

Rising was also found to occur before other kinds of non-terminal boundaries such as these between subject and verb.

Rising and falling tones were found at non-terminal boundaries before intercalated syntactic constituents such as prepositional phrases and appositive clauses.

Rising tones usually precede constituents which restrict or introduce a proposition which either complements or denotes a relationship between its

contents and the contents of the constituent from which it is separated by a boundary.

Falling tones usually occur at boundaries between content unrelated propositions and before parenthetical utterances.

Based on the findings, phonetic implementation rules were formulated.

SYNTHESIS

The Duration Model

The duration model is based on the process of modifying the intrinsic duration of each phoneme according to the context in which it occurs in a sentence by means of mutually independent rules.

This modification is worked out through multiplicative factors.

A set of rules was formulated to determine the duration of each phone.

The duration model operates by means of a product of coefficients where each rule represents a factor.

The product of rules applicable to a given phone is calculated for each phone of an utterance to be synthesized.

Maxima and minima values for each phone were established. This constraint was introduced to avoid possible distortions generated by model.

The process of setting up the rules followed basically that proposed by (Allen, Hunnicutt and Klatt, 1987) for the English language.

To cope with Brazilian Portuguese phonetic constraints, alterations and adaptations have been introduced.

Adjustments on the values of the relative coefficients were made based on the perceptual assessment of the synthesized sentences.

The rules are hierarchical: the higher constituent is the sentence, the lowest the phone.

In all 24 rules were formulated.

The application of the durational model to a sentence requires the phonetic transcription of its segments and the specification of their duration.

The pitch model

The pitch model assigns specific pitch contours to the prosodic domains depending on its attributes and its distribution within the sentence.

The rules determine the F_0 contour of each phone at the moment of the synthesis.

The pitch model follows a hierarchical approach. According to this, each level is constrained by the superior level and determines the inferior level in a tree-like structure.

Each prosodic constituent is governed by two linear functions which relate F_0 values to the time domain.

All F_0 contours of a given constituent must be placed between these two linear function graphics.

Within a prosodic constituent, F_0 values between consecutive words are set up.

All F_0 contours of a word must be placed between the graphics of the linear function pairs.

Within a word, the F_0 values are set up between its syllables.

The initial and final F_0 values of each phone were interpolated in a linear manner, that is, the curve is composed by a sequence of line segments, where each segment corresponds to a phone. Perceptually, this limitation has not been felt as causing obtrusive distortion.

As an example of F_0 rule, the following equation can be mentioned:

$$F_0 \text{ is } 100 + 2.5 * n$$

where "n" is the number of syllables of a prosodic constituent corresponding to the subject in a simple sentence.

F_0 is the medium value of the fundamental frequency and it can never be higher than 140Hz.

Another example is the rule which establishes a constant value to the F_0 at the beginning of each sentence.

CONCLUSION

The implementation of prosodic rules in the synthesizer has improved intelligibility and naturalness.

Figures 3 and 4 shows the speech waveform and the F_0 contour for the simple declarative sentence "O custo é pequeno" (The cost is small). Figure (3) refers to natural speech and figure (4) to synthesized speech with prosodic implementation.

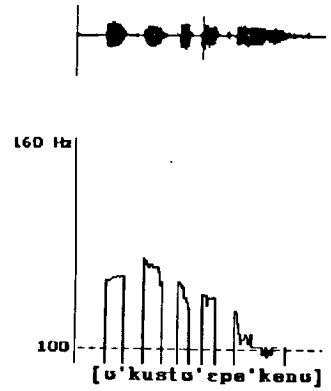


Figure 3. Natural speech

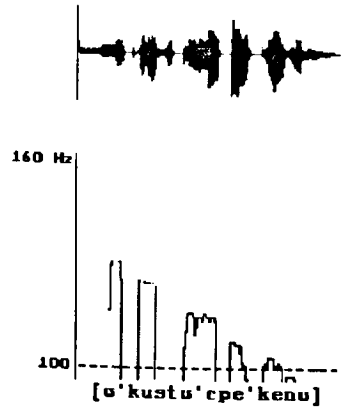


Figure 4. Synthesized speech

ACKNOWLEDGEMENT

This work was supported by CNPq grants n. 50.0400/90-7, 3500 14/93-0 and FAPESP grant 93/0565-2.

REFERENCES

[1] Albano, E. (1993) "Uma fonologia voltada para a implementação fonética", I Encontro sobre Processamento de Língua Portuguesa Escrita e Falada, INESC/ CLUL, Lisboa.

[2] Allen, J., Hunnicutt, S., & Klatt, D. H. (1987) *From text-to speech: The MITalk System*, Cambridge University Press, Cambridge, UK.

[3] Aubergé, V. (1992), "Developing a structured lexicon for synthesis of prosody". In: Bailly, G., Benoit, C., Sawallis, T. R. (eds). *Talking Machines: Theories, Models and Designs*, Elsevier Science Publishers, 39, 274-287.

[4] Bolinger, D. L. (1996) *Intonation and its parts*. Edward Arnold Publishers.

[5] Quené, H. & René, K. (1992) "The derivation of prosody for text-to-speech from prosodic sentence structure". In: *Computer Speech and Language* 6, *77-99.*

PRODUCTION AND PERCEPTION OF STATEMENT, QUESTION AND NON-TERMINAL INTONATION IN GERMAN

Hansjörg Mixdorff and Hiroya Fujisaki

Department of Applied Electronics, Science University of Tokyo

ABSTRACT

The present study adopts a quantitative model originally developed for Japanese to analyze and resynthesize F_0 contours of German utterances. In order to test the descriptive capacity of the model, a production experiment dealing with the realization of statement, question and non-terminal intonation was performed. Based on the analysis results of this experiment, synthetic stimuli for a perception experiment were produced. We examined which of the model parameters are most important for the classification of the sentence mode and found that the accent command offset time T2 largely determines whether an utterance is perceived as statement or unfinished, whereas unfinished and question intonation are distinguished by the accent amplitude.

1. INTRODUCTION

Research on the prosodic features of a language and their relationship with the underlying linguistic and paralinguistic information is important to improve speech analysis and synthesis as well as foreign language teaching [1]. An early experiment [2] by Isačenko and Schädlich showed that in the case of German most syntactic and semantic functions can be realized by manipulating the fundamental frequency (F_0) contour. In the present paper, we will use the term 'intonation' for the prosodic feature expressed by the F_0 contour, being aware of the fact that duration and intensity also play important roles.

2. THE APPROACH ADOPTED IN THE PRESENT STUDY

During recent years much effort has been made to describe the features of German intonation, and to formulate "prototypal" patterns for various sentence types and structures [3].

The present study applies the model by Fujisaki [4] in order to produce a quantitative description of the F_0 contour. The model has originally been developed for Japanese and has since been extended to other languages. It produces an arbitrary F_0 contour by superimposing global (phrase) and local (accent) components. Hence there are two kinds of input signals to the system: impulses (phrase commands) and stepwise functions (accent commands). These are derived by fitting the synthetic F_0 contour to the natural one.

We will try to relate the values derived for these input functions to linguistic units. Figure 1 shows a block diagram of the model.

3. SPEECH MATERIAL AND METHOD OF ANALYSIS

The speech material consists of utterances of the short sentence with declarative word order "Sie haben den Wagen geliehen."— "They rented the car". The sentence was uttered either in statement or question intonation with a narrow focus on one of the constituents 'sie', 'Wagen' or 'geliehen'. We selected an additional sentence where prominence is actually placed on a second clause added ("Sie haben den Wagen geliehen und sind TATSÄCHLICH gefahren" — "They rented the car and ACTUALLY drove away.") to examine the realization of non-terminal intonation. The expression 'non-terminal intonation' needs some explanation.

The final intonation pattern chosen for the first clause of a statement consisting of two clauses connected by 'und' depends on the degree of relatedness between clauses [5]. In the unmarked case where the contents of Clause 2 is based in some way on the contents of Clause 1 the F_0 pattern exhibits a rise followed by a plateau, whereas in the case of two independent clauses only indicating a sequence in time, both clauses may exhibit a fall at the end like in statement intonation. In our experiment, however, we only deal with the unmarked case which we henceforth mean by 'non-terminal intonation'.

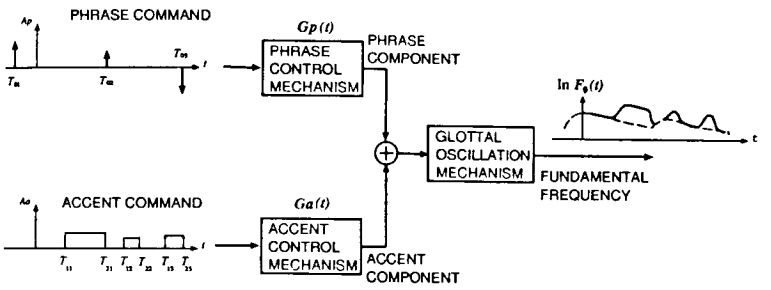


Fig. 1. *Quantitative intonation model.*

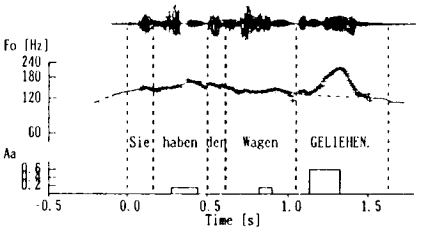


Fig. 2. *Example of analysis, "Sie haben den Wagen GELIEHEN." .*

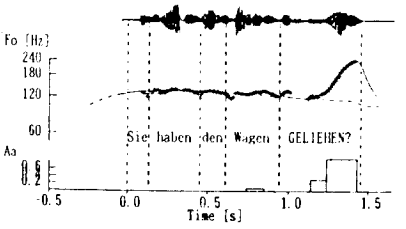


Fig. 3. *Example of analysis, "Sie haben den Wagen GELIEHEN ?".*

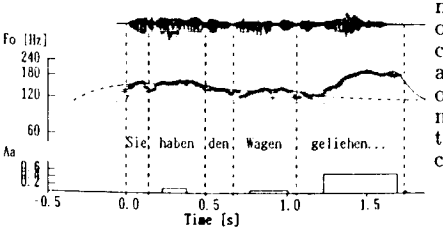


Fig. 4. *Example of analysis, "Sie haben den Wagen geliehen...".*

Twelve speakers from the northern part of Germany read the sentences at a medium speech rate. Words to be focused were hinted at by embedding the sentences in an appropriate discourse context. The utterances were recorded on a DAT and converted at 10kHz (16 bit). After editing the resulting sound files and marking the word boundaries, the F_0 contour was extracted. Errors were corrected by listening and visual inspection. The F_0 contour was then modeled in the Analysis-by-Synthesis approach using a graphic editing tool. In the procedure the initial positions and amplitudes of phrase commands are selected by approximately fitting the phrase component along local minima (the baseline) of the F_0 contour. The accent command on-and offsets are mostly aligned with major transitions of the F_0 contour ("tone-switches") connected to syllables bearing the word accents. The parameter values are then optimized by an iterative procedure for minimizing the mean square error in the ln F_0 domain with α and β set to constant values ($\alpha = 2.0$, $\beta = 20.0$).

4. RESULTS OF ANALYSIS

In this paper we only discuss the results of the production experiment which are relevant to the perception experiment. We have chosen the condition where a narrow focus is placed on "geliehen". Figures 2, 3 and 4 display examples of analysis for statement, question and non-terminal intonation. At the top of all figures, the speech waveform is displayed. The

Table 1. *Mean and standard deviation of accent command timing and amplitude.*
T1: Accent command onset time
T2: Accent command offset time
A_a: Accent command amplitude.

	statem.	question	non-ter.
T1 [ms] μ	88	170/290	220
σ	30	50/50	50
T2 [ms] μ	220	290/480	440
σ	20	50/40	20
A_a μ	0.55	0.45/0.95	0.44
σ	0.16	0.13/0.13	0.06

curve drawn using + symbols indicates the measured F_0 contour, the solid line the synthesized F_0 contour and the dashed line its phrase component part. The accent commands are displayed at the bottom. The statement condition in Figure 2 can be described by a single accent command which is assigned to "geliehen", causing a rise-fall movement of the F_0 contour. In question intonation (Figure 3) typically a lower accent command followed by a high one can be observed. The former has a delayed onset compared with statement intonation. Third, the non-terminal intonation (Figure 4) is characterized by a single accent command with on- and offset timing delayed compared with statement intonation. Table 1 gives mean values and standard deviations of accent command timing and amplitudes for the three conditions. The timing is normalized to the mean word duration for all tokens.

It is easily seen that A_a varies considerably within the group of speakers, though auditory check shows no differences in semantic function between the various realizations. The accent command timing is less subject to individual variation.

5. THE PERCEPTION EXPERIMENT

Taking the results of the production experiment into account we designed an experiment to examine the relationship between the perception of statement, question and non-terminal intonation and the placement and amplitude of accent commands.

A neutral utterance of the sentence

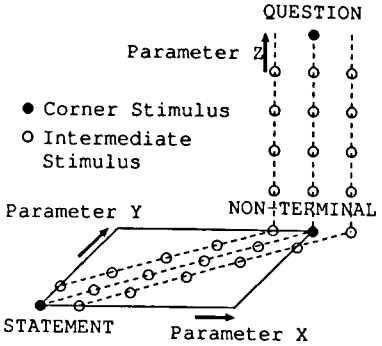


Fig. 5. *Locations of synthetic stimuli in the parameter space.*

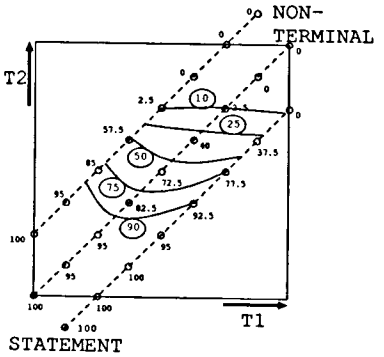


Fig. 6. *Probability of judgment 'statement' in the plane T1 vs. T2.*

"Sie haben den Wagen geliehen." is analyzed using LPC of order 14. The data is then resynthesized replacing the original F_0 contour by one produced with the Fujisaki model. The phrase component of the synthetic F_0 contour is copied from the original F_0 contour and remains unchanged for all stimuli. By using the averaged parameter values (see Table 1) for the three conditions (statement, question and non-terminal intonation) we generate three examplary "corner stimuli". These are locations in the multidimensional parameter space as shown for three parameters X, Y and Z in Figure 5. We create intermediate stimuli at equidistant points on the connecting lines between the corner stimuli and additional stimuli along two lines parallel

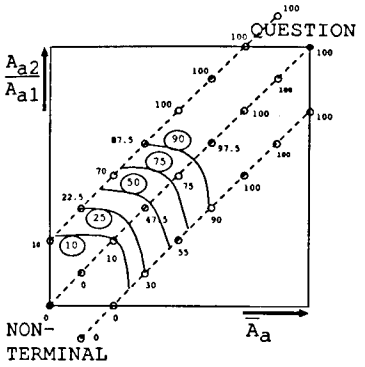


Fig. 7. *Probability of judgment 'question' in the plane $\bar{A}_a$ vs. A_{a2}/A_{a1}.*

to these. Statement and non-terminal intonation mainly differ as to T1 and T2, whereas the difference between non-terminal and question intonation is characterized by the accent command amplitude ratio A_{a2}/A_{a1} and the amplitude mean value $(A_{a1}+A_{a2})/2$.

20 German subjects (14 male, 6 female), two of whom were trained phoneticians, were exposed to each stimulus five times in random order and were asked to decide, if they perceived it as a statement, a question or an unfinished utterance. They could listen to the stimuli as often as they liked.

6. RESULTS

Most subjects consistently identified stimuli grouped around the corner stimuli as belonging to either one of the three cathogories. Figures 6 and 7 show the results of the perception experiment for one pair of corner stimuli each. The locations of the stimuli in the parameter space are marked by dots. In Figure 6, the probability of judgment 'statement' is written to every stimulus, whereas in Figure 7 the probability of judgment 'question' is displayed. By means of maximum likelyhood estimation, lines of equi-probability were determined from the data at 10, 25, 50, 75 and 90 percent levels. The curves suggest that in the case of statement vs. non-terminal the judgment is mainly influenced by the accent command offset time T2. For non-terminal vs. question intonation we find that increasing the accent command amplitude ratio has almost the same effect as increasing the accent command amplitude for both commands.

7. DISCUSSION AND CONCLUSION

As far as the data presented is concerned, the quantitative model has proved its applicability to the analysis and synthesis of F_0 contours of German. Our perception experiment shows that intonation types can be described by averaged parameter sets. The distinction between statement and non-terminal intonation is mainly determined by the accent command offset time T2. This corresponds to the results of a former study by the authors [6]. Question intonation is characterized by a high accent command amplitude and accent command splitting.

Our results encourage using the model for the formulation of a quantitative model of German intonation and its application to speech synthesis.

8. REFERENCES

[1] Fujisaki, H. (1993): From information to intonation. In *Proceedings of the 1993 International Symposium on Spoken Dialogue*, Waseda University, Tokyo., pp. 7-18.

[2] Isačenko, A.V., Schädlich, H.-J. (1966): Untersuchungen über die deutsche Satzintonation. In *Untersuchungen über Akzent und Intonation im Deutschen* (Akademie-Verlag, Berlin), pp.7-67.

[3] Altmann, H., Batliner, A. et al. (1989): *Zur Intonation von Modus und Fokus im Deutschen* (Niemeyer, Tübingen).

[4] Fujisaki, H., Hirose, K. (1984): Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *Journal of the Acoustical Society of Japan (E)*, 5, pp. 233-242.

[5] Pheby, J. (1981): Phonologie: Intonation. In *Grundzüge einer deutschen Grammatik* (Akademie-Verlag, Berlin), p. 890.

[6] Mixdorff, H., Fujisaki, H. (1994): Analysis of Voice Fundamental Frequency Contours of German Utterances Using a Quantitative Model. *Proceedings of the ICSLP '94*, Yokohama, vol. 4, pp. 2231-2234.

BEDOUIN ARABIC INTONATION PATTERNS IN SUPPORT OF THE SUPERPOSITIONAL APPROACH TO INTONATION

Judith Rosenhouse
The Dept. of General Studies, Technion, I.I.T., Haifa, Israel

ABSTRACT

This paper describes an approach to the system of intonation in an Arabic-bedouin dialect group in the north of Israel. The study is based on spontaneous narrative material as recorded "in the field". The material used here represents 2 male and 2 female speakers' intonation patterns. The analysis reveals basic intonation elements, comprising rises, falls and their combinations. The combinations appear in complex utterances (syntactically as well as intonation-wise) and seem to support the superpositional approach to intonation.

Background

One of the languages whose intonation has been little studied is Arabic. In part, this fact is due to the structure of Arabic which comprises a dichotomy into a literary register and many colloquial dialects, or even languages, according to some scholars. This means that the intonation of a certain dialect is not necessarily valid for another or all Arabic dialects, or for the literary variety. The study of Arabic itonation also uffers rom the need for computerized facilities as found in many other languages.

This paper brings the results of a new study of the intonation of a Bedouin Arabic dialect group from the point of view of basic and more complex structures. The data are of a hitherto uninvestigated language variety, and within it of a little-studied text type, i.e. stories.

Material and Method

Our study relies on spontaneous stories narrated by 4 speakers, 2 men and 2 women, from 3 tribes in the Galilee, in the North of Israel (cf. Rosenhouse, 1984). Although these Bedouin groups have been sedentary in the present century, they still preserve their traditional speech habits, including intonation which even now differs from that of sedentary populations. This follows fron the fact that intonation is part of the sociolinguistic differences between Arabic dialects.

The material was uninterrupted narrative monologues with plot, heroes, problems, climax and solution, as in any artistic story, though the contents were not of one and the same "kind".

The speakers' ages were 45, 55 (the women) and about 70 (the men). Each story includes between 300-500 words and supplies reliable dialectal, intonationally rich authentic material, by each speaker.

The work began with pitch analysis (by spectrograph) of the stories which were recorded in the "field". We then continued the study by analyzing various sentence parts in different sizes, structures and intonation patterns using especially written computer programs by I. Rosenhouse. This process has yielded some interesting results (in Hz. and Semitones) part of which are presented here. (More details see Rosenhouse, 1995.)

Results

The detailed inspection of the material yielded basic tonal elements - rises and falls. The roles of the rises and falls are similar to those in other languages, i.e., falls seem to indicate the end of an utterance, while rises indicate continuity. There seem to be 4 pitch levels in this dialect group, from lowest to highest. The range of each level is 4-6 semitones. In an utterance the rise/fall may be for 1,2,3 levels. Falls from the highest to level 3or 2 seem to create an indefinite air, or at least non-finality of the utterance. Thus, a fall not always indicates the end of an utterance in this dialect. This may also depend on the slope of the fall, i.e., on the time element, besides pitch changes.

A difference was found between men's and women's intonation patterns - not in the range (about 10.5-11 STs for both sexes), which rather depends on human physiology, but rather in the modifications indicated by standard deviations of men's vs. women's utterances.

Starting with the smallest relevant units, i.e. syllables, we now focus on intonation units and their combinations.

Colloquial Arabic words consist of syllables of the patterns CV, CVV, CVC, CVVC, CVCC, CCVC, CCVCC, CCVVC. Most of the words are bisyllabic, many are monosyllabic or trisyllabic, and fewer have four or five syllables. Since the peak of each syllable is the vowel at its center, where most of the energy is concentrated, its inherent pitch is usually higher than that of the adjacent consonant(s). Thus, a syllable's pitch has a "natural" rise-peak-fall energy shape. In bisyllabic words this pattern is repeated, usually with more energy and high pitch on the accented syllable. This can be considered the basic intonation structure. Intonation contours span, however, both on single word uterances, as well as longer utterances, up to complex sentences.

The basic elements appear to be repeated in more complex structures, i.e. such that include both rises and falls. These structures can be syntactically or semantically complex, since Arabic has a synthetic and morphologically rich structure, so that

one word (2 or 3 syllables) may be a complete (S:V:O) sentence.

But semantics does not stop at the syntactic level. Thus, a word (e.g. a noun) with a complex syllabic structure may get internal intonation curves according to the speaker's intent, in addition to the word-stress, placed according to definite rules.

In utterances longer than a word, the intonation contour may appear both more detailed and clearer, similar to a picture whose features get clearer the larger it is. In such cases intonation contours can be analyzed into separate rises or falls that do not form part of one and the same linguistic sub-element.

The system in our Bedouin samples seems thus to be nearer to a hierarchical superposition of intonation patterns than to a "linear" description: The complex patterns appear in these texts both in short (and simple) and long (and complex) linguistic units (i.e., phrases, clauses and sentences.

This approach is similar to the one presented already 20 years ago by R. Nash (1973) for Turkish intonation (cf. Fig. 1). Our examples demonstrate this approach from the bedouin Arabic texts we have analyzed (see below).

Discussion and Conclusions

Arabic intonation is complicated to study due to basic inherent linguistic features of the language: it includes numerous dialects which make the study of a certain dialect sometimes irrelevant for any other dialect; the difference between the colloquial and literary registers make it impossible to predict the intonation of a certain dialect (sedentary or bedouin). Moreover, Arabic has a rich synthetic morphological structure, so that though most words include 2 syllables, many others have more than 2 syllables due to various afffixed elements. Words such as nouns or verbs that carry word-stress in an utterance usually have basic intonation structures, i.e., rises or falls. But they may also have combinations of these elements in, e.g., rise-fall patterns. Longer words, with more syllables, as well as longer syntactic utterances, combine these elements once per word/utterance or more times per word/utterance. Thus, such longer-than-monosyllabic speech-units support the theory of superposition of intonation. The Figures below seem to corroborate this viewpoint.

References

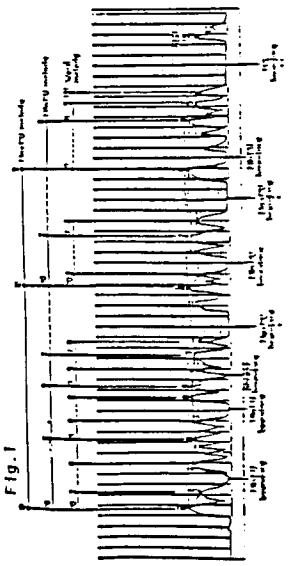
Nash, R. (1973) Turkish Intonation, The Hague: Mouton

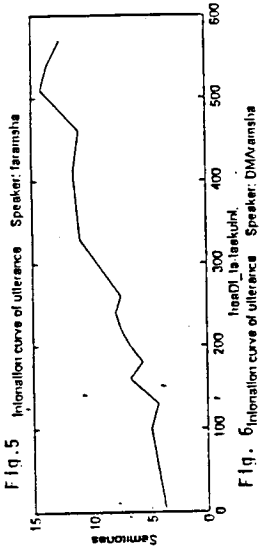
Rosenhouse, J. (1984) The Bedouin Arabic Dialects, General Problems and a Close Analysis of North Israel Bedouin Dialects, Wiesbaden: Harrassowitz.

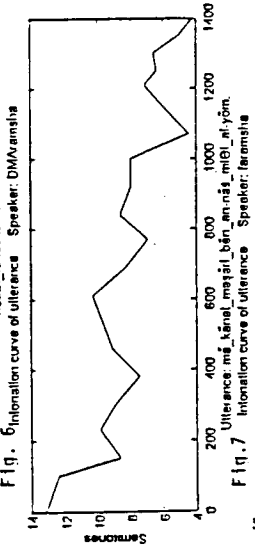
Rosenhouse, J. (1995) "Features of intonation in Bedouin Arabic narratives of the Galilee (North Israel)" in: Festschrift for H. Palva for his 60's birthday, Helsinki.

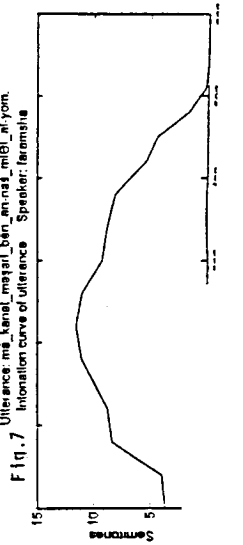
Fig. 1 - Fig. 9 in Nash, 1973:87

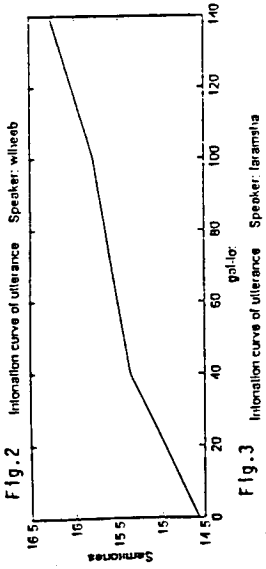
Fig. 2 - "he said to him"
Fig. 3 - "you bring"
Fig. 4 - "I slept"
Fig. 5 - "This one will eat me"
Fig. 6 - "There was no money among people as today"
Fig. 7.- "that is, there is nothing between them"

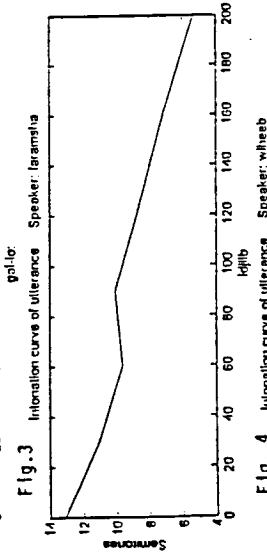


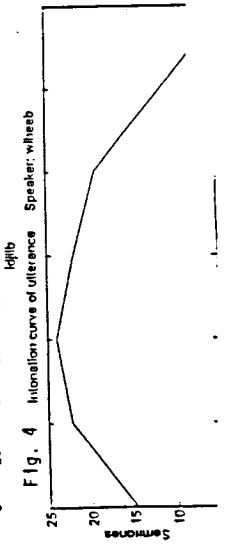












SPATIOTEMPORAL STABILITY OF THE TONGUE-JAW AND LIP-JAW COMPLEX: COMPARISONS ACROSS SESSIONS

Peter J. Alfonso
The University of Illinois at Urbana-Champaign, USA

ABSTRACT

The search for invariant motor control schemes most often is based upon articulatory data collected from a single session and within a single speech rate. Thus, we know little about the stability of gestural organization and coordination across relatively long periods of time. Articulatory data were collected over multiple sessions and rates. Only motor equivalence covariability, compared to a number of spatial and temporal measures, was stable across sessions for all subjects.

INTRODUCTION

Certain observed characteristics of the motor system, such as motor equivalence covariability, have lead to the idea that there is an underlying invariance in the motor control scheme despite the observed surface variations in performance. In particular, the notion of coordinative structures in dynamical models of speech production, and in dynamical models of movement in general, is based on the idea of underlying invariant motor control schemes to meet a specified task [1,2]. However, and in spite of the decades-long search for the so-called invariant characteristics of normal speech production, the majority of the often cited motor characteristics of speech [e.g., 3] are much more variable in repeated-trial tasks than they are invariant.

Most often the search for invariant control schemes focuses on data collected from a single session and within a single speech rate condition. The overall purpose of these ongoing experiments is to explore further the saliency of certain presumed invariant motor characteristics of speech, and by extension, the notion of coordinative structures, by comparing certain spatiotemporal characteristics of tongue-jaw and lip-jaw movements for stops and fricatives across multiple sessions and across speech rate.

For the purposes of the experiments reported here, motor coordination and organization are not synonymous. First, the primary criterion that a speech motor characteristic must meet in order to reflect a well-coordinated speech motor gesture is in its relative stability. Second, the organization of a speech motor gesture refers to the contribution of the components of an articulatory synergy toward the specified task, and does not in itself invoke assumptions about coordination.

METHODS

The movements of the tongue blade, lips, and jaw were transduced by electromagnetic midsagittal articulography (EMMA). A single session included twenty perceptually fluent repetitions of the target words /pap/, /tat/, and /sas/ imbedded in the Dutch carrier phrase "Zij zei CVC alveer." The 60 phrases were blocked by rate and produced first at a normal speech rate, than again at a fast rate, and finally at a slow rate. Sessions were repeated three times, and the interval between sessions was about two weeks. Thus, approximately 540 utterances (3 words X 20 repetitions X 3 rates X 3 sessions) per subject were collected. Seven native Dutch talkers completed either 2 or 3 sessions. Only data associated with closure movements for syllable initial /p/ and /t/ during the normal rate condition are discussed here. Considerable software development, calibration procedures, and hardware modifications were made to the Carstens EMMA system used here [4,5].

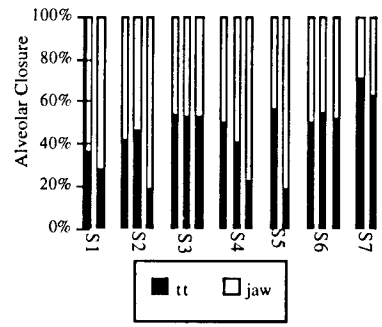
RESULTS

1. Spatial Organization and Coordination.

General trends in the organizational stability for /t/ closure are demonstrated in Figure 1, which shows the normalized vertical displacement for /t/ closure across two or three sessions for seven control subjects. While organizational patterns differ across subjects, for example, Subject 1 achieves /t/ closure primarily by jaw displacement while Subject 7 achieves closure primarily by tongue

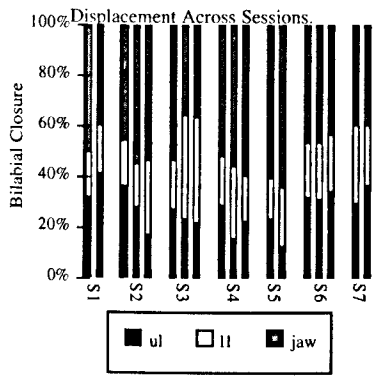
displacement, the relative displacement patterns within subjects are generally stable across sessions for subjects 1, 3, 6, and 7. In the remaining three subjects, only Subject 5 shows a reversal in the primary articulator for closure. That is, /t/ closure is achieved primarily by tongue displacement in session 1 but primarily by jaw displacement in session 2. Subjects 2 and 4 achieve closure primarily by tongue displacement, although the relative contribution of the tongue and the jaw for closure varies considerably across sessions.

Figure 1. /t/ Closure. Normalized Vertical Displacement Across Sessions.



General trends in the organizational stability for /p/ closure are demonstrated in Figure 2.

Figure 2. /p/ Closure. Normalized Vertical Displacement Across Sessions.



As in the case of /t/ closure, organizational patterns differ across subjects. For example, Subject 5 achieves /p/ closure primarily by jaw displacement and secondarily by lower lip displacement with little contribution of the upper lip. On the other hand, Subject 6 achieves closure primarily by jaw displacement, secondarily by upper lip displacement, with little lower lip displacement. However, the relative displacement patterns within subjects are generally stable across sessions for 4 of the 7 subjects.

Details of the closure gestures for Subject 2, the only subject who demonstrated unstable organizational patterns for both /t/ and /p/ closure, are shown in Figures 3 and 4.

Figure 3. /p/ Closure. Subject 2. Vertical Displacement Across Sessions.

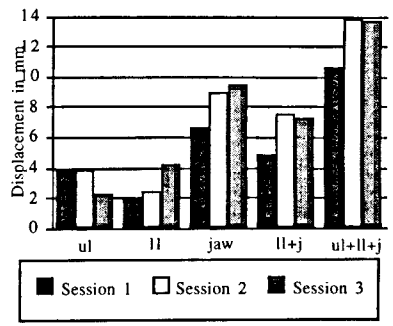
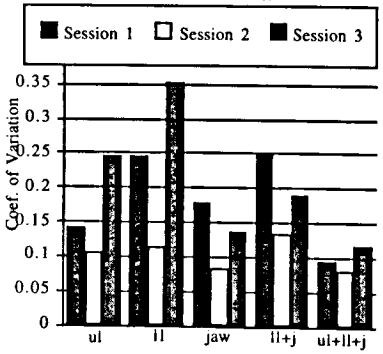


Figure 3 shows the average absolute vertical displacements of the upper lip (ul), lower lip (ll), and jaw for /p/ closure across three sessions. Note that this subject presents different closure strategies across sessions. For example, upper lip displacement is twice as great as lower lip displacement in sessions 1 and 2 but half as great in session 3. Also note that differences in average total displacement across the three sessions is about 3.5 mm. The peak velocity profiles generally correspond to the displacement profiles.

Figure 4 shows that although the control strategies vary across sessions, the lips and jaw demonstrate motor equivalence covariability and thus are well coordinated across sessions for /p/ closure. The figure shows that the variability of the labial gesture, that is, the combined ul+ll+j displacement, is less than the variabilities associated with either the upper lip, lower lip, or jaw sig-

nals for each of the three sessions. Thus, Subject 2, who is the least stable of the seven subjects for both /t/ and /p/ closure in regard to both the relative organizational patterns shown in Figures 1 and 2 and in the absolute displacements shown in Figure 3, demonstrates consistent motor equivalence covariability. That is, the gestural organization varies but gestural coordination remains stable, a result that is observed in all of the subjects regardless of the idiosyncratic organizational patterns that they exhibit.

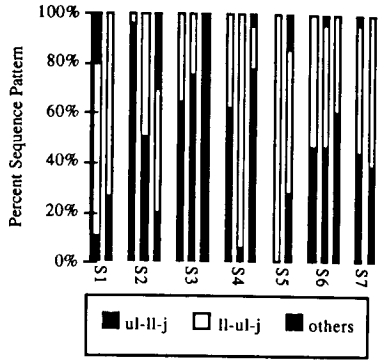
Figure 4. /p/ Closure. Subject 2. Coefficient of Variation Across Sessions.



2. Temporal Organization and Coordination.

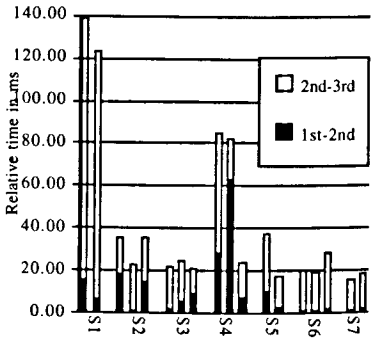
Figure 5 shows the within and across session relative distribution of the upper lip, lower lip, and jaw sequence patterns for /p/ closure for seven subjects. The closed portions of the bars represent the upper lip, lower lip, and jaw sequence, the open portions represent the lower lip, upper lip, and jaw sequence, and the hatched portions represent all others. By far, the predominant sequences are those in which lip movement occurs first and jaw movement occurs last. First, the figure shows that either lip lead sequence is equally likely to occur. For example, Subject 1 prefers the lower lip lead sequence while Subject 3 prefers the upper lip lead sequence. Second, the figure shows that some subjects, for example Subject 6, show no clear preference for either lip lead sequence. Third, the figure shows that two of the subjects, 2 and 4, show a clear reversal in the lip lead sequence across sessions.

Figure 5. /p/ Closure Across Sessions and Subjects. Relative Distribution of Sequences re Peak Velocity.



The temporal ordering of the movements of the speech articulators was thought to represent a motor invariant in the temporal domain in that an invariant upper lip, lower lip, and jaw sequence for bilabial closure was reported for a large group of control subjects [6]. However, recent research suggests that the initial conclusion in regard to invariance may have been over generalized [7]. The data reported here support the recent findings and, further, demonstrate that temporal ordering is quite unstable across sessions for some subjects.

Figure 6. /p/ Closure. See Text.



One of the reasons that the sequence pattern is not stable across time is that it does not reflect interarticulator relative time. Subjects who demonstrate tight coupling between lip movements, for example, would have a higher probability of producing both lip-lead sequences compared to subjects who demonstrate longer relative timing of the lip movements. This is demonstrated in Figure 6, which shows the lip relative time and the lagging lip to jaw relative time for the average sequences per session. (The relative time profiles that correspond to each of the sequences are similar to each other, and they are similar to the relative time profiles for the session average sequences shown in Figure 6.) Note that the relative time for lip movements in the case of Subjects 6 and 7 is less than five ms. Figure 5 shows that these subjects demonstrate nearly equal probability of producing either lip-lead sequences. Thus, an invariancy criterion based on consistent sequence patterns would exclude subjects 6 and 7 whereas a criterion based on tight interarticulator timing would include the same subjects.

DISCUSSION

The data shown in Figures 5 and 6 demonstrate that for some subjects neither temporal ordering nor relative time are stable across sessions and thus do not represent invariant speech motor characteristics. Further, it appears that subjects who are relatively unstable across sessions in regard to these two temporal measures are also relatively unstable in regard to displacement characteristics. For example, Subject 2 achieves closure with varying spatial strategies (Figures 1-4) and with varying temporal strategies (Figures 5-6) across sessions. Only motor equivalence covariability is stable across sessions for all subjects, that is, even in the case of unstable spatial and temporal organizational characteristics such as demonstrated by Subject 2.

A conclusion that could be drawn from a stable coordination index (motor equivalence covariability) regardless of the stability of the displacement pattern, temporal order, or interarticulator relative timing is that the spatial and temporal organization of functionally linked articulators is secondary to gestural specification, that is, bilabial closure in this example. This is precisely what would be predicted from a task dynamic point of view; that the spatial and temporal organizational characteristics of the individual articulators that comprise an articulatory complex represent the natural consequence of gestural coordination and therefore would not demonstrate stability across sessions.

REFERENCES

[1] Saltzman, E. (1991). "The task dynamic model in speech production." In H.F.M. Peters, W. Hulstijn, and C.W. starkweather (Eds). *Speech motor control and stuttering* (pp. 37-52). Amsterdam: Excerpta Medica.

[2] Saltzman, E., & Munhall, K. (1989). "A dynamical approach to gestural patterning in speech production. " *Ecological Psychology*, 1, 333-382.

[3] Gracco, V.L. (1994). "Some organizational characteristics of speech movement control." *Journal of Speech and Hearing Research*, 37, 4-27.

[4] Alfonso, P.J., Neely, J.R., van Lieshout, P.H., Hulstijn, W., & Peters, H.F. (1993). "Calibration, Validation, and Hardware-Software Modifications to the Carstens EMMA System.." *Proceedings of the ACCOR Workshop on Electromagnetic Articulography in Phonetic Research.*, Munich, Germany. Also in *Haskins Laboratories Status Report of Speech Research, SR-114*, 101-112.

[5] Van Lieshout, P.H.H.M., Alfonso, P.J, Hulstijn,W., & Peters, H.F.M. (1994). "Electromagnetic Midsagittal Articulography (EMMA)." In F. J. Maarse, A.E. Akkerman, A.N. Brand, L.J.M. Mulder, and M.J. van der Stelt (Eds.), *Computers in Psychology 5: Applications, Methods, and Instrumentation.* (pp. 62-76). Lisse, The Netherlands: Swets & Zeitlinger.

[6] Gracco, V.L. and Abbs, J.H. (1986). "Variant and invariant characteristics of speech movements. "*Experimental Brain Research, 65*, 156-166.

[7] Van Lieshout, P.H.H.M., Alfonso, P.J., Hulstijn, W., & Peters, H.F.M. (1993). "Significance of relative-timing towards an interpretation of articulatory sequencing." *ASHA, 35-10.*, 191.

COMPARING METHODS FOR QUANTIFYING THE VOICE SOURCE OF DIFFERENT PHONATION TYPES INVERSE FILTERED FROM ACOUSTIC SPEECH SIGNALS

Paavo Alku[1] and Erkki Vilkman[2]
1: Helsinki University of Technology, Acoustics Lab.
Otakaari 5 A, FIN-02150 Espoo, Finland
2: Helsinki Univ. Central Hospital, Dept. Phoniatrics
Haartmaninkatu 4, FIN-00290 Helsinki, Finland

ABSTRACT

This study compares quantification techniques that have been developed to parameterize the voice source obtained by inverse filtering. Quantification of the voice source of different phonation types was computed using altogether seven parameters. The results showed that phonation types could be separated from each other most effectively when quantification was based on parameters determined between the instant of the maximal glottal opening and the minimum peak of the flow derivative.

INTRODUCTION

Inverse filtering is widely applied in the analysis of voice production. Inverse filtering methods can be divided into two categories. The first category consists of techniques that are based on inverse filtering of the volume velocity signal that has been recorded at the mouth using a flow mask [1]. The resulting glottal volume velocity waveform can be calibrated in the amplitude domain and the DC-flow is also obtained. The second category of inverse filtering techniques is based on the estimation of the glottal source from the acoustic speech pressure wave that has been recorded in a free field (e.g.[2]). Glottal airflow waveforms estimated by these techniques are obtained on arbitrary amplitude scales with no indication of the DC-flow.

Glottal flows estimated by inverse filtering are usually quantified using certain parameters. Characterization of the voice source by time-based parameters that are extracted from the glottal volume velocity waveform has been widely used [3]. Time-domain quantification of the voice production using the derivative of the glottal airflow waveform has also been used [3, 4]. If flow mask is used in inverse filtering parametrization of the voice source can be done by measuring the absolute values of both the AC- and DC-flow. In the frequency domain the decay of the voice source spectrum can be parametrized using, for example, the harmonic richness factor (HRF) [5].

The aim of this research was to compare different methods that have been developed for quantification of the glottal airflow that has been inverse filtered without a flow mask. We were interested in exploring how changing the phonation type can be presented by different quantification techniques. By doing this comparison our purpose was to find the parameter that most clearly indicates changes in the phonation type.

MATERIAL AND METHODS

Time-domain quantification of the glottal airflow waveform was computed by using the following parameters [3]: open quotient (OQ), speed quotient (SQ), and closing quotient (CQ). Quantification of voice production using the derivative of the flow waveform was performed with the following time-domain parameters: return quotient (RQ) [6] and peak-to-peak quotient (PPQ) [4]. By referring to Fig. 1 these time-domain parameters can be defined as follows:

$OQ = (t_{o1} + t_{o2}) / T$
$SQ = t_{o1} / t_{o2}$, $CQ = t_{o2} / T$
$RQ = t_{ret} / T$, $PPQ = t_{pp} / T$

Amplitude domain quantification of the glottal source was not possible with absolute flow values because our approach was based on inverse filtering without a flow mask. However, the authors have recently presented a new amplitude-based quotient which can be used even though absolute flow values are not given be the recording apparatus [7]. This new parameter, amplitude quotient (AQ), is defined as the ratio of the AC-amplitude of the flow and the amplitude of the negative peak of the first derivative of the flow (Fig. 1):

$$AQ = A_{ac} / A_{min}$$

Frequency domain quantification of the voice source was computed using the harmonic richness factor [5]:

$$HRF = \frac{\sum_{i \geq 2} H_i}{H_1}$$

, where H_i denotes the amplitude of the ith harmonic computed from the spectrum of the glottal volume velocity waveform.

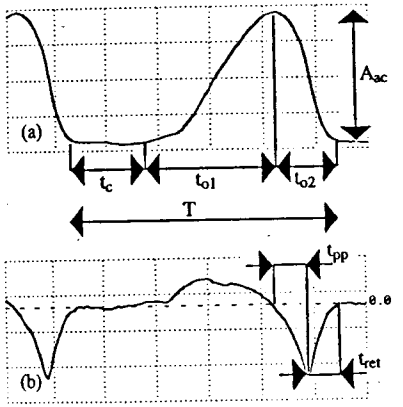


Fig. 1 (a): Glottal flow, t_c = closed phase, t_{o1} = opening phase, t_{o2} = closing phase, A_{ac} = amplitude of AC-flow, T = fundamental period
(b): Derivative of the glottal flow, A_{min} = amplitude of the negative peak, t_{ret} = return time, t_{pp} = time distance between the negative peak of the differentiated flow and the positive peak of the flow

The speech material consisted of vowels produced by five female and five male speakers. The speakers were asked to produce a sustained /a/-vowel using breathy, normal, and pressed phonation types. Recording of the signals was performed in an anechoic chamber using a condenser microphone (Brüel&Kjær 4133).

Estimation of the glottal airflow waveforms was performed with a new inverse filtering technique that is described in detail in [2]. All the estimated glottal airflow waveforms and their derivatives were analyzed using a computer cursor in order to mark time and amplitude values that were required for computation of parameters.

RESULTS

The obtained values for all the seven parameters are given in Tables 1 and 2 for female and male voices, respectively. Two relative changes were computed for each parameter by comparing normal phonation to breathy and pressed phonation to normal. In the case of OQ, for example, these relative changes were defined as follows:

$$\frac{OQ_{normal} - OQ_{breathy}}{OQ_{breathy}} 100\ \%$$

$$\frac{OQ_{pressed} - OQ_{normal}}{OQ_{normal}} 100\ \%$$

To compare deviation of the parameters between different speakers we expressed results using coefficient of variation (i.e. the ratio between the standard deviation and the mean). In the case of OQ, for example, coefficient of variation was defined as follows:

$$v = (sd_{OQ} / m_{OQ})\ 100\ \%$$

Female speakers:

Time-based parameters computed from glottal flows of female voices showed that the mean value of both OQ and CQ decreased while the mean of SQ increased when phonation was changed from breathy to pressed. Changing of the mean values of all these time-based parameters was monotonic when phonation was altered. The relative change was smallest for OQ. The mean value of SQ showed the largest relative change (38 %) among the three time-based quotients when breathy phonation was compared to normal. However, when phonation was further changed from normal to pressed the mean value of SQ increased only slightly (5 %). The mean value of CQ showed a clear descend when phonation was changed both from breathy to normal (-28 %) and from normal to pressed (-10 %). When these three time-based parameters were analyzed between different speakers it turned out that CQ was the only one whose value changed (decreased) monotonically for all the female subjects when phonation was changed.

RQ was the only parameter among all the analyzed quotients whose mean value did not change monotonically when phonation was altered from breathy to pressed. There were large variations between different speakers when the value of RQ was analyzed as a function

of phonation. PPQ decreased monotonically for all the female subjects when phonation was changed from breathy to pressed.

AQ decreased monotonically for all the female subjects when phonation was changed. The mean value of AQ showed large relative changes (-33 % and -20 %) when different phonation types were compared. The frequency domain parameter, HRF, showed the largest relative changes when phonation was altered (135 % and 57 %). HRF-values increased for all the subjects monotonically.

Among the three time-based parameters extracted from the glottal flows the value of OQ showed the smallest deviation from one speaker to another. The value of v averaged over the three phonation types (v_{av}) equaled 8 % for OQ, 15 % for CQ, and 16 % for SQ. Deviation of parameter values between female speakers was largest for RQ (v_{av} equaled 26 %). For PPQ deviation was smaller (v_{av} equaled 20 %). The coefficient of variation for AQ gave a value that was the second smallest among all the analyzed parameters of female subjects (v_{av} equaled 11 %). The value of HRF showed large deviations from one female voice to another (v_{av} equaled 26 %).

Male speakers:

Mean values of time-based parameters extracted from the flow waveforms of male subjects changed monotonically (OQ and CQ decreased, and SQ increased) when phonation was altered from breathy to pressed. Relative changes between different phonation types were larger than in female voices. The relative change in the value of OQ was again smallest. SQ yielded the largest relative change for time-based parameters of male voices (87 % when phonation was changed from breathy to normal). However, the value of SQ increased only slightly (by 1 %) when phonation was further changed from normal to pressed. The value of CQ showed large changes both in breathy-to-normal (- 40 %) and normal-to-pressed (-19 %) changes. CQ was the only time-based parameter that showed for all the male subjects a monotonic decrease when phonation was changed from breathy to pressed.

RQ showed also for male voices the largest deviation from one voice to another. The mean value of RQ did not change monotonically when phonation was altered. The mean value of PPQ decreased by 38 % and 35 % in breathy-to-normal and normal-to-pressed changes, respectively. For all the five subjects PPQ decreased monotonically when phonation was altered towards pressed.

The mean value of AQ decreased monotonically when the phonation type was changed. All the male subjects yielded the largest value of AQ in breathy phonation, the second largest in normal phonation and the smallest value in pressed phonation. The mean values of HRF increased also monotonically when phonation was changed towards pressed.

The time-based quotients extracted from the glottal flows showed deviations between subjects that were smallest in CQ (v_{av} equaled 11 %), second smallest in OQ (v_{av} equaled 13 %) and largest in SQ (v_{av} equaled 18 %). The value of RQ showed for male voices the most substantial deviation from one speaker to another (v_{av} was equal to 33 %). PPQ showed also quite large deviation (v_{av} equaled 21 %). Variation of the value of AQ from one speaker to another was larger in male voices than in female speech (v_{av} equaled 20 %). HRF showed also for male voices great deviation between different subjects (v_{av} equaled 27 %).

SUMMARY

Comparison of the parameters was done, first, by using as a criterion the change of the mean parameter value when phonation was altered from breathy to pressed. With this criterion the parameters could be sorted according to the following order of superiority both for female and male voices: HRF, AQ, PPQ, CQ, SQ, OQ, and RQ. Second, we analyzed for how many speakers the changing of parameters was monotonic when phonation was altered. It was found that AQ was the only parameter whose value showed a clear monotonic change (decrease) for all the subjects when phonation was altered from breathy to pressed. Hence, the parameters could be sorted using the following order for female voices: HRF, AQ, PPQ, CQ, SQ, OQ, and RQ. For male voices the order of superiority was as follows: AQ, PPQ, CQ, HRF, SQ, OQ, and RQ. Third, quantification methods were compared by using as a criterion deviation of the parameter values between different speakers. Using this criterion the following order of superiority was obtained for female voices: OQ, AQ, CQ, SQ, PPQ, HRF, and RQ. For male voices the order was slightly different: CQ, OQ, SQ, AQ, PPQ, HRF, and RQ.

We conclude that the phonation type can be characterized most effectively by using either frequency domain parameterization with HRF or time-domain parameterization that is based on values extracted during glottal closing phase, especially during the time that spans from the instant of maximal glottal opening to the instant of the negative peak of the flow derivative. If extraction is based on the flow waveform alone, then according to our experiments the best time-domain parameter is CQ. However, if time-domain parameterization is based on the flow derivative alone, then applying PPQ is recommended. According to the results of our experiments the most effective way to characterize the voice source in the time-domain is to apply both the flow and its derivative by using parameter AQ.

REFERENCES

[1] Rothenberg, M. (1973). "A new inverse-filtering technique for deriving the glottal air flow waveform during voicing", J. Acoust. Soc. Am., Vol. 53, pp. 1632-1645.

[2] Alku, P., Vilkman, E. (1994) "Estimation of the glottal pulseform based on discrete all-pole modeling", Proc. '94 Int. Conf. on Spoken Language Processing, pp. 1619-1622.

[3] Holmberg, E.B., Hillman, R.E., Perkell, J.S. (1988). "Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice", J. Acoust. Soc. Am., Vol. 84, pp. 511-529.

[4] Sundberg, J., Titze, I., Scherer, R. (1993). "Phonatory control in male singing: A study of the effects of subglottal pressure, fundamental frequency, and mode of phonation on the voice source", J. Voice, Vol. 7, pp. 15-29.

[5] Childers, D.G., Lee, C.K. (1991). "Vocal quality factors: Analysis, synthesis, and perception", J. Acoust. Soc. Am, Vol. 90, pp. 2394-2410.

[6] Price, P.J. (1989). "Male and female voice source characteristics: Inverse filtering results",Speech Communication, Vol. 8, pp. 261-277.

[7] Alku, P., Vilkman, E. (1994). "Amplitude domain quotient for characterization of the inverse filtered glottal flow", In review.

Table 1. Means (m) and standard deviations (sd) for parameters of female subjects.

phonation		OQ	SQ	CQ	RQ	PPQ	AQ	HRF
breathy	m	0.94	1.38	0.40	0.18	0.22	9.05	0.20
	sd	0.07	0.24	0.06	0.03	0.05	1.55	0.08
normal	m	0.84	1.90	0.29	0.14	0.16	6.07	0.47
	sd	0.04	0.34	0.04	0.06	0.03	0.37	0.09
pressed	m	0.78	1.99	0.26	0.16	0.12	4.85	0.74
	sd	0.10	0.28	0.04	0.03	0.02	0.46	0.13

Table 2. Means (m) and standard deviations (sd) for parameters of male subjects.

phonation		OQ	SQ	CQ	RQ	PPQ	AQ	HRF
breathy	m	0.96	1.15	0.45	0.14	0.32	22.16	0.20
	sd	0.04	0.16	0.05	0.07	0.06	4.33	0.04
normal	m	0.84	2.15	0.27	0.09	0.20	10.39	0.56
	sd	0.08	0.45	0.02	0.03	0.04	2.65	0.11
pressed	m	0.70	2.18	0.22	0.12	0.13	7.08	0.91
	sd	0.17	0.44	0.03	0.02	0.03	0.94	0.36

SPEECH MAPS INTERACTIVE PLANT "SMIP"

L.J. Boë[1], B. Gabioud[2] and P.Perrier[1]
[1]ICP URA CNRS n° 368 INPG/ENSERG Université Stendhal, France
[2]Institut d'Informatique, Lausanne, Suisse

ABSTRACT

The SMIP is an interactive and ergonomic software. On the basis of an articulatory model, this device delivers an output signal which is associated with geometric and acoustic informations. This articulatory model regenerates lip and vocal tract shapes, with seven articulatory parameters as input. By means of a set of coefficients, the midsagittal contour is converted into an area function with which the transfer function and/or formants of the vocal tract can be calculated. Reasonable quality sound is generated. In addition, the 37 vowel prototypes of the UPSID database are provided, and tools to compute macro-variations are implemented.

INTRODUCTION

In the frame of studies on the articulatory-acoustic relationship, it is of high interest to manipulate articulatory models which integrate morphological and articulatory constraints. Indeed, such anthropomorphic models offer the possibility to coherently vary area functions by modifying appropriate control parameters. It thus becomes possible to relate these geometric variations and associated formant changes to real behaviour in speech production systems. Such a model can be extensively used for designing vowel (Vallée, 1994) and syllable prototypes, for the prediction of speech sound systems, for articulatory-acoustic inversion, and for speech synthesis. The present paper presents briefly the different components of the SMIP, and describes in detail the various operations that can be performed with this software. The SMIP, i.e. the *Speech Maps Interactive Plant*, has been elaborated in the frame of the European ESPRIT/BR project Nr. 6975 *Speech Maps*, and constitutes the core of the *Articulotron* [1].

THE COMPONENTS

The SMIP is organised in four main modules: (1) vocal tract midsagital contours are delivered by the articulatory model from seven control parameters: lip height, lip protrusion, vertical position of the jaw, front-back position of the tongue body, vertical position of the tongue dorsum, position of the apex, and vertical position of the larynx; (2) from the midsagittal dimensions, the vocal tract area function is estimated with a set of coefficients derived from radiographic measurements; (3) vocal tract acoustic transfer functions and formants and bandwidths are calculated by means of an acoustic model; (4) sustained vowel sounds are computed using a cascade formant synthesiser excited by an appropriately shaped glottal pulses.

Maeda's articulatory model

The core of the SMIP is Maeda's articulatory model [2] together with a variant proposed by Gabioud [3]. It was built from a thorough statistical analysis of 519 hand-drawn midsagittal contours sampled at a rate of 50 frames/sec, obtained from synchronized radiographic and front/profile labiographic films shot at the *Strasbourg Institute of Phonetics* for one subject (PB) uttering ten meaningful French sentences [4]. The midsagittal contours were described as the 28 x-y coordinates of the intersection points of the two contours describing the vocal tract with a semi-polar grid. Seven parameters enable to explain 88 % of the variance of the observed variance of tongue contours. The percentages of explanation of the variance for tongue shape and jaw are distributed as follows: 15 % for the vertical position of the jaw, 43 % for the tongue body displacement, 23 % for the tongue dorsum, 7 % for the apex position. A linear combination of the seven parameters enables the re-construction of the vocal tract midsagittal contour. The inner (intero-labial) and outer (external arches of the vermilion) contours of the lips seen from front are based on the last version of lip models developed at ICP [5].

From the midsagittal contour to the area function

The 2D midsagittal function is converted into 3D area functions using a set of (α, β) coefficients derived from general data on vowels, and from cast and scanner measurements on constriction zones obtained at the ICP [6] in collaboration with the Grenoble University Hospital Center (maxillo-facial surgery service, Dr. Lebeau; radiographic service, Pr. Crouzet). The cross-sectional area S is computed from the midsagittal distance d as $S = \alpha\, d(x)^{\beta}$, where β has a constant value of 1.5, and α depends on the region in the vocal tract (glottis, lower part of the pharynx, upper part of the pharynx, oro-pharynx region, velar region, hard palate zone, alveolar region, intero-labial lip region), and on the dimension of the cross-section width (separate sets of coefficients for dimensions less than 1 cm and greater than 2 cm, with interpolation for the intermediate values). An alternative possibility is proposed [7] where the α coefficient depends continuously on the coordinate along the vocal tract midline x, and have been optimised for vowels and and fricative sounds.

In addition, the three traditional parameters that characterize area functions are computed: the lip area (Al), the position of the intra-oral constriction (relative to the glottis Xcg, or better to the teeth Xct) and the constriction cross-sectional area .

From area functions to formants, bandwidths and transfer functions

Three possibilities are offered to compute formants, bandwidths and transfer functions: (1) line analog frequency domain simulation of the vocal tract acoustic transfer function, and extraction of the corresponding complex conjugate poles [8]; (2) estimation of the resonances by means of a variationnal method [9]; (3) direct estimation of the formants expressed as 3rd order polynomes of the seven articulatory parameters, with coefficient optimized from a codebook generated by the model [10].

A given vowel sound can be located in the Maximal Vowel Space MVS [11], i.e. an nD space (with $2 \leq n \leq 5$). The MVS of the model has been evaluated by performing an extensive exploration of combinations of input parameters. Thus, each vowel calculated by simulation can be plotted in the F1-F2 and F2-F3 projections of the MVS. If an occlusion appears in the vocal tract, a warning message is sent and no calculation is performed. The formant bandwiths needed for sound synthesis are extracted from the complex poles, when the analog model is selected. For the others two cases, the bandwidths are estimated as proposed by Båvegård et al. [12]

The transfer functions (251 frequency points over the 0-5 kHz range, amplitudes in dB) are computed using an analog simulation [8]. The boundary conditions are: (1) a closed glottis, (2) distributed wall impedances, (3) a lip radiation impedance simulated as a piston in an infinite baffle. Heat conduction and friction losses are also included.

From formants and bandwidths to acoustic signal

The vowel sounds are computed using a simplified synthesis model consisting of a cascade of five formant filters (center frequencies F1-F5 and bandwidths B1-B5), excited by an appropriate glottal waveform [13]. Particularly, each glottal pulse is pitch-synchronously damped to simulate the effect of the periodic glottis opening. In order to obtain a more natural sound, Fo excitation variations and Intensity Level envelopes have been extracted from natural isolated vowel sounds uttered by a speaker. Finally, the signal can be listen to directly through loudspeakers, and a sound file is also stored on disk. Reasonable quality sound is generated (8 bits, 22 kHz).

SOFTWARE IMPLEMENTATION

Environment and platform

Several criteria have been considered for the choice of the platform: standard configuration with no additional hardware for graphics and audio output, reasonable processing time, ease of distribution and update, and reasonable chances of longevity, considering anticipated software and hardware evolution. A Macintosh platform (Macintosh II, Powerbook, and Quadra series) has been selected.

All software has been rewritten in the C language (Think C) on the basis of original software written in Fortran (the acoustic model and the sound generation). The interface has been developed in the HyperTalk language, in a HyperCard environment. Several facilities of "object-like programming" in HyperTalk have been used: additional menus, palette; the number of XCMD (Hypercard External commands) has been limited to the strict minimum in order to ease portability. The display is automatically adapted to the size of the available screen (from 14 inches to 21

inches), and sound is output through standard Macintosh sound resources (8 bits, 22 kHz). The coprocessor (68882) is necessary for reasons of processing speed.

Macintosh platforms

All platforms with 68020-68030-68040 microprocessors, 68882 coprocessor, with at least 4 Megabytes of RAM can be used. A 14 (or more) inch screen is required, and no special sound I/O card is necessary: the SMIP generates Macintosh sound resources. The SMIP requires a system 7.x version. SMIP does not run on a PowerMac so far.

AKNOWLEDGEMENTS

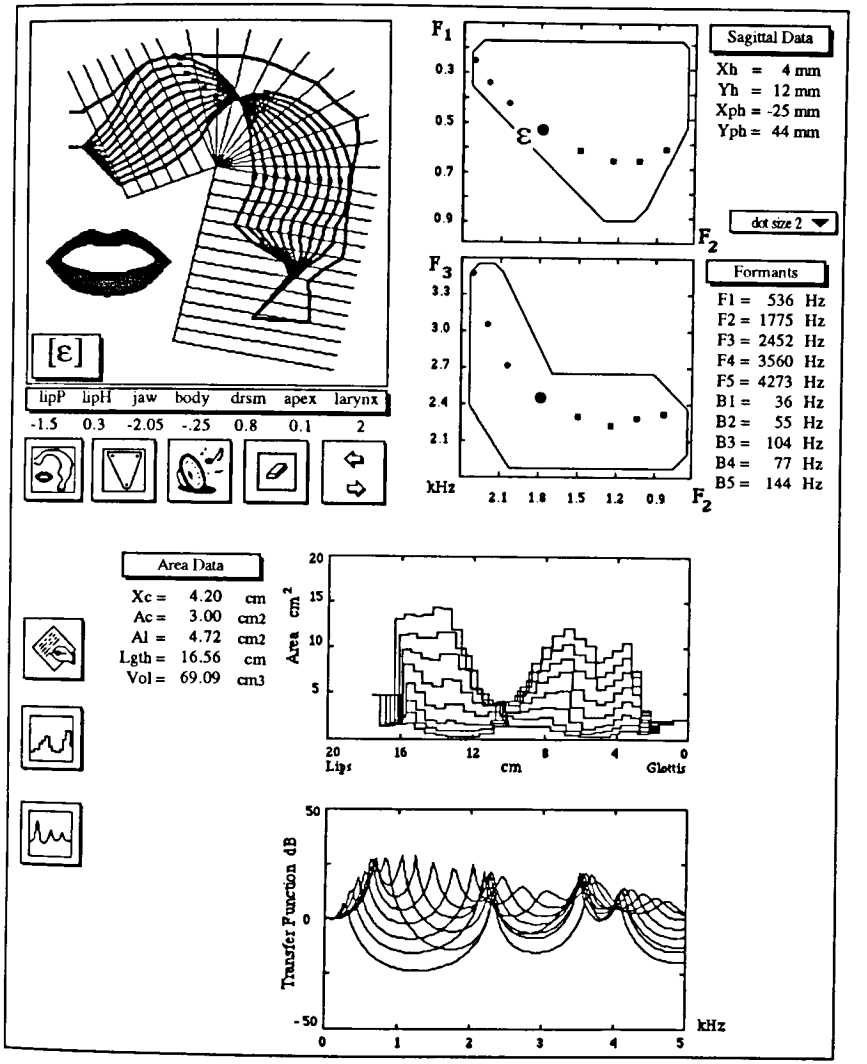
We would like to thank Paul Jospa of the institute of Phonetics in Brussels who provided us with the formant calculation program (variational method). Thanks a lot to Pierre Badin, Sonia Kandel, and Caroline Smith for improvements to the manuscript and to all the first users for their helpful criticisms

THE CONTRIBUTORS

The SMIP is the result of fruitful exchanges between Shinji Maeda, the ICP and the Institute of Informatics of Lausanne. Many direct or indirect contributors have been involved:
Adaptation of the Maeda model: P. Perrier and V. Jacquart. Translation of the sofware in C language: B. Gabioud. *Lip models:* C. Abry, L.J. Boë, P. Perrier, C. Benoît and T. Guiard-Marigny. *Coefficients from sagittal dimensions to area function*: L.J. Boë, Pascal Perrier, Rudolf Sock, D. Beautemps, P. Badin, R. Laboissière. *Harmonic simulation and formant estimation:* P. Badin and G. Fant, P. Jospa, A. Morris and E. Reynier. *Pole simulation*: G. Feng. *Translation of the sofwares in Think C:* B. Gabioud. Vowel Prototypes: N. Vallée and J. Payan. *HyperCard general conception:* L.J. Boë and B. Gabioud. HyperCard and XCMD developments: L.J. Boë, B. Gabioud, S. Bernier, P. Vacchino, F. Pinet, D. Guillem, L. Galmiche, A. Dumay, P. Déquier, M. Chaize, E. Grimont, and J. de Combret (Diadème Society).

REFERENCES

[1] Abry C., Badin P. & Scully C. (1994) *Sound-to-gesture Inversion in Speech:* The Speech Maps *Approach.* ESPRIT Research Rept. 6975. In *Advanced Speech Applications.* Varghese K., Pfleger S. & Lefèvre J.P. (Eds.), 182-196. Springer Verlag, Berlin.
[2] Maeda S. (1989) *Compensatory Articulation during Speech: Evidence from the Analysis and Synthesis of Vocal-Tract Shapes using an Articulatory Model.* In *Speech Production and Modelling*, 131-149. W.J. Hardcastle & A. Marchal (Eds.), Academic Publishers, Kluwer.
[3] Gabioud B. (1994). *Articulatory Models in Speech Synthesis.* In E. Keller (Ed.), *Fundamentals of Speech Synthesis and Recognition:*, 215-230, Chichester, John Willey.
[4] Bothorel A., Simon P., Wioland F., & Zerling J.-P. (1986). *Cinéradiographie des voyelles et des consonnes du français.* Institut de Phonétique, Strasbourg.
[5] Guiard-Marigny T. (1992). *Modélisation des lèvres.* DEA, Institut National Polytechnique de Grenoble.
[6] Perrier P., Boë L.-J. & Sock R. (1992) Vocal Tract Area Functions Estimation from Midsagittal Dimensions with CT Scans and a Vocal Tract Cast: Modelling the Transition with two Sets of Coefficients. *J. of Speech and Hearing Research*, 35, 53-67
[7] Beautemps D., Badin P., & Laboissière R. (1995). Deriving vocal-tract area functions from midsagittal profiles and formant frequencies. *Speech Communication*, 16, 27-47.
[8] Badin P., & Fant G. (1984). Notes on vocal tract computations. *STL Quaterly Progress Status Report*, 2 -3, 53-108.
[9] Jospa P. (1992). Caractérisation variationnelle des modes de résonance dans le conduit vocal. *Rapport d'Activité de l'Institut de Phonétique de Bruxelles*, 28, 13-30.
[10] Morris A. (1992) Least-squares fit to Maeda model dictionary. *Technical report, ICP*, Grenoble, 8 p.
[11] Boë L.-J., Perrier P., Guérin B., Schwartz J.-L. (1989) Maximal Vowel Space. *EuroSpeech 89*, 2, 281-284.
[13] Feng G. (1983). Vers une synthèse par la méthode des pôles et des zéros. *13e J* EP (GFCP, SFA) 155-157.
[12] Båvegård M., Fant G., Gauffin J., & Liljencrants J. (1994). Vocal tract sweeptone data and interpretation. In S. Maeda (Ed.) *From Speech Signal to Vocal Tract Geometry.* PPR 2, European ESPRIT/BR N° 6975 *Speech Maps* project. Vol. III.



For prototypic [ε]:
• The sagittal contour of the vocal tract generated by the Maeda model, and macro-variations calcultaded for ± 3σ variations of the tongue body.
Lip shape contours are calculated by using the ICP models. The upper point of the tongue body (Xh, Yh), and the furthest point of the tonge root in the pharynx (Xph, Yph) are indicated by dots (■).
• Area function derived from contour In addition, the three traditional parameters that characterize area functions are computed: the lip area (Al), the position of the intra-oral constriction (Xc) relative to the teeth, the constriction cross-sectional area (Ac), the length and the volume of the vocal tract.
• The transfer function derived from the area function, the associated F1-F4 formants and B1-B4 bandwidths, and the the location of [ε] in the Maximal Vowel Space of the model.

INTERACTION BETWEEN GLOTTAL AND VOCAL-TRACT AERODYNAMICS IN A COMPREHENSIVE MODEL OF THE SPEECH APPARATUS

Paul Boersma
Institute of Phonetic Sciences, Amsterdam, The Netherlands

ABSTRACT

Our computer model of the speech apparatus converts muscle activities into the resulting tissue movements, air pressures, air velocities, and sound.

THE MODEL

The entire speech apparatus (lungs, glottis, mouth, nose) is modelled with 80 tube sections that contain air (fig. 1).

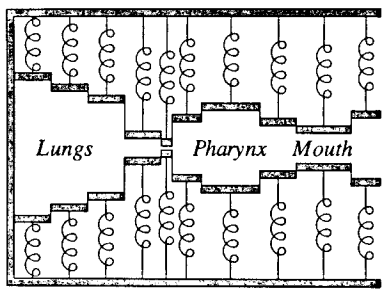


Figure 1. The whole speech apparatus, including the glottis, is modelled with the same kind of tubes. Only a few of these tubes are shown here.

Each tube section has (a) two parallel stiff walls, (b) two near-parallel walls that can move under aerodynamic and myoelastic forces and whose equilibrium positions and tensions are controlled by the muscles, and (c) two boundaries that connect the tube to the rest of the world.

Figure 2 shows the four different forms of these boundaries: they are either (1) closed, (2) connected to one other section, (3) connected to two other sections, or (4) open to the outer air.

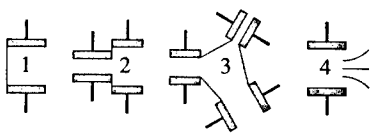


Figure 2. Types of tube boundaries.

The sound-generating algorithm is not specific to speech: it works equally well for an arbitrary structure of ducts.

All tube sections can have different and time-varying widths and lengths. The finite-differencing integration scheme that solves the resulting aerodynamic problem without approximating away any pumping and sucking effects, was presented in [1]. In the present paper, I will show examples of some of the phenomena that the model can describe realistically.

Our model speaker is characterized as an average adult woman.

LUNG PRESSURE

Because the lungs are described in the same way as the vocal tract, subglottal pressure is a direct consequence of controlling the equilibrium (target) width of the lungs.

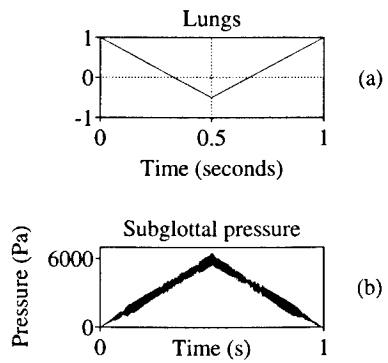


Figure 3. Lungs: input (target width) and output (subglottal pressure), if the vocal cords are adducted (phonation).

Figure 3 shows how this relation is realized during the production of the vowel [a]. The change in target width is much more reflected in the lung pressure than in the lung volume. The intensity of the uttered sound is an increasing function of lung pressure: the slope is 10 dB per kPa for pressures below 1.3 kPa, and 2 dB per kPa for pressures above. The fundamental frequency varies by 100 Hz and 20 Hz per kPa, respectively.

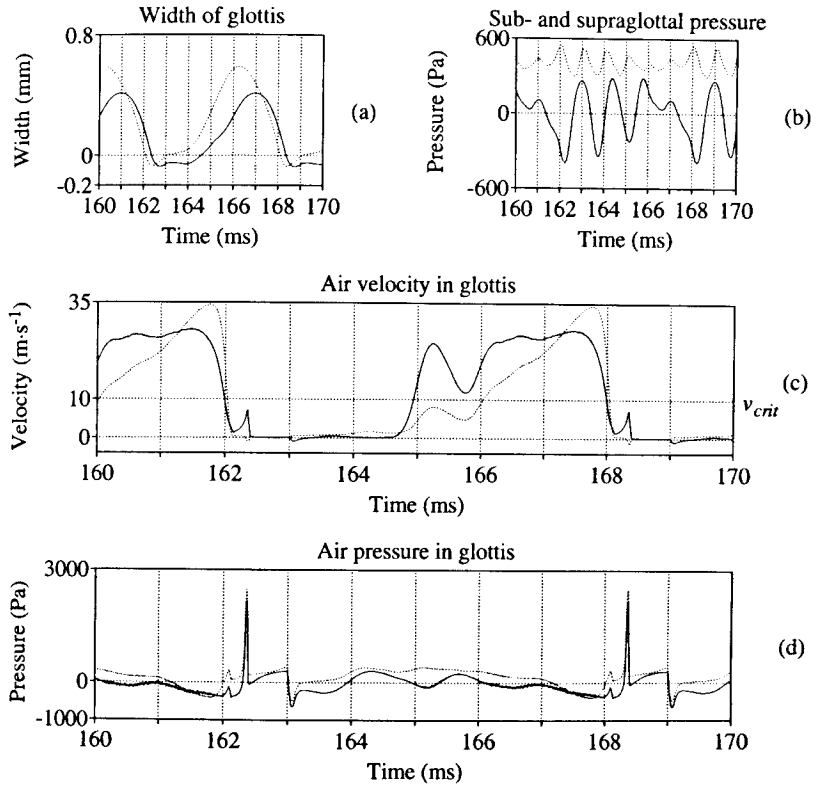


Figure 4. Events in the glottis during phonation. The dotted curves refer to the lower part of the glottis or to the subglottal pressure, and the solid curves refer to the upper part of the glottis or to the supraglottal pressure.

EVENTS IN THE GLOTTIS

Because the glottis is described in the same way as the vocal tract, we get a detailed view of the air pressures and particle velocities in the glottis. Phonation (fig. 4) automatically results when the lungs contract while the vocal cords are adducted and the supralaryngeal pathway is unobstructed.

Figure 4a shows that the glottis is closed about half the time.

Fig. 4b clearly shows the subglottal and supraglottal first formants. They are more strongly damped when the glottis is open than when it is closed.

Fig. 4c shows that between 165 and 166 ms, the air velocity in the glottis is a direct result of the difference between the formant pressures. We also note in this figure the velocity drop when the glottis closes at 168.1 ms, and the sudden velocity peak arising at 168.3 ms between the closing upper parts of the vocal cords when the last amount of air is forced into the pharynx while the lower glottis is already closed.

If the velocity is above 10 m/s, noise is generated, as we see in the upper glottis between 166 and 168 ms (fig. 4d).

Also in fig. 4d, we see the reason why the upper parts of the vocal cords hesitate to open between 163 and 164 ms (fig. 4a), thus causing the long closure interval: as the lower glottis opens, the air is rarefied there and the pressure drops to negative values; this sucks the upper parts of the vocal cords together. In the two pressures of fig. 4d, we also see the formants, and positive pressure peaks when the vocal cords collide.

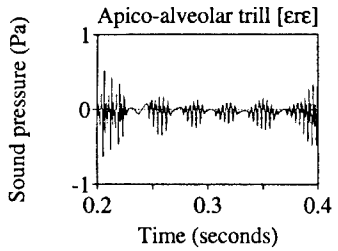


Figure 5. Vibration outside the glottis.

TRILLS

Because the vocal tract is described in the same way as the glottis, our model speaker can generate apical, labial, and uvular trills as easily as vibrating vocal cords. This is shown in fig. 5.

FALSETTO

Our model speaker's voice enters the falsetto register when, other things staying equal, the lung pressure becomes very low. Fig. 6 shows that the lower parts of the vocal cords do not close. The register break occurred when the pressure had fallen to 250 Pa during sustained phonation. Incidentally, our speaker managed to phonate for 20 seconds starting from only 400 Pa without adjusting her lung width, which is much longer than our *male* model speaker can do (11 seconds) and also longer than in reality. This sugggests that a realistic modelling of the amount of air that leaves the lungs during phonation, can only be achieved if we model a substantial leak parallel to the glottis, especially for female speakers. Our model can easily handle this if we add two three-way boundaries (fig. 2), but this was not used for the present paper.

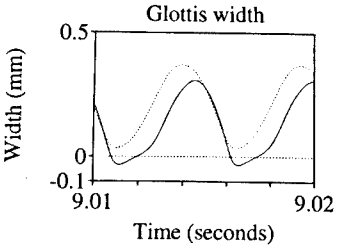


Figure 6. Falsetto at a mean transglottal pressure of 200 Pa. The upper part of the glottis (solid curve) closes, but the lower part (dotted curve) does not.

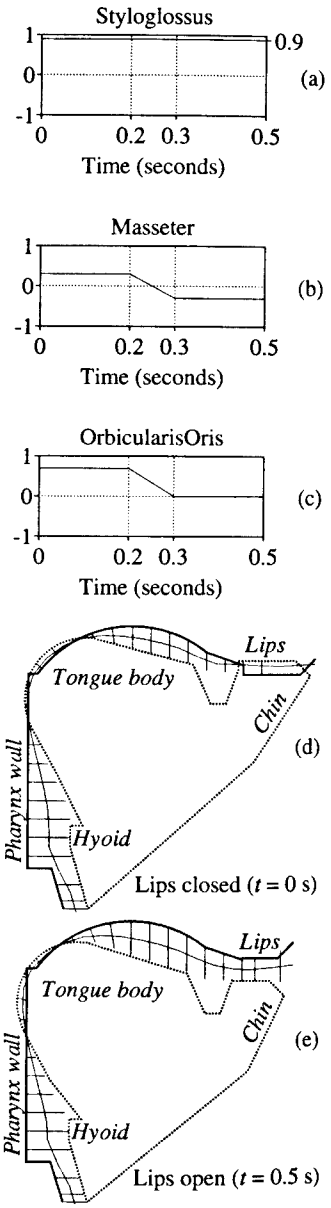


Figure 7. Input for a bilabial click: muscle activities and target shape. Styloglossus pulls the tongue back up. "Masseter" stands for all the muscles that close or open the jaw. Orbicularis oris rounds and protrudes the lips.

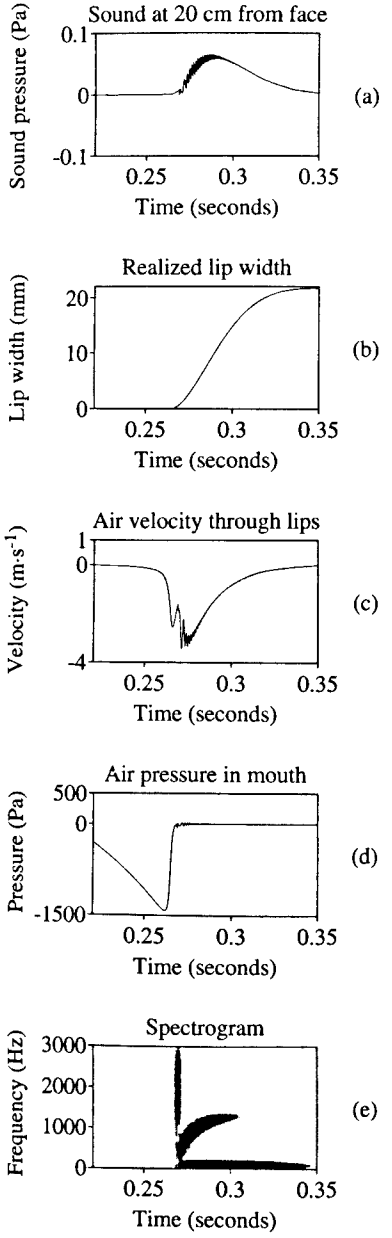


Figure 8. Output of a bilabial click: sound, movements, and aerodynamics.

VOICING IN OBSTRUENTS

Our model speaker can make an [aba]-[apa] contrast by varying her oral wall stiffness only [1]. This suggests that for implementing voicing contrasts languages can use supralaryngeal myoelastic features, apart from more drastic measures like aspiration or constriction.

A BILABIAL CLICK

Because of the realistic modelling of sucking effects, our model speaker can produce click consonants. Figure 7 shows how our model speaker lowers her jaw and unrounds her lips while maintaining a velar closure. The lungs and glottis are not involved.

Figure 8 shows the acoustic, aerodynamic, and myoelastic results. Jaw lowering starts at 0.20 seconds (not in the figures), which causes the air pressure in the mouth to fall to -1500 Pa relative to the atmospheric pressure (fig. 8d). At 0.27 seconds, the lowering of the jaw causes the lips to separate (fig. 8b), which causes air to flow from outside into the mouth through the lips (fig. 8c). This air flow, which reaches a velocity of 3 m/s, quickly restores the air pressure to the atmospheric pressure (fig. 8d). The resulting sound (fig. 8a) has a tiny burst, which is seen as a vertical band in the spectrogram (fig. 8e; we used a Gaussian window with a -22 dB length of 10 ms, 30 dB dynamic range, and 6 dB/octave pre-emphasis). After the burst, the sound (fig. 8a) shows, superposed on the DC flow, a sine wave with a frequency that rises from 300 Hz to 1200 Hz, which is also reflected in the velocity (fig. 8c) and more clearly seen in the spectrogram (fig. 8e). This *formant transition* is what we would expect for an opening gesture of the lips from an [u]-like position to the [a]-like position of fig. 8b, where the lips end up 22 mm apart.

The auditory impression from the sound (fig. 8a) is correct: the seven listeners that were asked to identify this sound, spontaneously reproduced a non-affricated bilabial click.

REFERENCES

[1] Boersma, P. (1993), "An articulatory synthesizer for the simulation of consonants", *Proceedings Eurospeech '93*, pp. 1907-1910.

ASYNCHRONY MEASURE OF LIP-TONGUE-JAW MOVEMENTS

H-H. Bothe [1], C. Mooshammer [2], S. Kuhrt [1], and B. Pompino-Marschall [2]
[1] Technical University of Berlin, Electronics Institute, Berlin, Germany
[2] Forschungsschwerpunkt Allgemeine Sprachwissenschaft, Berlin, Germany

ABSTRACT

This paper describes a method of analyzing timing correlations between characteristic *independent movements* of the lips and the tip of the tongue after model based elimination of jaw movements. The Dynamic Time Warping algorithm was applied to time series of articulographic measurements. The investigations lead to a complex similarity measure for the articulatory processes as a degree of coordination as well as to time series of coproduction data. Future goal is to add realistic tongue movements to an existing facial animation computer program.

DATA ACQUISITION

Text Corpus

For data acquisition, one German speaker produced five repetitions of a nonsense-word corpus of the form /ge-CVC-e/ with C=/p, t, k/ and V=all full vowels of German in the carrier-sentence 'Ich habe ... gesagt.' ('I said ...') in randomized order. In this study, we analysed the subset of sequences with lax /a, œ/ and the consonants /p/ and /t/.

Sensor Mounting

To monitor articulatory movements Electromagnetic Articulography (AG100, Carstens Medizinelektronik; see [1]) was used. One sensor $<c_0>$ was mounted on the lower incisors (jaw), one $<c_1>$ on the lower lip, and one on the midline of the tongue, 1 cm from the tip of the tongue $<c_2>$ (see Figure 1). To compensate for head movements two reference coils were attached to the upper incisors and the bridge of the nose. After measuring the occlusal plane the data were translated and rotated with respect to this new (x,y)-coordinate system. Whereas the x-axes is given by inter-section of occlusal and sagittal plane, the y-axes is scaled othogonal to it. The x-offset of which is defined by the coil position of the upper incisors.

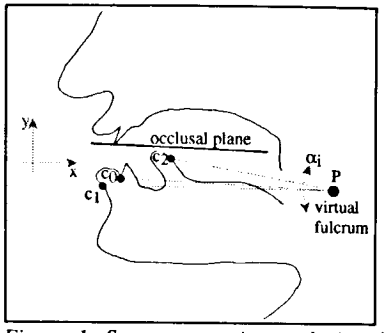


Figure 1. Sensor mounting and virtual fulcrum.

Segmentation of Time Signals

Elimination ofjaw movements and DTW was applied to articulatory time signals, beginning at the onset of the first consonant and ending at the offset of the second consonant (see Figure 2).

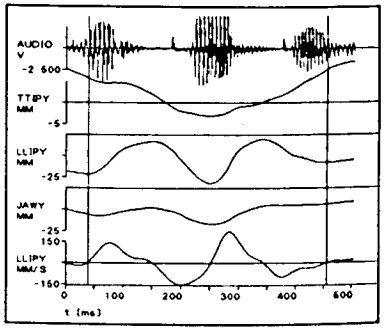


Figure 2. Segmentation of the analysed articulatory time signals for /pap/.

On- and offsets were determined by means of zero-crossings of the vertical velocity signals of the corresponding articulator, i.e. tongue tip coil for apical stop and lower lip for the bilabial stop.

ELIMINATION OF JAW MOVEMENTS

The directly measured cartesian $<c_1>$- and $<c_2>$-coordinates x' and y' were corrected by the corresponding $<c_0>$-rotation around a virtual fulcrum (see Figure 1). Reference was the rest position of the lower incisors $<c_0>_{rest}$.

Goal of the correction is to separate *independent movements* of lips and tip of the tongue from those induced by the jaw motion.

The calculation model of the virtual fulcrum (vf) extends the idea of pure vertical shift and proposes that $<c_0>$ moves approximately on a circle as shown in Figure 3.

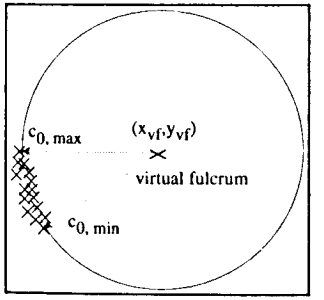


Figure 3. Calculation model of the virtual jaw fulcrum by circle approximation.

The (x_{vf},y_{vf})-location is calculated by curve fitting of sample jaw data (×) by means of a Genetic Algorithm. This iterative optimization method assumes a family of parameter vectors, eachof which composed of the free circle parameters x_{vf} and y_{vf} which are lined up in a bit string and Gray-coded ([2]; Figure 4).

The parameters are optimized by random change with the help of *mutation* (single bit change) and *crossover* (changes of longer parts of the string within a family of parameter sets). The used algorithm is described in [3].

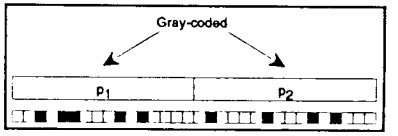


Figure 4. Gray-coded parameter string.

It shows up that the (vf)-position becomes stable at (82.4,-6.1) after a high number of iterations (>15000). The courses of the corrected secondary positions (x_i,y_i) were then compared by applying the DTW algorithm.

DYNAMIC TIME WARPING (DTW) OF 2D TIME DISCRETE SIGNALS

The DTW was applied to two prototypes of the same two-dimensional CVC time series of interest. Two principle courses of discrete (x,y)-positions over time τ with a sample rate of 250 [Hz] are shown in Figure 5.

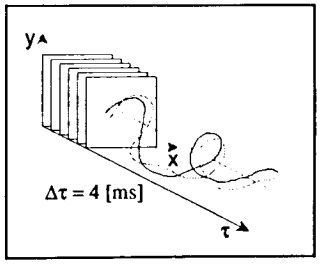


Figure 5. Time series of discrete (x,y)-positions of two sensors over time τ.

Applying DTW to a time discrete reference curve $C_1(\tau)$ and a test curve $C_2(\tau)$ results in a nonlinear projection between them together with a global distance measure of similarity. The nonlinear stretching is necessary with respect to different speed of the articulatory movements and different lengths of the phonemes. The principle of the algorithm is shown in Figure 6 (see also [4]).

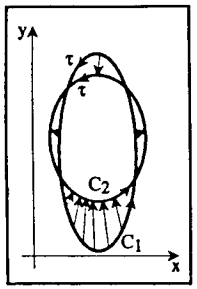


Figure 6. Nonlinear mapping of two sample curves $C_1(\tau)$ on $C_2(\tau)$ in (x,y) head coordinates over time τ.

In order to find the optimum projection at first a matrix **d** of local distances $d(\tau_i,\tau_j)$ between all C_1 (τ_i) and $C_2(\tau_j)$ has to be calculated. Applying the Euclidian distance measure $d_E(\tau_i,\tau_j)$ results in

$$d_E(\tau_i,\tau_j)^2=[x_1(\tau_i)-x_2(\tau_j)]^2+[y_1(\tau_j)-y_2(\tau_j)]^2.$$

The basic idea behind DTW is to find the path from the starting point $\tau_0=0$ to the end point τ_I on which the accumulated local distances become a minimum.

The sum $D(\tau_i,\tau_j)$ of the distances can be calculated by the recursive formula

$$D(\tau_i,\tau_j) = d(\tau_i,\tau_j) + \min [D(\tau_{i-1},\tau_{j-1}), D(\tau_{i-1},\tau_j), D(\tau_i,\tau_{j-1})]$$

as the sum of the local distance $d(\tau_i,\tau_j)$ and the minimum of the accumulated distances

$D(\tau_{i-1},\tau_{j-1})$, $D(\tau_{i-1},\tau_j)$ and $D(\tau_i,\tau_{j-1})$.

Solving the above equation requires a column oriented calculation in the three-dimensional (x,y,τ)-universe.

For the time being we are interested only in the vertical articulatory movements $y_{lips,\ tongue}(\tau)$ in order to position a two-dimensional bit pattern of the tongue in the later computer animation.

The calculation scheme for the o-path is shown in Figure 7.

The value $D(\tau_i,\tau_j)$ is interpreted as a measure of global similarity.

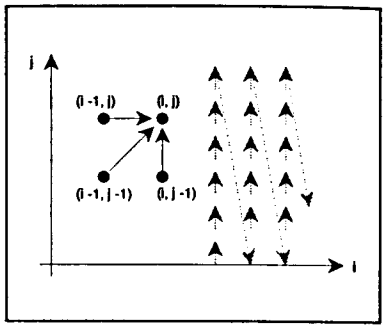


Figure 7: Column oriented calculation of the accumulated distance $D(\tau_i,\tau_j)$.

The projection of $C_2(\tau)$ on $C_1(\tau)$ is related to the minimum path calculated from the end point to the starting point of the matrix of accumulated distances **D** (a principal example of a one-dimensional mapping of $y_1(\tau)$ and $y_2(\tau)$ is shown in Figure 8).

In order to reduce the calculation time, a desired area of interest is taken into account around the diagonal of **D**.

The nonlinear stretching coefficients $d_{OP}(\tau_i,\tau_j)$ of the optimum path (o-path) represent the local similarities of the projection of $C_2(x,y,\tau)$ on $C_1(x,y,\tau)$.

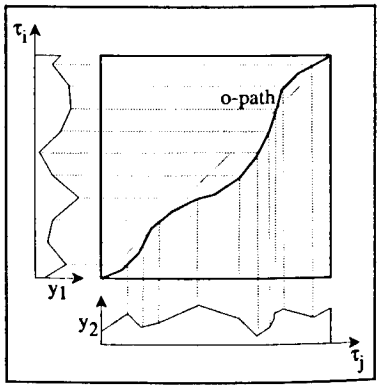


Figure 8. Principal of 1D- DTW mapping.

Restricting these coefficients to relatively small ε-values by $d_{OP}(\tau_i,\tau_j) < \epsilon$ can be interpreted as fixing the zones of similarity.

Some Exemplary Results

The DTW results in a diagonal either for total similarity of both curves or for uncorreleted curves.

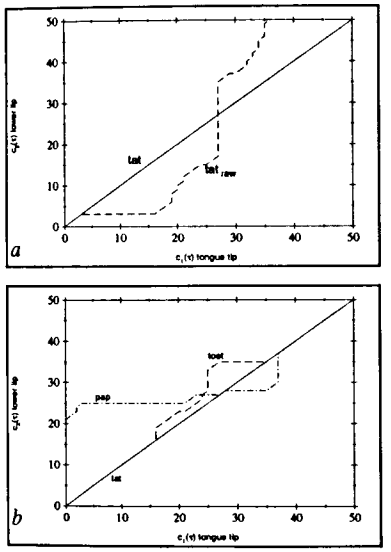


Figure 9. DTW path for a. /tat/raw, /tat/, b. corrected curves /tat/, /pap/, /tøt/.

Figure 9a shows that for uncorrected $/tat/_{raw}$ the time intervall $\Delta\tau=[16, 35]$ in msec of the tongue tip is mapped on [4, 50] of the lower lip, whereas after correction the movements are similar. DTW curve /pap/ in Figure 9b shows that [0, 40] of the tongue tip is mapped on a nearly constant lower lip position at ca. 25, and thus, determines highly independent movement during this time. The course of similarity for /tøt/ can be interpreted as an achievment of lip rounding.

Summary and Conclusion

The above work shows basic investigations for the design of a model based computer animation program that displays visual articulation movements with moving lips, teeth and tongue tip on the computer screen.

Whereas in the existing program the grey-scale film is created by a codebook of key-pictures and a morphing algo-rithm for lips, skin, and teeth [5], the above investigations make possible the implementation of coordinated movements of the tongue.

Acknowledgement

The text corpus has been designed and recorded within a DFG project (DFG = German Research Council) under grant Ti 69/29-2. Many thanks for help especially to Phil Hoole.

References

[1] Perkell, J.S., Cohen, M.H., Svirsky, M.A., Matthies, M.L., Garabieta, I., Jackson, M.T.T. (1992), "Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements", *J. Ac. Soc. of America*, vol. 92, pp. 3078-3096.

[2] Mathias, K.E., Whitley, L.D., *Transforming the search space with Gray-coding. Proc. 1st IEEE Conf. on Evolutionary Computation*, pp. 513-518, Orlando, USA.

[3] Thierens, D., Goldberg, D. (1994), *Elitist Recombination: An Integrated Selection Recombination GA*. Proc. 1st IEEE Conf. on Evolutionary Computation, pp. 508-512, Orlando, USA.

[4] Bothe, H.H., Rieger, F., Tackmann, R. (1993), *Visual coarticulation effects in syllable environment*, Proc. EUROSPEECH '93, pp. 1741-1744, Berlin, Germany.

[5] Bothe, H.H., Rieger, F. (1993), Visual speech and coarticulation effects. Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP) '93, pp. V634-V637, Minneapolis, USA.

ARTICULATORY CHARACTERISTICS OF SINGLE AND BLENDED LINGUAL GESTURES

Dani Byrd†
UCLA Department of Linguistics
†*now at Haskins Laboratories, 270 Crown Street, New Haven, CT 06511-6695 USA*

ABSTRACT

This study uses electropalatography to investigate the temporal and spatial organization of lingual consonants in English consonant clusters. Reduction differences between onsets and codas and between stops and fricatives are observed. Gestural overlap explains contact patterns in juncture geminates.

INTRODUCTION AND METHOD

The sequences considered are (1) juncture geminates—[d#d], [s#s], & [g#g]—which are realized with a single raising and lowering of the tongue, and (2) sequences in which one lingual consonant occurs with a labial consonant —[d#b], [b#d], [s#b], [b#s], [g#b], & [b#g]. The sequences were read in the phrase *"Type baC Cab again"* and recorded using electropalatography. Seven tokens of each sequence from each of five speakers are analyzed. In all the sequences, only a single consonant articulation is made against the palate, and no other consonantal constriction interferes with it. Based on EPG contact profiles, metrics were calculated indicating the spatial and temporal extent of the lingua-palatal contact. Effects of syllable position, place, and manner are tested. For a more detailed description of the method see [1], [2].

RESULTS

The contact profiles for the front region for [d#d] (geminated), [d#b] (coda), and [b#d] (onset) are shown for one of the five speakers in Figure 1. The y-axis represents the percent of the pseudopalate region (front or back) registering lingual contact at a particular point in time; the x-axis represents time in frames of .01 seconds. The null hypothesis, clearly not supported, is that the three contact profiles for a consonant—juncture geminate, coda, and onset—are the same.

Comparison of Means

First let's consider differences in the amount of contact in the front region for [d]. This is indexed by the maximum contact expressed as a percentage of the total possible contact in the front region. Repeated measures ANOVA determines there to be a significant effect of sequence on the maximum front contact for [d] ($F(2,8)= 5.75$, $p=.0283$) such that the coda [d]'s have less maximum contact than [d]'s in the other two sequences. There is also a significant interaction with speaker, with Speaker B having the reverse pattern. The other speakers have a group mean of 58% contact for the onset [d]'s and 42% for codas. The combined mean maximum contact for [d#d] across speakers is

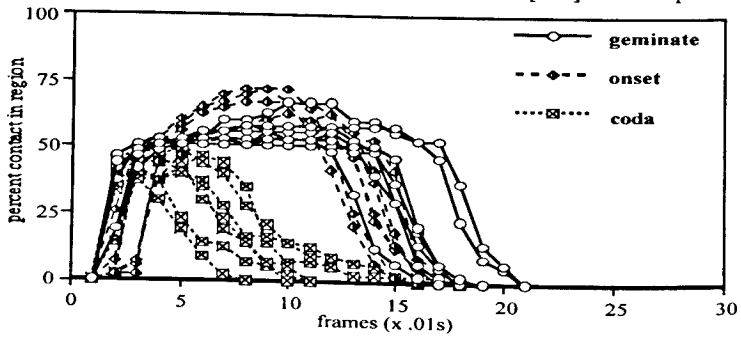


Figure 1. Contact profiles in the front region for [d#d], [d#b] & [b#d], Speaker K.

56%. This parallels results reported elsewhere on reduction of final tongue tip gestures [3], [4].

The duration of linguapalatal contact for [d] is longer in [d#d] than for the other sequences ($F(2,8)=12.55$, $p=.0034$). For three Speakers, A, K, and M, codas were shorter than onsets as well. Next, consider the shape of the profile, or temporal distribution of contact. One measure of this is skewness. Roughly speaking, a greater (positive) skew indicates a shorter closure formation in the contact profile than closure release. For these same three speakers, codas had a greater positive skew than onsets. There was a significant interaction of sequence and speaker in affecting skew ($F(8,90)= 19.460$, $p=.0001$), with Speaker K showing the strongest effect. The time taken in forming the contact was shorter than that needed for the release (*i.e.* positive skew) for both coda and onset [d]'s with the asymmetry being greater in coda position. Another measure of shape is the FLATNESS of the contact profile, indexed here by the mean contact divided by the maximum contact. There was a significant interaction of speaker and sequence on FLATNESS ($F(8,90)= 6.464$, $p=.0001$). All speakers had flatter profiles for [d#d] than for the other sequences. Three speakers, K, A, and B, also had flatter onsets than codas.

The next sequences considered are [s#s], [s#b], and [b#s]. The contact profiles for these are shown for Speaker K in Figure 2. First consider the degree of contact. ANOVA shows there to be no difference among the sequences [s#s], [s#b], and [b#s], in maximum contact in the front region. Next, duration of contact is of interest. There are significant differences in duration ($F(2,8)= 18.35$, $p=.001$). The juncture geminates have the longest durations of contact, and, excepting Speaker S, onsets are longer than codas. For all speakers, contact profiles for codas are less flat and have a more positive skew than either the geminated or onset consonants, as determined by the main effect of sequence on FLATNESS ($F(2,8)= 8.125$, $p=.0118$) and SKEW ($F(2,8)=16$, $p=.0016$). In fact, the coda [s] was the only one of the three [s]'s to have a positive skew for all subjects; only two subjects had a positive skew for the onset [s]. This parallels the findings for [d], suggesting that while coda [s]'s may not undergo spatial lenition, they, like coda [d]'s, are shorter and have faster constriction formation in coda position than in onset position.

The dorsal stop consonant [g] was examined in the sequences [g#g], [g#b], and [b#g]. Here the relevant articulatory region is the back one. The contact profiles are shown for Speaker K in Figure 3. ANOVA determines there to be a significant effect of sequence on maximum displacement in the back region ($F(2,8)=5.476$, $p=.0318$). For all speakers, except Speaker B, displacements decrease from onset to geminated to coda [g]'s. For Speaker B, onsets rather than codas have the lowest maximum contact; codas are still less displaced than geminated [g]'s. There is also a significant difference between the duration of [g] in the three sequences ($F(2,8)=9.51$, $p=.0077$). All speakers'

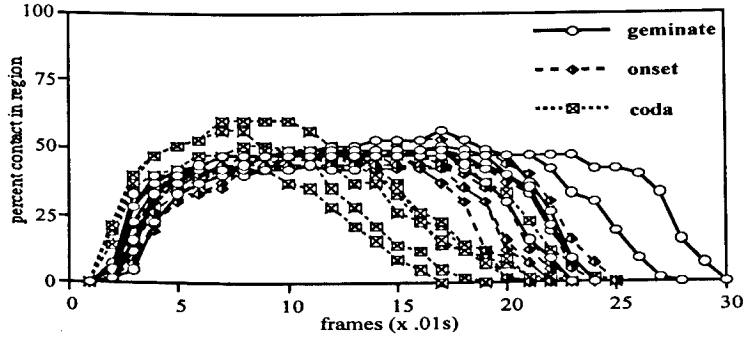


Figure 2. Contact profiles in the front region for [s#s], [s#b] & [b#s], Speaker K.

geminated sequences are longest, and all, except Speaker B, have longer onset [g]'s than coda [g]'s. There is also a significant effect on FLATNESS ($F(2,8)=5.788$, $p=.0279$). Contact profiles for four speakers are flatter for codas than onsets. There is no main effect on SKEW for the velar consonant, although there is a significant interaction of speaker and sequence ($F(8,90)=8.511$, $p=.0001$) with Speakers K, A, and S having onsets more skewed to the right than codas, although most speakers' skews for all three sequences are negative.

Summary

Graphical comparisons illustrating differences in reduction and shortening are shown in Figure 4. (However, recall that Speaker B's data often pattern opposite to those of the other speakers, making an examination of the group means less representative of the general behavior.) In summary, for the stop consonants, onsets are generally more displaced (*i.e.* have greater maximum lingua-palatal contact) than codas. For most speakers, the stops in onset are also longer than in coda. For [s], contact in onset position is longer than in coda position, but there is no difference in spatial extent. Contact profiles for front consonants are generally flatter and less (positively) skewed in onset position than in coda position. This difference is probably due to the need for the jaw to lower for the immediately following vowel in the onset sequences but the possibility of a slower release due to a longer high jaw position for the following bilabial consonant in the coda sequences. (Changes in jaw height would have a much smaller influence on [g] skewness due to the posterior hinge location of the jaw.) For the back stop consonant, contact profiles are flatter in coda position than in onset position. Finally, for all three consonants, the juncture geminates are longer and flatter than both onset and coda consonants.

JUNCTURE GEMINATES

Munhall and Löfqvist [5] examined the blending of two laryngeal gestures separated by a word boundary. This situation is analogous to our geminated sequences where two lingual gestures are canonically present. They observed that a single smooth movement occurred as the gestures overlapped at fast speaking rates. Lingua-palatal contact profiles for our juncture geminate sequences also show a single smooth movement. Both studies find the coproduced movement for juncture geminates to be longer than the non-coproduced movement for a single gesture. Additionally, Munhall and Löfqvist [5] found no consistent tendency for the combined single movement to be larger than an individual (non-coproduced) movement, although a simulated summation of the gestures predicts such a difference. At medium speech rates one of their speakers showed larger geminated movements but this behavior reversed at fast rates. Their other speaker showed no consistent difference. Our data above also show no consistent increase in maximum contact for the geminated consonants, suggesting that a summation process is not at work. (In fact, data here and in [6] suggest some

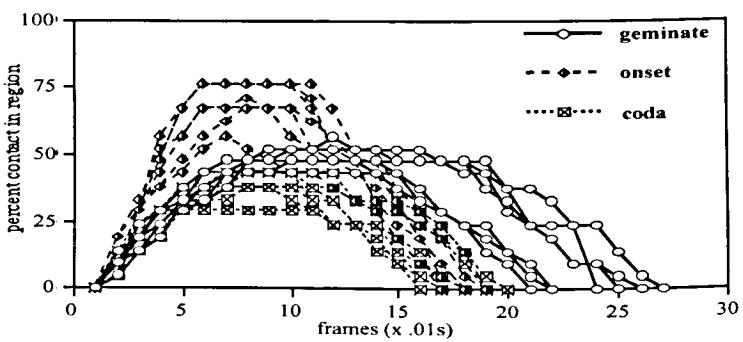


Figure 3. Contact profiles in the back region for [g#g], [g#b] & [b#g], Speaker K.

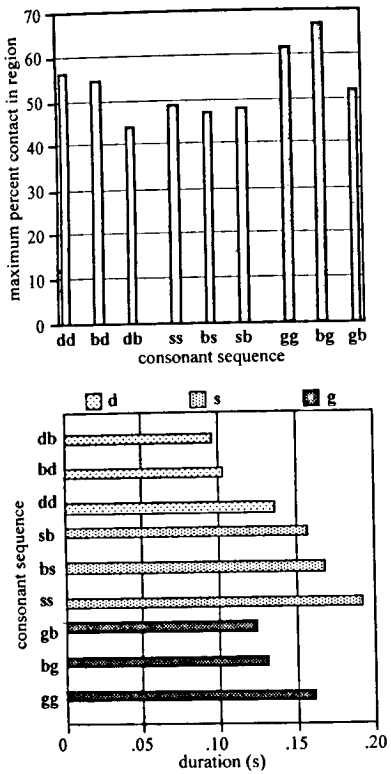


Figure 4. (top) Mean maximum percent contact in region. (bottom) Mean duration of contact in region.

tendency for the onset to be larger than the geminate.) Similarly, Kelso and Tuller [7] (cited in prepublication form in [8]) report that a larger gesture would typically have both increased amplitude and steeper onset and offset slopes. Partly on this basis, Browman and Goldstein [8] argue for the presence of two overlapping bilabial gestures in Chaga [mp] sequences because of the similarity in slope and amplitude to single bilabial closure gestures. The result of the overlap is simply a longer movement. Likewise, we observe a pattern of similar slopes and amplitudes with longer durations for our [C#C] sequences as compared to the single consonants, arguing against gestural summation in these sequences. Furthermore, by virtue of Occam's Razor alone, the overlap account is preferable to the supposition of a new mechansim [8] or to the modification of a gestural score by the substitution of a single 'macro' gesture for the abutting lingual gestures (cf. [9]).

ACKNOWLEDGEMENTS

This work was supported by an NSF graduate fellowship & NSF grant DBS9213604 to the author & Pat Keating. Some of the above work can be found in [2]. Many thanks to Pat Keating, Peter Ladefoged, and Louis Goldstein.

REFERENCES

[1] Byrd, D., Flemming, E., Mueller, C.A., and Tan, C. C. (in press), "Using regions and indices in EPG data reduction", *JSHR*.

[2] Byrd, D. (1994), *Articulatory Timing in English Consonant Sequences*, UCLA Ph.D. diss.

[3] Barry, M. (1992), "Palatalisation, assimilation and gestural weakening in connected speech", *Sp. Comm.*, 11, pp. 393-400.

[4] Browman, C. & Goldstein, L. (1995), "Gestural syllable position effects in American English", in F. Bell-Berti and L. Raphael, eds., *Producing Speech: Contemporary Issues for Katherine Safford Harris*.

[5] Munhall, K. G. & Löfqvist, A. (1992), "Gestural aggregation in speech: Laryngeal gestures", *J. Phon.*, 20, pp. 93-110.

[6] Dunn, M. H. (1993), *The phonetics and phonology of geminate consonants: A production study*, Yale Univ. Ph.D. diss.

[7] Kelso, J. A. S. & Tuller, B. (1987), "Intrinsic time in speech production: Theory methodology, and preliminary observations", in Keller & Gopnik, eds., *Sensory and Motor Processes in Language*. Hillsdale, NJ: Erlbaum, pp. 203-222.

[8] Browman, C. & Goldstein, L. (1986), "Toward an articulatory phonology", *Phonology Yearbook*, 3, pp. 219-252.

[9] Recasens, D., Fondevila, J., Pallarès & Solanas, A. (1993), "An electropalatographic study of consonant clusters", *Sp. Comm*, 12, pp. 335-356.

PATTERNS OF LINGUAL VARIABILITY IN GERMAN VOWEL PRODUCTION

Philip Hoole[1] and Barbara Kühnert[1,2]
[1]Institut für Phonetik, Munich University, Germany
[2]Department of Linguistics, Cambridge University, England

ABSTRACT

This study aimed to assess the relative importance of biomechanical and linguistic constraints on articulatory precision by analyzing contextual and token-to-token variability in tongue postioning for vowels. Contextual variability proved greater for lax vowels. Back vowels showed substantially increasing variability towards more front tongue locations. Regarding token-to-token variability, lax vowels were more variable for front vowels, but less so for back ones. Again, back vowel variability increased towards the front. The main distinction in variability was thus between palatal vowels (whole tongue constrained) and non-palatal vowels (anterior tongue unconstrained).

INTRODUCTION

This study analyzes patterns of tongue configuration variability in the articulation of German vowels. The guiding assumptions are, firstly, that a physiologically realistic theory of vowel production must account for magnitude of both contextual (e.g coarticulation) and token-to-token (henceforth "T-T") variability, and secondly, that it is essential to view the vowels as a system. Consider here some major hypothetical influences on magnitude of variability. Firstly, finite-element modelling of the tongue (see [1] for discussion) suggests that for high vowels bracing of tongue against hard palate helps attain a stable tract configuration. Secondly, tense vowels may be more tightly controlled than lax vowels - in any case, better understanding of the precise physiological substrate of this distinction is an important issue in German [2]. Thirdly, crowded regions of a vowel system may be less variable than less crowded ones. If, however, the complete system is not examined the weight to be accorded these potential influences is difficult to assess. In fact, there are few articulatory studies investigating multiple repetitions of complete vowel systems. One exception is a glossometer study by Bohn et al. [3] of token-to-token variability in German. Surprisingly, the tendency was for high vowels and tense vowels to show more variability. The present study reviewed these results using a different technique (EMA) and extends them by examining contextual in addition to T-T variability.

METHOD

Six German speakers spoke 5 repetitions of a nonsense-word corpus of the form /gəCVCə/ with C1=C2=/p, t, k/ and with V consisting of 7 pairs of tense-lax vowels (/i: ɪ, y: ʏ, e: ɛ, ø: œ, ɑ: a, o: ɔ, u: ʊ/) embedded in a carrier phrase. The corpus was recorded at both normal and fast speech rates. Electromagnetic articulography (AG100, Carstens Medizinelektronik) was used to monitor movement of tongue (4 sensors mounted approx. 1 to 6 cm from the tongue tip), lower lip and jaw. Sensors on upper incisors and bridge of nose were used to compensate for head movement. Articulatory configurations were determined at the mid-point of the target vowels using a minimum-velocity criterion. Measures of contextual and T-T variability were derived in the following way: At each sensor position (on the tongue) a principal components analysis of the two-dimensional coordinates was performed: the variability measure was defined as the area in mm^2 of the 2-sigma ellipse oriented with its main axis along the first principal component of variation (cf. [1]). For contextual variability the ellipse area was simply calculated over all tokens of each vowel in turn. For token-to-token variability, the area was calculated separately for p-, t- and k-context, and then averaged over the three consonants.

RESULTS

The two different speech rates produced very similar variability patterns so we will present here only those obtained at the normal rate.

Contextual Variability

The 3 panels of Fig. 1 display the results first for each vowel averaged over sensor positions (top), and then for each sensor position individually with the vowels grouped into a front group (middle) and a back group (bottom). We will consider the vowels under three headings:

(i)The front high vowels /i:, ɪ, y:, ʏ, e:, ɛ/ (Fig.1, top and middle)

For these 3 pairs the tense member shows less variability than the lax at all sensor positions, and indeed the lowest variability of any vowels.

(ii) The pair /ø:, œ/ (Fig.1, top)

This is an anomalous pair (left out of the front group in the middle panel) as it has unusually high variability for the tense member compared to the other front vowels. This in turn means that no very clear answer emerges as to whether front rounded vowels show more lingual variability than the unrounded counterparts. (While these two vowel categories differ reliably in tongue position, it might have been hypothesized that tongue position in the rounded vowels is a subsidiary feature and thus liable to vary more).

(iii) The low and back vowels /ɑ:, a, o:, ɔ, u:, ʊ/ (Fig.1, top and bottom)

These vowels all show lowest variability at the rearmost sensor location. The variability at this position is somewhat higher than the minimum variability found for the front coils, but for the low back vowels the least variable sensor is probably rather further away from the actual vocal tract constriction than is the case for the front vowels. Thus while this group of vowels is clearly overall more variable than the front vowels (Fig.1, top), it would be hazardous to claim that the tongue is less tightly controlled at the site of maximum constriction.

The most striking feature of the results for this third group of vowels is the steady and extensive increase in variability from back to front sensor location, with the tense-lax distinction in variability becoming less clear-cut in the process.

Nonetheless, the higher variability for the lax vowels as a whole can be assumed to be a natural consequence of their shorter duration and the concomitant greater overlap with the adjacent consonantal articulations.

Token-to-Token Variability

Analogously to Fig.1, the results are summarized in the 3 panels of Fig.2. A similar grouping of the vowels also proves convenient. Regarding first the tense-lax distinction, the high front vowels show a pattern of slightly but consistently higher variability for the lax vowels (Fig.2, top and middle); the /ø:,œ/ pair (Fig.2, top) shows marginally more variability for the tense vowel; the low and back group (Fig.2, top and bottom) shows consistently more variability for the tense vowels, thus contrasting notably with the contextual variability results in Fig.1, especially at the more front sensor locations. The results thus provide neither a simple confirmation nor disconfirmation of the results in [3]. It is not immediately clear why the tense-lax distinction should be

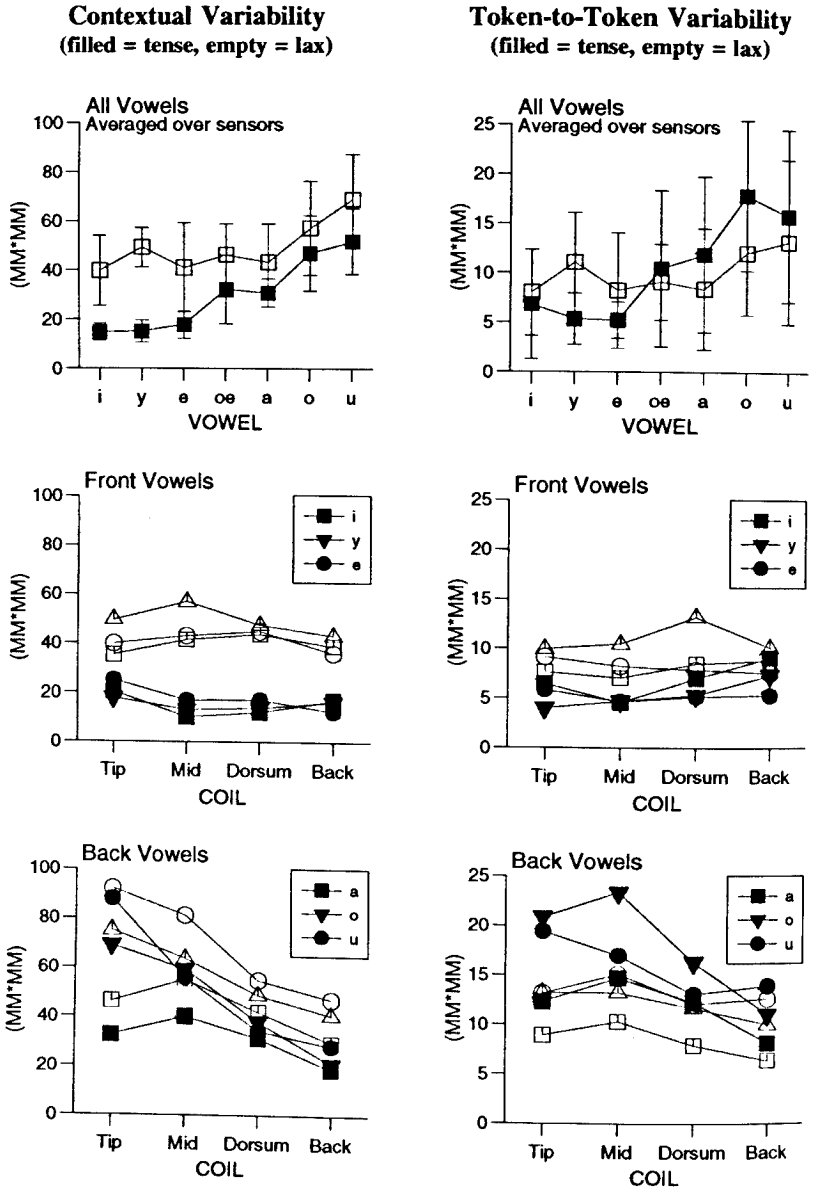


Fig.1 Contextual variability averaged over speakers and sensors (top, n=24) and over speakers (n=6) for the front and back vowel group (middle, bottom).

Fig.2 Results for token-to-token variability. Details as in Fig.1.

coupled with different T-T variability patterns for the front and the back vowels. One possible factor is the distance travelled by the tongue in the CVC movements. In t-context, for example, the tense variants of the low and back vowels have further to travel, while the reverse is the case for the high front vowels. However, the precise extent to which T-T variability may be explainable by articulator displacement remains to be explored in detail.

In considering whether high front vowels may be able to profit from the proximity to the hard palate to achieve a relatively invariant configuration, the first point is that we clearly cannot confirm Bohn et al.'s contrary finding of *more* variability on the high vowels, at least not for the front vowels (cf. Fig.2, top). On the other hand, the simple alternative conclusion of increasing variability with decreasing tongue height is also not completely warranted. If we inspect the values in Fig.2 (top) after ordering tense and lax vowels separately into two rows with respect to tongue height (i.e /iː, yː, eː, øː, ɑː/ and /ɪ, ʏ, ɛ, œ, a/) then no dependency of variability on tongue height is found for the lax vowels. For the tense vowels we do indeed find a difference between the 3 highest and the 2 lowest, but no obvious gradual increase from high to low. In fact, the question of high vs. low vowels may be wrongly posed. It may be more profitable to point to the one major parallel between contextual and T-T variability, and to reformulate the question in terms of a distinction between a palatal group and a velar-pharyngeal group.

CONCLUSIONS

As just mentioned, analysis of the patterns of contextual and T-T variability suggested the existence of two main vowel groups. This distinction is undoubtedly partly a biomechanical one: palatal constrictions constrain the whole of the tongue, whereas constrictions further back leave the mobile anterior tongue much freedom to vary. However, additional factors could underly the variability in back vowels: firstly, the acoustic consequences of variability remote from the main constriction may be rather slight, particularly when coupled, secondly, with the relatively uncrowded back vowel region in German. Assessment of these factors awaits the completion of the acoustic counterpart to this investigation.

The second conclusion is that the slightly untidy results found for the comparison of tense vs. lax vowels and rounded vs. unrounded vowels underline the importance of investigating sound systems as nearly as possible in their entirety, as otherwise the danger of spuriously clear-cut results may be considerable.

ACKNOWLEDGEMENTS

Work supported by German Research Council grant Ti 69/29-2

REFERENCES

[1] Perkell, J.S. (1990), "Testing theories of speech production: implications of some detailed analyses of variable articulatory data", In: W.J. Hardcastle & A. Marchal (eds.), *Speech Production and Speech Modelling*, Dordrecht: Kluwer Academic Publishers, pp. 263-289.

[2] Hoole, P., Mooshammer, C. & Tillmann, H.G. (1994), "Kinematic analysis of vowel production in German", *Proc. ICSLP 94*, Yokohama, 1:53-56.

[3] Bohn, O.; Flege, J.E.; Dagenais, P.A. & Fletcher, S.G. (1991), "Differenzierung und Variabilität der Zungenpositionen bei der Artikulation deutscher Vokale", *Zeitschrift für Dialektologie und Linguistik (Beihefte 72)*, pp. 1-26.

A GESTURAL PRODUCTION MODEL BUILT BY ACOUSTIC-ARTICULATORY INVERSION OF FORMANT TRAJECTORIES.

Paul Jospa, Martine George & Alain Soquet.
Inst. des Langues Vivantes et de Phonétique, Université Libre de Bruxelles,
50 av. F.D. Roosevelt, 1050 Bruxelles, Belgium.

ABSTRACT

The proposed model defines the articulatory gesture as the result of a competition between coactivated, invariant articulatory targets. The consonant gesture is defined as a deformation of a vowel gestural continuum. The model is governed by a gestural score which generates the involved target activation functions. Currently, targets are associated to isolated phonemes. The identification of structural parameters is carried out using an efficient acoustic-articulatory inversion technique.

THE GESTURAL MODEL

The proposed model defines the articulatory gesture as the result of a competition between coactivated, pseudo invariant articulatory targets (see figures 1 and 2). The activation levels of these targets change in time. The consonant gesture is defined as a deformation of a vowel gestural continuum.

The model is governed by a gestural score which generates the involved target activation functions. The temporal evolution of the articulatory profile (the area function or some simple transformation of it) is generated by a competition law between coactivated targets. The target attraction level is locally weighted by acoustic sensitivities of articulatory parameters characterising the target. Currently, the targets are typical (extremal) articulatory profiles associated with isolated phonemes (phonetic targets).

The model presents two main functional levels (analogous to those of the Task Dynamics model [1]). At the top level, activation functions are generated from a gestural score [2]. The temporal evolution of the articulatory profile is generated at the bottom level. The articulatory configuration is expressed in terms of the area of vocal tract sections (the area function); then it can be further expressed in terms of parameters of a given articulatory model. We have adopted here the Distinctive Region Model (DRM) [3] to describe the area function because: i) the DRM exhibits a large monotonicity in the acoustic-articulatory link around the neutral configuration, ii) the DRM is able to generate a large acoustic space (in terms of the three first formant frequencies). We plan to use other articulatory models --especially Maeda's model which exhibits also a large monotonicity in the acoustic-articulatory link-- for studies more closely connected with the articulatory level.

The gestural score level

Activation functions (one for each target) are generated at the gestural score level. They are either generated by means of autonomous (non linear) dynamic regimes, or defined by means of given time functions (in analytical or tabulated form). Currently, activation functions are simple time functions (sigmoid-like functions), or extracted from speech signal segments by acoustic-articulatory inversion (see fig.1).

The articulatory level

a) The vocalic V1-V2 transition rule

In the framework of the DRM model, articulatory target T_v^* of a vowel v is

defined by means of:

T_v^*: $\left\{A_{k,v}^*\right\}$ (for: $k = 1,...,8$), and L_v^*,

where $A_{k,v}^*$ is the area (or some simple transform of it like a square root) of region k of target v (see figure 2.c), and L_v^* is the tract length of target v. For a vocalic transition V1-V2, we use three targets: T_{v1}^*, T_{v2}^* and the « neutral » (schwa like) target T_Φ^* which is associated with the tract rest configuration. The V1-V2 competition rule is given by:

$$A_k(\text{v1v2};\, t) = \frac{\sum_{v\in\{v1,v2,\Phi\}} \alpha_v(t) p_{k,v} A_{k,v}^*}{\sum_{v\in\{v1,v2,\Phi\}} \alpha_v(t) p_{k,v}} \quad (1)$$

with: $0 \le \alpha_v(t) \le 1 \;\; \forall$ v and $\forall\, t$.

$A_k(\text{v1v2};\, t)$ is the area (or a transform of it) of region k at time t for the transition v1v2, $\alpha_v(t)$ is the activation function of the v target, and $p_{k,v}$ is a target region specific weight, which is a function of the acoustic sensitivity [4] of region k of target v; we have chosen:

$$p_{k,v} = \sum_{n=1}^{3} q_n \left(\frac{\partial f_n}{\partial A_{k,v}^*} \right)^2 ,$$

where $\{f_n\}$ (n=1,2,3) are the first three formant frequencies, and q_n are some convenient weights.

b) The V1-C-V2 transition rule

Let be $\{C\}$ the set of tract regions (consonant constriction regions) affected by consonant gesture c. Let $\left\{A_{l,c}^*\right\}\; l \in \{C\}$ be the consonant constric-tion target, and $\alpha_c(t)$ the consonant gesture activation function. The V1-C-V2 transition law is given by:

$$\begin{aligned} A_k(\text{v1cv2};\, t) &= A_k(\text{v1v2};\, t) \quad \text{if } k \notin \{C\} \\ A_k(\text{v1cv2};\, t) &= A_k(\text{v1v2};\, t) + \\ &\quad \alpha_c(t)\left(A_{k,c}^* - A_k(\text{v1v2};\, t)\right) \quad \text{if } k \in \{C\} \end{aligned} \quad (2)$$

with: $\left(0 \le \alpha_c(t) \le 1\right)$.

The consonant gesture is thus defined as a deformation of a vocalic continuum.

MODEL PARAMETERS IDENTIFICATION

The identification of the model structure parameters (including the activation functions) proceeds from formant trajectory data and from some a priori knowledge of the acoustic-articulatory link, rather than from articulatory data which are difficult to obtain. For this purpose, an efficient acoustic-articulatory inversion technique has been developed which is capable of gaining a priori knowledge from a limited but well designed set of acoustic-articulatory links [4]. This technique consists uses a neural controller [5] and a variational method to compute fastly the acoustic-articulatory link [6]. The identification process is carried out by steps. Firstly, articulatory targets are identified, and normal modes and acoustic sensitivities of the chosen target configurations are computed. Then, the temporal evolution of the activation functions for typical V1-V2 and V1-C-V2 transitions is extracted using acoustic-articulatory inversion of formant trajectories. This occurs in the framework of the adopted gestural competition model. As a last step, not currently implemented, a sigmoidal model (which can be expressed as a non-linear autonomous dynamic model) is

adjusted to the resulting activation functions. This identification procedure enables us to adapt our gestural production model to speech signal segments. As a result, it becomes a speech signal (formant trajectories) analyser in terms of activation functions, gestural scores (activation parameters), or parameters of the sigmoidal model .

CONCLUSION.

We propose a simple gestural production model in terms of competitions between « extremal » articulatory (phonetic) targets. We have described a procedure to identify the structural parameters of this model by means of an acoustic-articulatory inversion technique applied to selected V1-V2 and V1-C-V2 acoustic logatomes. By this way, we are able to built the model firmly on an acoustic basis. Moreover, this identification procedure enables us to use our model not only for articulatory movement synthesis, but also for analysis of formant trajectories in terms of activation functions, gestural scores, or other parameters of the gestural model.

ACKNOWLEDGEMENT

We wish to thank Marco Saerens for his numerous and valuable comments. This work was partially supported by grant SC1-CT92-0786, and by the ARC 92/97-160 project of the Communauté Française de Belgique.

REFERENCES

[1] E. Saltzman (1986): "Task dynamic coordination of the speech articulators: a preliminary model". In H. Heuer and C. Fromm (eds.) *Generation and modulation of action patterns.* Springer-Verlag, Berlin, pp.129-144.

[2] C. Browman, L. Goldstein (1992): "Articulatory phonology: an overview". Phonetica 49, pp.155-180.

[3] Mrayati M., Carré, R. Guerin B. (1988): "Distinctive regions and modes: A new theory of speech production". Speech Comm. 7, 257-286.

[4] A.Soquet, P. Jospa, (1994): "The acoustic-articulatory mapping and the variational method", ICSLP-94 Proc. -2 pp. 595-598.

[5] Saerens M & Soquet A. (1991): "Neural Controller Based on Back-Propagation Algorithm". IEE Proc.-F, 138 (1), pp. 55-62.

[6] P. Jospa, A. Soquet, M. Saerens (1995): "Variational formulation of the acoustic-articulatory link and the inverse mapping by means of a neural network", in: C. Sorin & al. (eds.).: *Levels in Speech Comm.: Relations and Interactions.* Elsevier. pp. 103-113.

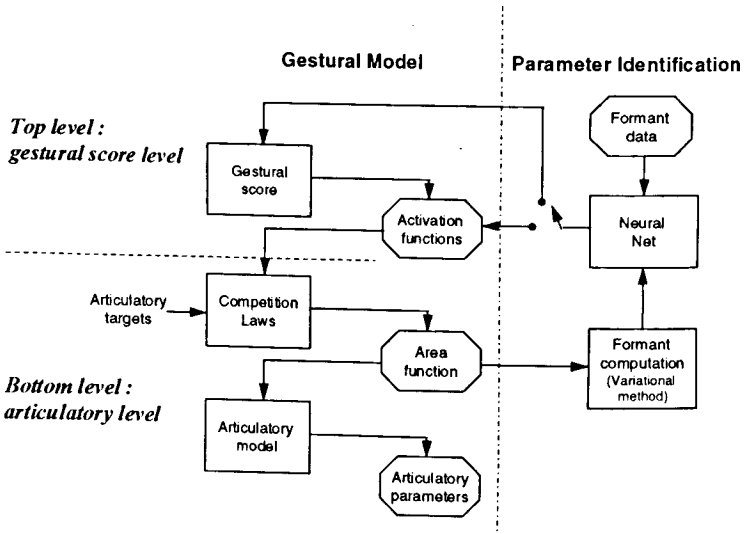


Figure 1. The gestural model embedded in the acoustic-articulatory inversion system.

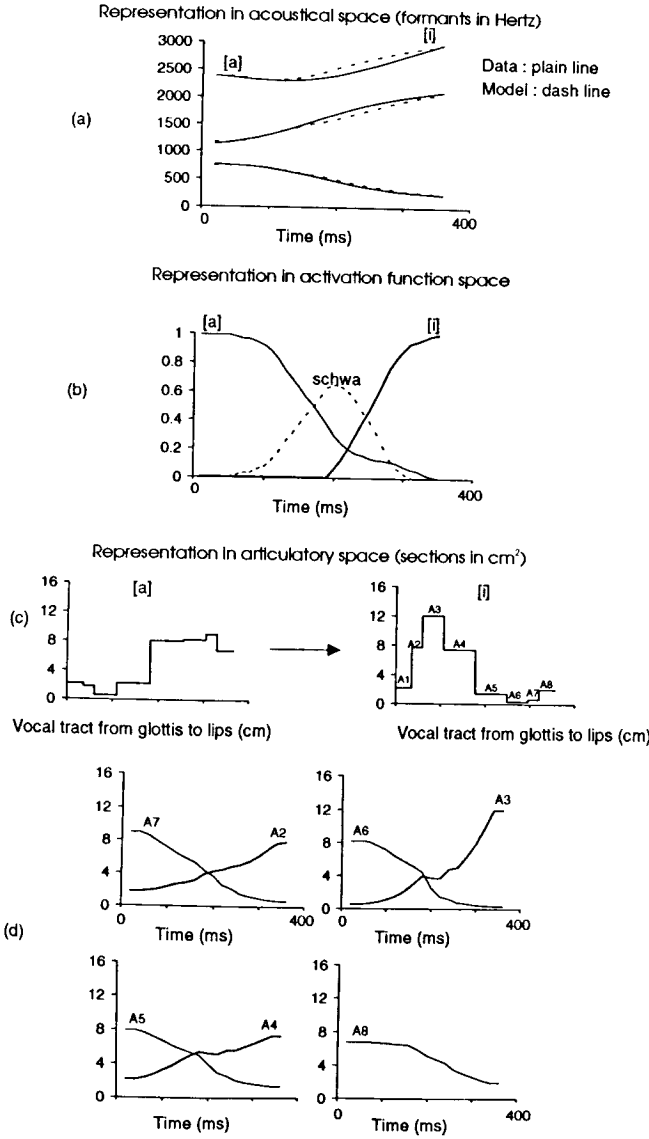


Figure 2: Model outputs for a transition /a-i/.
a) Formants generated (----) (after acoustic-articulatory inversion of formant data (——)).
b) Activation functions.
c) Area functions at the begining and end of the transition (articulatory targets).
d) Articulatory parameters (tube region areas of the Distinctive Region Model).

NEW TECHNIQUES OF VOCAL TRACT MODELING FOR ARTICULATORY SPEECH SYNTHESIS

Matti Karjalainen and Vesa Välimäki
Helsinki University of Technology
Laboratory of Acoustics and Audio Signal Processing
Otakaari 5 A, FIN-02150 Espoo
Finland

ABSTRACT

The theory of fractional delay waveguide filters (FDWF) is presented that is proved to solve the fundamental problem of the fixed sample-position lattice found in traditional discrete-time simulation of transmission-line type systems. Thus the approach is well suited to vocal tract models, in particular to articulatory modeling and speech synthesis, where a simple relation between model parameters and articulator positions is desired.

INTRODUCTION

In the days before digital computers the human vocal tract was modeled by analog circuits and physical equivalents of the real system. Transmission-line type models were considered as being continuous, at least in time but generally in space as well. The corresponding analytic mathematical models of speech production, e.g. [1], reflect the same non-discrete character inherent in many macroscopic physical systems. It was natural to simulate the vocal tract by a chain of tube sections (e.g. two and three-tube models) that had relatively direct correspondence to the positions of the articulators.

Computers and digital signal processing changed the picture by providing very accurate and flexible numeric methods to deal with vocal tract modeling [2]. One specific characteristic in digital simulation of continuous-time systems, however, was not flexible at all. Unit delay, the interval between subsequent samples, dictates much of the behavior in discrete-time systems at high frequencies. Signals must be bandlimited and the high-frequency characteristics of analog systems can in a general case be only approximated.

This limitation is particularly prominent in transmission-line models of the vocal tract. When the sampling frequency is fixed, the physical positioning of known sample points in space is fixed as well. Even a detailed adjustment of the vocal tract length to this digital grid has been somewhat difficult and discussed only in relatively few papers. Earlier techniques for this partial solution of the problem are found in [3], [4], and [5].

We may well accept the discrete-time nature (since speech and hearing are bandlimited) but we do not have to accept the discrete-space characteristics in transmission-line modeling. What we need is fractional delays, mathematically equivalent to ideal bandlimited interpolation. This turns out to be possible to approximate and implement using digital filters. Allowing somewhat increased computational costs one can have a whole bunch of discrete-time models that are virtually continuous in space.

We have developed some basic principles and building blocks for modeling transmission-line type systems, such as acoustic tubes of arbitrary and varying shapes, including the human vocal tract [6]–[10]. In this paper we present how to use digital signal processing to build tube sections of varying lengths and Kelly–Lochbaum type tube models with variable junction positions. It will be shown that not only cylindrical but also conical tube sections can be used, thus leading to more natural shape approximations. The only addition that conical tubes introduce to the KL scattering junction is that a simple reflection filter is needed instead of a real-valued reflection coefficient.

The new fractional delay filter structures (we call them Fractional Delay Waveguide Filters, FDWF) are natural candidates when building vocal tract models for articulatory speech synthesis since moving articulators may be associated directly to subsections of the system. The increased computational cost can be compensated by the ever increasing performance of modern signal processors.

GENERALIZED KELLY–LOCHBAUM MODELING

We start by considering a generalized version of the Kelly–Lochbaum (KL) vocal tract model [11]. Figure 1 depicts a one-multiplier KL scattering junction where the traveling wave components are summed, multiplied by the reflection coefficient r, and injected back into the delay lines. This is a traditional formulation and easily implemented as far as the junctions are aligned with the natural sampling positions.

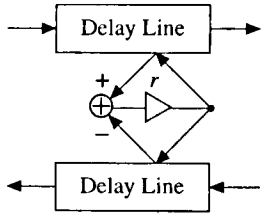


Fig. 1. A one-multiplier Kelly–Lochbaum scattering junction.

Let us consider a case where the scattering junction is located in an arbitrary position along the delay lines. Taking a signal value out of a delay line between sample positions is equivalent to *interpolation*. An ideal bandlimited interpolator is an infinitely long FIR filter with coefficients according to the sinc function [10]. (If the junction is precisely at a sampling point then only that single coefficient remains non-zero.)

The insertion of a signal back into the delay lines, when the insertion point is between the sampling points, is called *deinterpolation* [6]. As an FIR filter it is the transpose of the interpolation filter. The value to be inserted is multiplied by each filter coefficient, these results are added to the sample values in the delay line, and the sums are written back into the sample positions. With ideal interpolators and deinterpolators this precisely implements a bandlimited KL junction in any fractional position.

The two-port KL junction of Fig. 1 can be generalized to a three-port junction that is applicable to the modeling of the nasal tract branch [10].

FRACTIONAL DELAYS

In practice we cannot realize ideal interpolated junctions but have to approximate the infinite series of sinc coefficients by finite order digital filters. Two specific cases are of special interest: (a) Lagrange interpolators of FIR type [3], [5] and (b) allpass filters with maximally flat (at zero frequency) phase delay. Allpass filters have an ideal magnitude response and some other advantages [10]. However, since FIR interpolators are straightforward and conceptually more intuitive, we will consired only them below.

Figure 2 shows the implementation of third-order FIR filters for (2a) an interpolator and (2b) a deinterpolator.

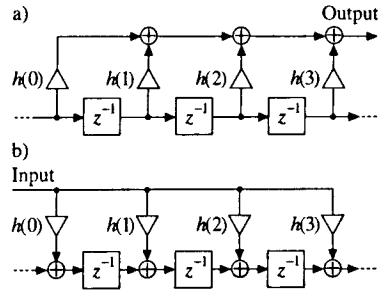


Fig. 2. Third-order FIR filter implementation of (a) an interpolator and (b) a deinterpolator.

When the position of an interpolated junction is moved, the locations of the FIR taps on the delay lines and the values of the filter coefficients must be updated. Notice that there may be any number of fractional delay KL junctions attached on a pair of delay lines. The tap regions of the junctions can also overlap as far as the junction operations are done on a 'read all, compute the scatterings, write all' basis [6].

The filter structure including the delay lines and any number of interpolated

scattering junctions is called the *fractional delay waveguide filter* (FDWF). A special case is a terminating junction where only one interpolation (out) and one deinterpolation (in) is needed. The interpolator of Fig. 2a as such finds many applications where a fractional delay is needed.

The application of finite order interpolators introduces appoximation errors in the waveguide filters. With odd-order Lagrange interpolators the maximum error occurs when the junction is in the middle of two sampling points. An error analysis [10, pp. 93-95] shows that third-order Lagrange interpolators yield good results in speech synthesis up to about 5 kHz when the sampling rate is 22 kHz.

CONICAL TUBE MODEL

The approximation of the vocal tract by cylindrical sections only has been another fundamental limitation of traditional transmission-line models. We have generalized the KL model to allow for conical sections as well (see Fig. 3), since this makes a better match to typical vocal tract shapes that are continuous and relatively smooth functions [9], [10].

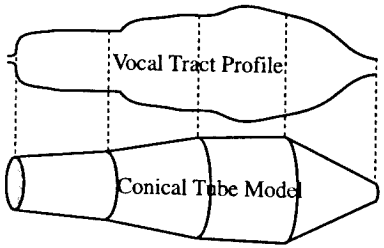


Fig. 3. Conical section approximation of a vocal tract profile.

The difference between cylindrical and conical sections is that instead of plane waves the conical sections carry spherical waves. The derivation of pressure wave scattering at a junction between conical tube sections results in the traditional KL formulation of Fig. 1 except that the reflection coefficient r is replaced by a reflection filter $R(z)$ and the signals from the upper and lower delay line are added (instead of subtraction). The filter $R(z)$ is a first-order IIR filter that is computationally efficient and does not essentially add to the complexity of implementation.

A natural step towards further generality is to combine the freely movable fractional delay ports and the conical sections [9].

ARTICULATORY N-TUBE SYNTHESIS MODEL

Vocal tract modeling using the FDWF techniques is a natural candidate for articulatory speech synthesis [7]. The traditional thinking of the tract as a decomposition of sections related to the articulators [1] is well supported since the section lengths and cross-sectional areas are freely adjustable. Figure 4 shows the case of a three-tube model and the related control parameters.

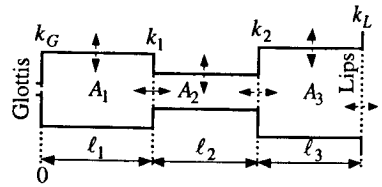


Fig. 4. A three-tube model of the vocal tract. Arrows indicate the variable parts.

The n-tube model parameters have a relatively straightforward mapping from articulatory parameters. In [7] we have proposed a case for a set of five articulatory parameters and a five-tube vocal tract model.

One inherent problem in speech synthesis with models of variable length sections is that there is no simple method known for the analysis of the parameter values from speech signals. Thus iterative or nonlinear estimation techniques must be used.

IMPLEMENTATION ISSUES

We have experimented with the new principles of vocal tract modeling in order to implement real-time synthesis on the TMS320C30 DSP processor.

The computational complexity of the fractional delay waveguide models is relatively high due to the need of interpolation and deinterpolation as well as a relatively high sampling frequency. On the other hand, less tube sections are needed to match a vocal tract profile than with the original KL model.

We have implemented four and five-tube models on the TMS320C30 (33 MHz) floating-point signal processor. Updating of the model parameters has been carried out by a host computer (Apple Macintosh). There is a graphical user interface where the user can move the sections, boundaries, or related articulators by a mouse. We have noticed that the update rate of the parameters should be relatively high (at least every 15 samples when the sampling rate is 22 kHz) in order to avoid audible transient problems. From a theoretical point of view there is need for an analysis of transient-free section length and port position controls in a similar way as was done in [12] for fixed junction models.

So far we have synthesized primarily vowels as well as nasals including a nasal tract. The mouse-controlled vowel synthesizer has been found a useful device for demonstrations and experiments in articulatory phonetics. A full-scale synthesizer with all phoneme classes remains to be developed.

SUMMARY

A non-mathematical introduction to fractional delay waveguide filters has been given and the theory of vocal tract modeling based on them has been presented. The approach allows for a flexible method to implement articulatory speech synthesis using variable-length tube sections. Modeling of the vocal tract by conical tube sections is introduced as another major extension to the theory. A real-time synthesizer with manual tract control has been demonstrated.

REFERENCES

[1] G. Fant, *Acoustic Theory of Speech Production*. Mouton, The Hague, 1960.

[2] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Prentice-Hall, Englewood Cliffs, 1978.

[3] H. W. Strube, "Sampled-data representation of a nonuniform lossless tube of continuously variable length," *JASA*, vol. 57, no. 1, pp. 256–257, Jan. 1975.

[4] H. Y. Wu, P. Badin, Y. M. Cheng, and B. Guerin, "Continuous variation of the vocal tract length in a Kelly–Lochbaum type speech production model," in *Proc. XIth ICPhS*, pp. 340–343, Tallinn, Estonia, Aug. 1987.

[5] U. K. Laine, "Digital modelling of a variable-length acoustic tube," in *Proc. 1988 Nordic Acoustical Meeting*, pp. 165–168, Tampere, Finland, June 1988.

[6] V. Välimäki, M. Karjalainen, and T. I. Laakso, "Fractional delay digital filters," in *Proc. IEEE Int. Symp. Circuits and Systems (ISCAS'93)*, Chicago, IL, vol. 1, pp. 355–358, May 3–6, 1993.

[7] V. Välimäki, M. Karjalainen, and T. Kuisma, "Articulatory control of a vocal tract model based on fractional delay waveguide filters," in *Proc. IEEE Int. Symp. Speech, Image Processing and Neural Networks (ISSIPNN'94)*, Hong Kong, vol. 2, pp. 585–588, April 13–16, 1994.

[8] V. Välimäki, M. Karjalainen, and T. Kuisma, "Articulatory speech synthesis based on fractional delay waveguide filters," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP '94)*, Adelaide, Australia, vol. 1, pp. 571–574, April 19–22, 1994.

[9] V. Välimäki and M. Karjalainen, "Improving the Kelly–Lochbaum vocal tract model using conical tube sections and fractional delay filtering techniques," in *Proc. 1994 Int. Conf. Spoken Language Processing (ICSLP'94)*, vol. 2, pp. 615–618, Yokohama, Japan, Sept. 18–22, 1994.

[10] V. Välimäki, *Fractional Delay Waveguide Modeling of Acoustic Tubes*. Report 34, Helsinki Univ. of Tech., Lab. of Acoustics and Audio Signal Processing, Espoo, Finland, 1994.

[11] J. L. Kelly and C. C. Lochbaum, "Speech synthesis," in *Proc. Fourth Int. Congr. Acoustics*, paper G42, pp. 1–4, Copenhagen, Denmark, Sept. 1962.

[12] J. Liljencrants, *Speech Synthesis with a Reflection-Type Line Analog*. Doctoral thesis. Royal Inst. of Tech., Dept. of Speech Communication and Music Acoustics, Stockholm, Sweden, 1985.

AN ACOUSTIC ANALYSIS OF THE RETROFLEX FLAP

Knut Kvale

Telenor Research
P.O.Box 83
N-2007 Kjeller, NORWAY
E-mail: knut.kvale@tf.telenor.no

Arne Kjell Foldvik

Dept. of Linguistics
University of Trondheim
N-7055 Dragvoll, NORWAY
E-mail: arne.foldvik@avh.unit.no

ABSTRACT

This paper investigates the acoustics of the retroflex alveolar flapped stop [ɽ] and demonstrates that different phonetic contexts systematically affect its realization.

1. INTRODUCTION

Retroflex flap [ɽ] is found in numerous Indian languages, in a number of African languages and in some languages in Australia and Mexico. In Europe retroflex flap occurs in Albanian, North-Western Italian and among speakers in large areas of Norway and Sweden.

Taps and flaps are similar in that the active articulator moves more rapidly than in their non-tapped or non-flapped stop-counterparts. The main difference between tapping and flapping is the direction of the tongue movement after the brief closure.

In Norwegian and possibly also in Swedish retroflex flap was considered sub-standard or non-standard, but with the present trend of more general acceptance of regional and social dialects the status of retroflex flap has risen. The previous low status of [ɽ] possibly explains why little or no research has been done on it. However, for speech technology purposes acoustic studies of retroflex flap are important both to produce natural sounding synthetic speech and to succeed in automatic speech recognition.

Typically the Norwegian [ɽ] is pronounced as a *voiced or partly voiced retroflex alveolar flapped stop*, but we are aware that in some languages [ɽ] may be realized as a retroflex alveolar *lateral* flapped stop. For a *phonemic* discussion of [ɽ], see [1].

2. CONTEXTUAL EFFECTS

In this section we first investigate the realisation of [ɽ] in *logatomes* (i.e. artificial words with a fixed syllable structure), containing diphones for Norwegian text-to-speech synthesis. The logatomes were read singly with an average articulation rate of 100 $^{ms}/_{phoneme}$.

In Norwegian the retroflex flap occurs in pre-, inter-, and post-vocalic position in some contexts, but not word initially nor after the front vowels [i], [e] and [y]. In the present logatome material the retroflex flap was also realised in some of these non-occurring positions.

2.1 Intervocalic position

Figure 1 shows the waveform and the broad band spectrogram of the Norwegian words [ɔ: ɽ ə] and [ɔ: ɾ ə], illustrating some similarities and differences between the retroflex flap and the apical alveolar tap in *intervocalic position*. In both words F_1 in the vowels is relatively constant, whereas the low F_2 of the back vowel [ɔ:] is shifted upwards due to the succeeding alveolar closure. (Alveolar phones are characterized by high F_2 [2]). The closure phase of the retroflex flap is acoustically similar to the corresponding part of the apical alveolar tap [3].

The differences between the tap and the flap are manifested in the F_3 and F_4 of the neighbouring vowels. In the *tap* realisation, the formant trajectories can be smoothly interpolated from the vowel preceding the tap to the following vowel, suggesting that the tongue tip returns to the onset position after the closure. In the *flap* realisation, the tongue makes a brief (alveolar) closure in the passing to another position, resulting in an abrupt change in F_3 across the flap closure. In addition, F_4 before the flap closure turns down, whereas before the tap closure F_4 turns slightly upwards.

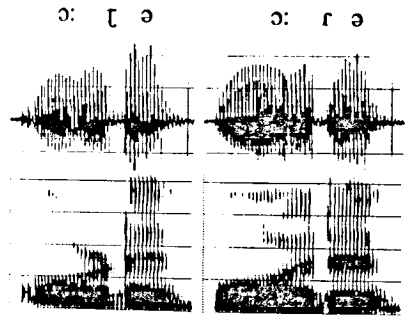


Figure 1 *Waveform and broad band spectrogram of [ɔ: ɽ ə] and [ɔ: ɾ ə] pronounced by a male speaker. Neighbouring vertical lines are 50 ms apart. In the spectrogram the frequency difference between neighbouring horizontal lines is 1 kHz.*

Generally, in isolated words the retroflex flap lowers F_3 and F_4 of the neighbouring vowels, especially in the preceding vowel. For front vowels (with high F_2) the second formant is lowered but is raised for back vowels (with low F_2). The flap has no noticeable effect on the F_1 of the neighbouring vowels.

For the retroflex flap, the curling up of the tongue tip before the forward flap movement takes place shows up in the spectrogram as a fairly long formant transition before the voicebar portion. Typically the whole preceding vowel was diphthongized. Although these formant transitions are important for the perception of [ɽ], they are not included in the [ɽ] segment. (This segmentation approach is similar to the one used for plosives which are defined as starting when the closure begins, and not when the formants of the preceding vowel change [4]).

2.2 Succeeding a consonant

With [ɽ] in a C_V context, an interval of voicing and formant structure is seen in the spectrogram *before* the apex touches the alveolar ridge.

When a bilabial stop or nasal precedes the retroflex flap, the tongue tip is free to curl up and back during the bilabial closure phase. When the built up pressure for the stop is released the tongue tip can flap forward quickly on the egressive airflow. In these consonant clusters the period of voicing between the consonant and the brief flap closure was relatively short (i.e. the duration of the epenthetic schwa, [ə], was about 30 ms after [p] and 50 ms after [b] and [m]).

Also with labiodental fricatives preceding the retroflex flap, the period of voicing between the fricative and the flap closure was short, resulting in an [ə] of about 40 ms duration.

When [ɽ] was preceded by alveolar, postalveolar or retroflex consonants or the apical alveolar tap, the duration of the schwa before the flap closure became noticeable longer; from about 80 ms after [s], up to 130 ms after [ɳ]. When articulating these consonants the tongue tip is engaged in making a closure or a constriction and cannot be curled backwards for the flap till the first closure or constriction is released. These combinations are therefore articulatorily inconvenient and may explain why they do not occur in Norwegian.

When retroflex flap is articulated after the alveolar stops, especially [d] and [n], the [ə] shows up in the spectrogram with a

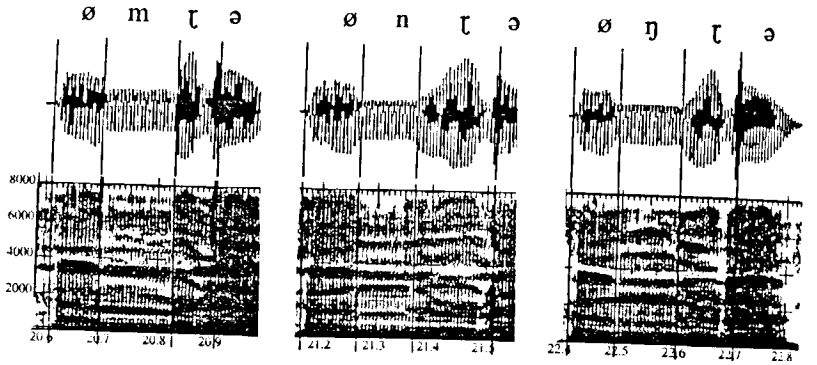


Figure 2 *Bilabial, alveolar and velar nasals preceding the retroflex flap.*

characteristic shift in F_3 from about 2500Hz immediately after the alveolar stop release to about 1700 Hz, (or close to F_2), before the flap closure.

Velar stops involve the dorsal part of the tongue, allowing the tongue tip to start curling backwards during the velar closure phase. Thus, the duration of schwa was only about 50 ms after [k], but longer for [ŋ] (70 ms) and particularly for [g] (80 ms).

Typically, the waveform envelope of [ə] decreases evenly towards the [ɽ] closure, whereas the closure release gives an abrupt change in the waveform, as exemplified in *figure 2*. Notice also the characteristic jumps in F_3 and F_4 from the [ə] before the [ɽ] closure to the following neutral vowel.

2.3 Preceding a consonant

With [ɽ] in a V_C context, the tongue has finished its flapping movement and is free to start its movement towards the consonantal target. Usually the voicing is not turned off after the flap closure release, so a schwa-type vowel appears *after* the flap closure. However, the intensity and duration of this [ə] is much less than when the consonant precedes the retroflex flap. Thus, for [ɽ] in this phonetic environment the duration of [ə] varied from about 40 ms before [b] and [d]) to 70 ms (before [ʃ]). See *figure 3*.

When an apical alveolar tap succeeds the retroflex flap, the tongue tip moves up to produce the tap after the [ɽ] closure release. This take time, and the epenthetic schwa became rather long (about 80 ms). Since this is articulatorily inconvenient, the [ɽ] + [ɾ] combination only occurs across morpheme boundaries in Norwegian.

The F_3 and F_4 usually turn down towards the flap closure, yielding a characteristic jump in these formants to the following epenthetic schwa.

2.4 Continuous speech

In continuous speech the contextual effects between the retroflex flap and the neighbouring sounds are in principle similar to those for the logatomes. However, intervocalically the closure phase of [ɽ] may be very brief or even non-existing. Thus, in these cases only the change in formant-transitions in the neighbouring vowels convey the perceptual cues for [ɽ].

When [ɽ] is preceded or succeeded by a consonant, an extra schwa may appear, e.g. the word "flaske" (=bottle) was realised with:

- a clearly articulated, short epenthetic [ə] and a distinct [ɽ] closure phase,

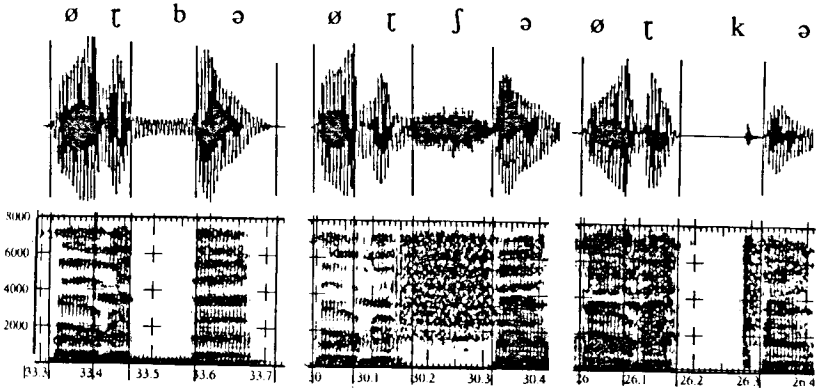


Figure 3 *Bilabial, post-alveolar and velar consonants succeeding the retroflex flap.*

- a very weak [ə] and no [ɽ] closure phase,
- no extra [ə] and partly or totally devoiced [ɽ] with no closure phase.

In the latter case the F_3 of the following back, open vowel was lowered.

The differences between the logatome and continuous speech material were *non-systematic* in that the same person might pronounce words like "flaske" in all the three ways described.

3. CONCLUSIONS

The contextual effects of the retroflex flap can be grouped into three main phonetic environments:

(*i*) [ɽ] in V_V context has an lowering effect on F_3 and F_4; especially in the preceding vowel,

(*ii*) [ɽ] in C_V context is realised with an epenthetic schwa type vowel *before* the flap closure,

(*iii*) [ɽ] in V_C context is realised with a schwa *after* the flap closure.

In (*iii*) the epenthetic schwa is significantly weaker and shorter than in (*ii*).

In continuous speech we found the same contextual effects of [ɽ] as in the logatomes, but with reduction effects: partly devoiced [ɽ], weaker and shorter [ə] and shorter or non-detectable closure phase.

Since the closure phase of the retroflex flap is acoustically very similar to that of the apical alveolar tap [ɾ], we suggest applying the same segmentation approach to these sounds. That is, in linear, phonemic segmentation the epenthetic schwa should be included in the [ɽ]-segment.

We believe that in addition to acquiring more acoustic-phonetic knowledge, such acoustic analyses of speech sounds are important for automatic speech recognition of different dialects and sociolects and for more natural sounding Norwegian text-to-speech synthesis.

REFERENCES

[1] Sandøy, H. (1992), *Norsk dialekt-kunnskap*, Novus forlag.

[2] Zue, V.W. (1989), *Speech Spectrogram Reading - An Acoustic Study of English Words and Sentences*, Course at the University of Edinburgh.

[3] Kvale, K., Foldvik, A.K., (1992), "The multifarious r-sound", Proc. International Conference on Spoken Language Processing, pp. 1259-1262.

[4] Kvale, K. (1993), *Segmentation and Labelling of Speech*, Doctoral thesis, Norwegian Institute of Technology.

THE GROOVE PRODUCTION OF SWEDISH SIBILANTS - AN EPG ANALYSIS

Per Lindblad and Sture Lundqvist
Dept of Linguistics, Göteborg, Sweden
& Dept of Prosthetic Dentistry, Göteborg, Sweden

ABSTRACT

The sibilantic groove was EPG analysed in Swedish /s/, /ɕ/ and [ʂ]. Groove position and width of each sibilant varied between but not within speakers. No correlation was found between groove width and sibilantic identity. /s/ was produced clearly frontmost, but /ɕ/ had the same groove front opening position as [ʂ] in phrases. The articulatory place of /ɕ/ and [ʂ] is not their primary distinguishing articulatory feature. The description of sibilants must attend more to the size of the groove anterior cavity.

INTRODUCTION

Acoustic modelling of fricative production has advanced in the last decade [1, 2, 3]. For further development, the need for empirical production data is great [4]. One important practical application of this growing knowledge is to give a scientific phonetic base to dental prosthesis constructing [5]. Especially [s] is often deteriorated by prostheses [5].

With three phonemically contrasted front tongue sibilants - /s/, /ɕ/ and [ʂ] - Swedish is especially suitable for an investigation with the aim to further the development of the fricative modelling work. These sounds are acoustically and perceptually closely related. /ɕ/ is intermediate perceptually in brightness, and acoustically in spectral energy distribution [6]. Detailed articulatory descriptions of these sounds are given in [6, 7]. [ʂ] is a common allophone of the Swedish /ʃ/ phoneme, which also has a common, non-sibilantic variant, [ɧ].

The best way available to analyse the sibilantic groove - one of the two crucial articulatory features in sibilants - is by electropalatography (EPG) [8]. This method has been used in several studies of sibilantic production e g in English [8, 9]. Swedish sibilants have been treated in two EPG investigations: of /s/, based on ten speakers [10], and of a large number of consonants, including /s/, /ɕ/ and [ʂ] - for only one speaker, however [11].

The other crucial articulatory feature in sibilants is the incisors, being hit by an air jet emanating from the groove. About this phenomenon, neither EPG nor any other existing method gives direct information. However, the combination of EPG data with jaw movement and acoustic information, and dental casts of the upper and lower jaws, will be able to contribute to the advancement of the understanding of the role of this feature. We have procured this combination of data and have short-time plans to work with it, with due attention to important new theoretical aspects in [1] of alveolar ridge and tooth contribution to the sibilantic source generation.

METHOD AND MATERIAL

Our equipment was of the Reading EPG type. For a thorough description, see [12]. In short terms, the speaker wears a thin palate, extending from the upper teeth back to the velum. In this palate, 62 electrodes are placed in a regular pattern. In the alveolar region, where the sibilantic groove is produced, both longitudinal and transverse inter-electrode distances are about 4 mm. The electrode diameter is 1.4 mm. The tongue contact pattern is registered 100 times/sec and stored in a computer.

Each EPG registration frame is a kind of map, representing the tongue-palate contact every 10 milliseconds. In this map, each electrode is represented by a specific point as either touched or free (untouched). The map points are arranged in a pattern, similar to the electrode pattern, with eight transverse rows and eight longitudinal columns of points. In our sibilantic groove analysis, we decided in which row the frontmost minimum constriction was (constriction place, CP), and counted the number of free electrodes in that row (constriction width, CW). Also back and front groove opening shapes were measured. All groove measure parameters were taken from [8, 11].

In parallel with EPG registration, also optoelectronic recording of jaw movements and acoustic registration were made, at the Dept of Prosthetic Dentistry, University of Göteborg.

Our investigation was based on 10 Swedish speakers, 4 women and 6 men (mean age 31 years, range 23-49 years). All had normal speech without strong dialect or hearing deficits. Six spoke varieties of central Swedish, 4 spoke south Swedish varieties. South Swedish lacks [ʂ], but each of /s/ and /ɕ/ are produced in the same way in all Sweden, just as [ʂ] in central Swedish [6].

The subjects had worn dummy palates, similar to the EPG ones, during a whole fortnight two years before, in connection with another study. In this study, they wore these dummy palates for four hours or more before each of three registration sessions.

The material consisted mainly of various long, natural phrases with the three sibilants in systematically varied vowel context - /i a u/, produced long or short. (For /s/, the consonant context, stress and phrase position were varied, too, but the effect of these parameter changes are not reported here.) Also isolated pronunciations of the sibilants were registered. The whole material was produced nine times.

RESULTS AND DISCUSSION

Inter- and intraindividual variation

Interindividual variation was great and intraindividual variation small for groove position and width of each sibilant. This agrees with other sibilant studies, e g [6, 9, 10, 13]. The main explanation of the interspeaker variability is that speakers with different shapes and sizes of teeth, alveolar ridge, jaw, and front tongue must reasonably produce the sibilantic groove in different ways as concerns the details, in order to achieve similar acoustic and perceptual results [6, 10]. The small intraspeaker variability is probably mainly explained by the strong demands on preciseness in directing the air jet against the front teeth in sibilants [6, 10].

Groove position and especially the /ɕ/-[ʂ] distinction

Not unexpected, /s/ was produced furthest in front, generally with a clear distance to /ɕ/ and [ʂ]. The groove front end position in /s/ ranged from immediately behind the upper front incisors to about 10 mm behind. The average distance between [s] and the other two sibilants was about 4 mm.

In phrases, /ɕ/ and [ʂ] had the same groove front end position, always mid to back alveolar - about 8-16 mm from the upper front incisors - except in one marginal case. Uttered in isolation, [ʂ] was however produced distinctly further back than /ɕ/ in 4 speakers and close to /ɕ/ in 3 out of 7. Also, the minimum groove width of these sounds was similar (usually below 7 mm). Often, their groove length, and back and front groove orifice width change shapes were also similar.

These facts support the hypothesis that the articulatory place of /ɕ/ and [ʂ] is not their primary distinguishing articulatory feature [6]. Instead, the description of sibilants must attend more to the size of the groove anterior cavity. Perhaps also the groove posterior cavity shall be considered [8], but according to [14], only the anterior cavity is important for the resonance shaping.

The size of this front cavity has been shown to be larger in [ʂ] than in /ɕ/, in two different respects [6, 7]. First, [ʂ] tends to be lip-rounded, wheras /ɕ/ has spread lips, like [i]. The [ʂ] rounding is analogous to [ʃ] rounding in English and other languages.Second, the sublingual cavity is larger in [ʂ] than in /ɕ/. The sagittal horizontal width of the upper part of this cavity is about 10 mm in [ʂ] and half that length in /ɕ/; the depth of this pocket is around 25 mm and 15 mm, respectively [6]. This sublingual cavity difference is produced by different overall tongue gestures. The tongue body is brought forwards and upwards in /ɕ/. The dorsum is convex, and the subapical tongue wall is perpendicular and tense [6]. In [ʂ], the tongue body is lower and further back. The dorsum is concave, and the subapical wall is concave and lax [6]. There is a close connection between these aspects of the upper and lower tongue walls [6], which will hopefully soon be accounted for by the developing, anatomically detailed tongue models, e g [15].

The fact that two different phonemes have their constriction in the same position in a single speaker, and also often at the same time have similar groove width and length, has implications for the general system of consonant description, as expressed in the universally used IPA

two-dimensional scheme of articulatory places and manners. This scheme is obviously the best general frame for consonantal classification, but it it not equally suitable for an adequate treating of distinctions within all classes of sounds. The sibilants are an evident example of this.

Secondary and primary palatal /ɕ/ constriction

Behind the alveolar groove in /ɕ/, and separated from it by a usually considerable widening of the vocal tract, an almost equally narrow secondary palatal constriction was found in 2 speakers, 15 to 20 mm back. In one of them this constriction was general in phrases, in the other it occurred before /i u/ but not /a/. Two other speakers had a related but much wider secondary palatal constriction. Still another speaker pronounced /ɕ/ before /i/ - but not before /a, u/ - with a primary palatal constriction, which was quite narrow - on average between about 4 and 10 mm. In this exceptional case, /ɕ/ was palatal. Otherwise, /ɕ/ was alveolar, with a groove equal in width to /s/ and [ʂ], and equal in length to in /s/, but tending to be longer than in [ʂ].

EPG data for one single central Swedish speaker in [11] disagrees with this general alveolar /ɕ/ pronunciation of 6 central and 4 south Swedish speakers. In [11], the /ɕ/ constriction was consistently palatal and wide, with a position much further back than [ʂ]. This is similar to the exceptional /i/ context case above, except for the wide constriction. It is evident that the most common Swedish /ɕ/ pronunciation is alveolar.

Groove length

In most cases, the groove length of all sibilants was less than about 7 mm. A longer groove (up to about 11 mm) was found in all /s/ productions of 2 speakers, and in all /ɕ/ productions of 2 others, and also in some speakers' production of these sounds before high vowels. However, it was not found in [ʂ]. This tendency for a somewhat shorter groove in [ʂ] and in /s/ and /ɕ/ in /a/ context appears to be caused by the lower tongue position in these sounds. Due to it, the front tongue has to be raised more, and a smaller part of it will make contact with the alveolar crest.

Groove width

On average, the groove width in /s/ was a little narrower than in the other two sibilants, which were quite similar. However, the differences were not significant. Each sound occurred with the closest groove in at least one subject, both in phrases and as isolated. In phrases, /s/ was closest in 4 subjects, and /ɕ/ and [ʂ] in one case each. However, in 4 subjects, the sibilants had fairly equal average groove widths. In isolation, the corresponding pattern was related, but the combination of individual speakers and closest sibilant was only partly identical. For example, as pronounced isolated, /ɕ/ had the narrowest grove in 3 subjects.

In phrases, the average width of each sibilant was near 2 CW units (i e 2 free electrodes, which corresponds to 5-11 mm) in 6 subjects. Two subjects had a generally closer constriction, around 1.5 units. Two subjects had a generally wider constriction around 3 units in /ɕ/ and [ʂ] - excluding /s/, with around 2 units. Obviously, each subject tended to have a general width style for all sibilants in phrases. This tendency was found also in isolated sibilants, but less pervading: Three subjects lacked this pattern there. The average groove width in isolated sibilants was however similar to the phrasal data.

This fairly constant groove width pattern in Swedish sibilants differs from English sibilants, where [ʃ] is significantly wider than [s] [9]. This difference has to be analysed more closely.

Groove width variation, related to vowel context, was found in Swedish. The pattern was complicated. In [ʂ], the variation was great, but with no general pattern. For each of /s/ and /ɕ/, the variation was small in five subjects and considerable in five (whereof three subjects are the same). For /s/, there was a general pattern: The width was smallest before /a/ and greatest before /i/. The /ɕ/ variation pattern was partly similar, with greatest narrowness in /a/, but not greatest width in /i/.

Apparently, this contextual /s/ and /ɕ/ groove width variation in several speakers had connection with tongue body position, especially height: Low tongue position was connected with a narrower groove. The same pattern was found in [10], where the first /s/ in *Å sadist* - [osa'dɪst] - was significantly narrower and had a lower tongue body position context than the second /s/. It appears that when the tongue mass is lower, the conditions for the narrow shaping of the groove are more favourable.

One possible explanation of this pattern has to do with conditions for muscular cooperation: When the authors' tongue bodies are high and front like in /i/, the tongue blade feels stiff. In /a/ on the other hand, it is slack. To shape the sibilantic front tongue groove is probably the most complicated of all articulatory gestures: All seven tongue muscle groups cooperate with a delicate balance [16]. To create a narrow groove with a stiff front tongue should be especially difficult.

A more penetrating explanatory analysis of this kind of phenomenon will hopefully soon be possible, within the framework of the now developing, detailed tongue models, e g [15]. Empirical data patterns like the groove variation above may also serve as touchstones for parts of such models.

Another factor which might contribute to the observed pattern has to do with variation in mechanical resistance: When the tongue mass is close to the oral ceiling and pressed against it, the effort to lower its median longitudinal front part will meet more resistance than otherwise. Therefore, the muscular effort to create the groove may be distributed horisontally to a greater extent.

ACKNOWLEDGEMENTS

Part of this research has been supported by grants from the Faculty of Arts, Göteborg University, and Landstinget, Kronoberg.

REFERENCES

[1] Shadle, C. (1990), "Articulatory-acoustic relationships in fricative consonants", in W Hardcastle & A Marchal (eds), *Speech production and speech modelling*, pp. 187-209. Dordrecht: Kluwer.

[2] Badin, P. (1991), "Fricative consonants: acoustic and X-ray measurements", *J of Phonetics* vol 19, pp. 397-408.

[3] Scully, C., Castelli, E., Brearley, E. & Shirt, M. (1992), "Analysis and simulation of a speaker's aerodynamic and acoustic patterns for fricatives", *J of Phonetics*, vol 20, pp. 39-51.

[4] Baer, T., Gore, J., Gracco, L. & Nye, P. (1991), "Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels", *J Acoust Soc Am*, vol 90, pp. 799-828.

[5] Lundqvist, S. (1993), "Speech and other oral functions". Dissertation, *Swedish Dental Journal*, Supplement 91.

[6] Lindblad, P. (1980), *Svenskans sje- och tje-ljud i ett allmänfonetiskt perspektiv (Some Swedish sibilants)*. Lund: Gleerup.

[7] Lindblad, P. (1978), "On the production of the Swedish [ɕ]-sound", *Aripuc* , Dept of Phonetics, Copenhagen, vol 12, pp. 31-42.

[8] Hoole, P., Ziegler, W., Hartmann, E. & Hardcastle, W. (1989), "Parallel electropalatographic and acoustic measures of fricatives", *Clinical Linguistics and Phonetics*, vol 3, pp. 59-69.

[9] Fletcher, S. & Newman D. (1991), "[s] and [ʃ] as a function of linguapalatal contact place and sibilant groove width", *J Acoust Soc Am*, vol 89, pp. 850-858.

[10] Lundqvist, S., Karlsson, S., Lindblad, P. & Rehnberg, I. (1995), "An electropalatographic and optoelectronic analysis of Swedish [s] production", *Acta Odontol Scand.* Accepted for publication.

[11] Engstrand, O. (1989), "Towards an electropalatographic specification of consonant articulation in Swedish", *Perilus*, Dept of linguistics, Stockholm, vol X, pp. 115-156.

[12] Hardcastle, W., Jones, W., Knight, C., Trudgeon, A. & Calder, G. (1989), "New developments in electro-palatography: A state-of-the-art report", *Clinical Linguistics and Phonetics*, vol 3, pp. 1-38.

[13] Bladon, R. & Nolan, F. (1977), "A video-fluorographic investigation of tip and blade alveolars in English", *J of Phonetics*, vol 5, pp. 185-193.

[14] Stevens, K. (1991), "Speech perception based on acoustic landmarks: Implications for speech production". *Perilus*, Dept of linguistics, Stockholm, vol XIV, pp. 83-87.

[15] Stone, M. (1991), "Toward a model of three-dimensional tongue movement", *J of Phonetics*, vol 19, pp. 309-320.

[16] Hardcastle, W. (1976), *Physiology of speech production*, London: Academic Press.

EQUILIBRIUM POINT HYPOTHESIS AND ARTICULATORY TARGETS IN SPEECH: A DESCRIPTION FROM SIMULATIONS OF EMPIRICAL DATA USING A BIOMECHANICAL MODEL OF THE JAW

LŒVENBRUCK Hélène[1], PERRIER Pascal[1] & OSTRY David J.[2]
[1]Institut de la Communication Parlée, URA CNRS 368, INPG & Université Stendhal
46, avenue Félix Viallet, 38031 GRENOBLE Cedex 01, France
[2]Department of Psychology, Mc Gill Unversity,
1205 Dr Penfield Avenue, MONTREAL Qc, Canada H3A 1B1

ABSTRACT

A previous study based on a simple model of speech articulator motion supported the idea that articulatory trajectories in vowel sequences could be produced from a succession of targets defined in terms of invariant equilibrium points. The prosodic variability was accounted for by modification in the timing of the commands and the level of force involved. These findings are compared with those obtained using a more sophisticated biomechanical model of the jaw.

INTRODUCTION

Controversies surrounding the notion of targets in speech production have centred on their nature (see eg [1] for a short review), on their timing ([2] *vs* [3]) and even on their existence [4].

The Equilibrium Point Hypothesis (EP Hypothesis) proposed by Feldman [5] for the control of limb movement provides some additional insight into this debate. According to this model, movements arise from changes in posture. In this context, speech targets may be associated with specific postures of the articulators and successive postures could correspond to a representation of the articulatory task at the level of control.

The notion of articulatory targets in speech production has been used in the debate on speech invariance and variability to support the idea that for a given phoneme, independent of the phonetic context, each articulator tends to approach a single position [6]. This idea that articulatory movements are intended towards spatial positions is related to MacNeilage's proposals [7].

The EP Hypothesis provides a neurophysiologically based account of how target positions can be specified [8]. In a previous study a simple second order model of the articulators, controlled in terms of equilibrium shifts, was used to examine how articulatory targets may be specified and how they may be related to an invariant linguistic task in different prosodic conditions [9].

In this paper, hypotheses based on simulations which used this simplified model are tested using a more sophisticated biomechanical model of the jaw [10].

STARTING HYPOTHESES

Lœvenbruck & Perrier [11] showed that the EP hypothesis can shed light on the control of vowel reduction. A second order model, consisting of two springs which simulated agonist and antagonist muscle groups was used to provide a simple way of controlling the gesture's dynamic parameters: the stiffness ratio specified the position of the intended spatial targets, whereas the cocontraction level (sum of the stiffnesses) and the timing of the commands determined the dynamic behaviour of the model.

The [iai] sequence was studied in the French sentence "Il y a immédiatement". The sequence was tested in three different conditions: (1) slow speech rate and stressed [a]; (2) slow rate and unstressed [a]; (3) fast rate and stressed [a]. Data from one French native speaker were analyzed. Articulatory trajectories were inferred from the acoustic signal by an inversion procedure involving an articulatory model. Condition (1) was supposed to be the "ideal" one, in the sense that the observed articulatory positions for [i], [a] and [i] correspond to the intended articulatory targets. In both the other cases, the movement extent was reduced.

Lœvenbruck & Perrier showed that the three different articulatory trajectories could be generated using the same successive intended articulatory targets for [i], [a] and [i]. The differences between the trajectories in the three conditions could be simulated by modifying the cocontraction level and the timing of target shift. In order for the model to replicate the kinematic patterns of stressed and unstressed vowels produced at slow speech rate it was necessary that cocontraction (and thus total force) was greater for the stressed vowel; no differences were observed in the timing of equilibrium shift. In contrast, the main difference between slow and fast stressed conditions was the timing of the shift; the cocontraction levels were almost identical. In the present paper, these findings serve as initial hypotheses for tests carried out with a more sophisticated model. Specifically, the timing of the commands should remain fairly constant at a given speech rate; movements for stressed vowels should involve greater total force than for unstressed vowels.

SIMULATIONS WITH A BIOMECHANICAL MODEL OF THE JAW

The jaw model

The model proposed by Laboissière et al.[10] has seven muscles (or muscle groups) and four kinematic degrees of freedom. Besides a more elaborate biomechanical formulation, this model has the advantage over the one used in the previous work of including a fuller account of the neurophysiological control mechanism. The essential control variables are independent changes in the membrane potentials of motoneurones (MNs) which establish a threshold muscle length (λ) at which the recruitment of MNs begins. Muscle activation and hence force vary in relation to the difference between the actual and the threshold muscle lengths and the rate of muscle length change. Thus, by shifting λ through changes to the central facilitation of MNs, the system can produce movement to a new equilibrium position.

In the jaw model, movements are not controlled directly in terms of commands to individual muscles. Rather control signals, which are based on different combinations of λs, are organised at the level of the system's kinematic degrees of freedom. This enables production of jaw rotation, horizontal jaw translation, vertical hyoid translation and horizontal hyoid translation. The level of cocontraction is also controlled by specifying the global level of force involved in the movement. These control signals may act alone or in combination.

Methods

The corpus was the same as the one used for the previous work. A different native speaker of French was tested.

Horizontal jaw translation and rotation in the midsagittal plane were tracked using the Optotrak system which captures the light emitted by infrared emitting diodes (IReds). An acrylic dental appliance accurately fitting the lower teeth was built from the dental impression of the subject. A lightweight but rigid dental wire was attached to the front of the appliance and was shaped so that its two ends came out of the mouth horizontally at the corner of the lips. A total of five IREDs were attached to bamboo sticks which were glued to the dental wire. These IREDs were used to track the motion of the jaw. An additional six IREDS were attached to a head mounted acrylic frame and were used to correct for head motion. IRED positions were sampled at 100 Hz and low pass filtered using a Butterworth filter. The acoustic signal was simultaneously recorded and sampled at 10 kHz. The orientation angles and positions which characterise the motion of the jaw were reconstructed from the IRED motions.

The [iai] sequence was extracted from the entire sentence using the Vocalic Voiced Onset and Vocalic Voiced Termination criteria [12].

Simulations

Only rotation of the jaw was considered because jaw rotation is the most relevant articulatory variable in the [i-a] transition, in the sense that the transition is characterised largely in terms of differences in jaw opening angle.

The same strategy as in the previous work was used, *ie* inferring the intended targets from the articulatory signal observed under the slow and stressed condition. We specifically tested the idea that stressed movements were associated with a high total force or high cocontraction level and changes in speaking rate were associated with changes in the duration of the command.

The results of the simulation for the slow and stressed condition are presented in figure 1, where the data and the simulated trajectories are plotted in dotted and solid lines respectively. As can be seen the fit to the data is rather good. Note that the actual position for the second [i] is influenced by the incoming context which is not taken into account in our simulations. The discrepancy between the data and the simulation for [a] is due to the dynamic coupling of sagittal plane rotation and horizontal translation. The actual [a] position slightly undershoots the target defined by the intended equilibrium position.

The fast and stressed condition is simulated by a reduction in the control signal underlying [a]. The amount of force is the same as in the previous condition. The fit here is likewise relatively good (figure 2).

Finally, the slow and unstressed condition was simulated by setting the global force to its minimum possible value and by reducing the equilibrium shift rate. The decrease in shift rate effectively eliminates a stationary position of [a]. Under these conditions, one may observe a clear reduction (-3 deg.) in the amplitude of the simulated movement. Under these conditions a suitable fit to the data was not possible (figure 3).

CONCLUSION

The simulation obtained for the fast and stressed condition could be obtained in a manner consistent with our initial hypotheses. A simple reduction of the hold duration for the vowel [a] produces a target undershoot comparable to that observed in the empirical data. The high level of global force as well as the relatively fast transition rate enable the simulation of the sharp transition which is observed. Moreover the reduction of movement amplitude obtained by decreasing the global force in the simulation of the slow and unstressed condition, confirms the role of the cocontraction level as an efficient parameter for dynamic control.

However reducing the force and the transition rate were not sufficient to produce an undershoot comparable to that observed empirically. This argues therefore against our starting hypothesis that articulatory targets remain the same independent of stress. A better fit to the empirical trajectory was obtained by reducing the amplitude of the equilibrium shift from [i] to [a] (figure 4). This can be interpreted in two ways:

(1) a change in stress corresponds to a change in the intended articulatory target.

(2) a decrease in the equilibrium shift rate induces an undershoot at the level of the control variables.

The absence of stationary target position for [a] in the slow and stressed condition favours the second hypothesis. However, further tests should be carried out, by comparing, for example, the articulatory patterns for [ɛ] and [a], which are acoustically similar under conditions of vowel reduction.

ACKNOWLEDGEMENT

This work is supported by the Esprit B.R. Project n° 6975, Speech Maps, by the cooperation France-Québec (Projet n° 07-01-92) and by the NIH grant DC-00594 from the National Institute of Deafness and Other Communication Disorders.

REFERENCES

[1] MacNeilage P. (1980). Distinctive properties of speech motor control. In G.E. Stelmach and J. Requin (eds.)*Tutorials in motor behavior* ,607-621. Amsterdam, The Netherlands: North Holland publishing company.

[2] Lindblom B., Lubker B, Branderu P. and Holmgren K. (1987). The concept of target and speech timing, 161-182, Channon R. and Shockey L. (eds). Honor of Isle Lehiste, Foris: Dordrecht, Holland.

[3] Fowler C.A. (1980). Coarticulation and theories of intrinsic timing. *J. Phonetics*, 8, 113-133.

[4] Pols C. W. & Van Son R.J.J.H. (1993). Acoustic and perception of dynamic vowel segments. *Speech Comm.* 13, 135-147.

[5] Feldman A.G. (1966). Functional Tuning of The Nervous System with Control of Movement or Maintenance of a Steady Posture-II Controllable Parameters of the Muscles. *Biophysics, 11*, 565-578.

[6] Lindblom B. (1963). Spectrographic study of vowel reduction. *J. Ac. Soc. Am.* 35, 1773-1781.

[7] MacNeilage P. (1970). Motor control of serial ordering of speech. *Psy. Rev. 77, 182-196.*

[8] Perrier P. & Ostry D.J. (1994), "Dynamic modelling and control of speech articulators. Application to vowel reduction." In Keller E. (Ed.), *Fundamentals in Speech Synthesis and Speech Recognition* ,231-251. London, U.K. : J. Wiley and Son.

[9] Perrier P., Lœvenbruck H. & Payan Y. (submitted), "Control of tongue movements in speech: The Equilibrium Point hypothesis perspective.", *J. Phonetics.*

[10] Laboissière R., Ostry D.J. & Feldman A.G. (submitted). The Control of Human Jaw and Hyoid Movement. *J. of Neurophysiology.*

[11] Lœvenbruck H. & Perrier P. (1993). Vocalic reduction: prediction of acoustic and articulatory variabilities with invariant motor commands. *Acts of the 4th European Conference on Speech Communication and Technology* , 85-88. Berlin, RFA.

[12] Abry C , Benoit C, Boë L.J. & Sock R. (1985). Un choix d'événements pour l'organisation temporelle du signal de parole. Actes des 14èmes JEP Paris, 133-137.

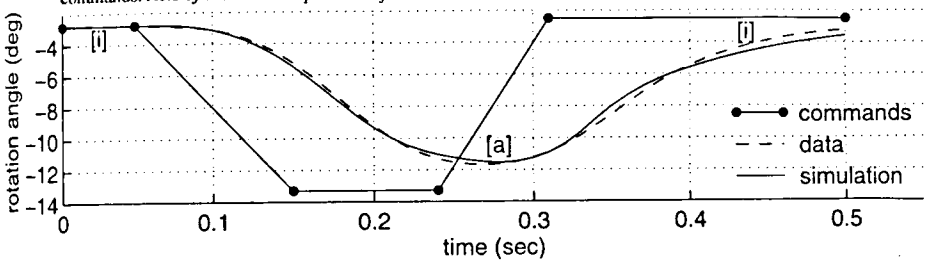


Figure 1 : Simulation under the slow and stressed condition. The equilibrium targets were E(i1) = -2.8 deg for the first [i], E(a) = -13.3deg for [a] and E(i2) = -2.5deg. for the last [i]. The level of force (computed for the first [i]) was F = 78.4N, the hold time for the first [i] was Thold1 = 0.05s, the [i-a] transition time was Tt1 = 0.1s, the hold time for [a] was Thold2 = 0.09s and the [a-i] transition time was Tt2 = 0.07s.

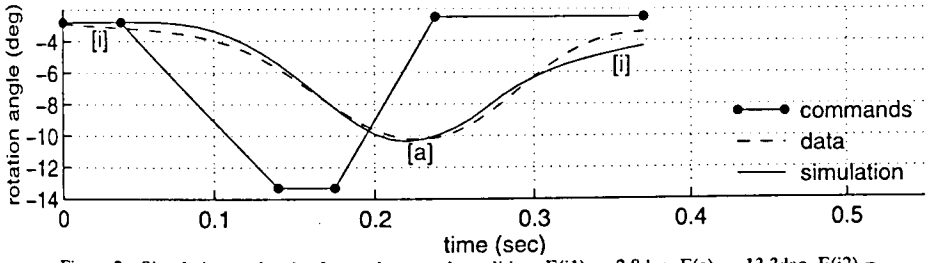


Figure 2 : Simulation under the fast and stressed condition: E(i1) = -2.8deg, E(a) = -13.3deg, E(i2) = -2.5deg., F = 78.4N, Thold1 = 0.04s, Tt1 = 0.1s, Thold2 = 0.035s and Tt2 = 0.063s.

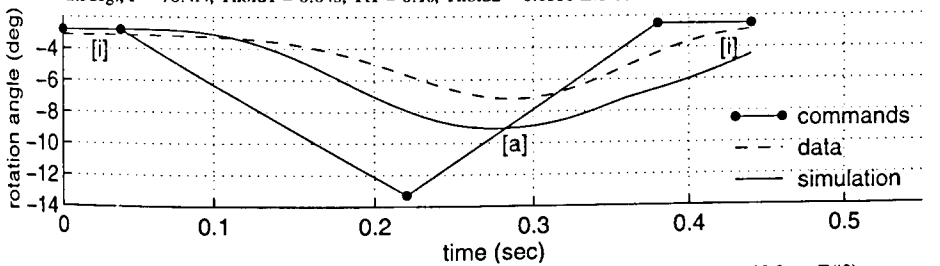


Figure 3 : Simulation under the slow and unstressed condition: E(i1) = -2.8deg, E(a) = -13.3 eg, E(i2) = -2.5deg., F = 10.6N, Thold1 = 0.04s, Tt1 = 0.179s, Thold2 = 0.001s and Tt2 = 0.16s.

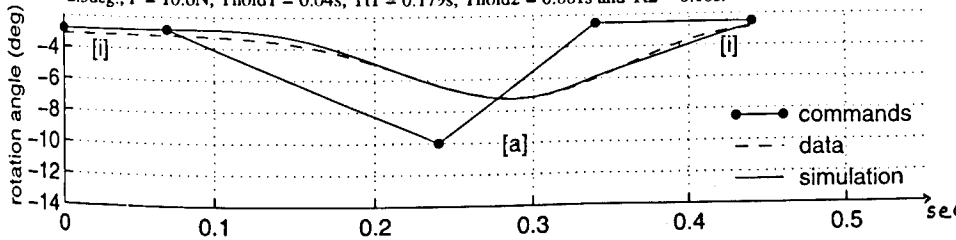


Figure 4 : Simulation under the slow and unstressed condition: E(i1) = -2.8deg, E(a) = -10deg, E(i2) = -2.5deg., F = 10.6N, Thold1 = 0.07s, Tt1 = 0.169s, Thold2 = 0.0001s and Tt2 = 0.1s.

ACOUSTIC MODEL FOR AN ARTICULATORY-FORMANT SPEECH SYNTHESIZER

A. Miller [1] *and V. Sorokin*
Institute for Information and Transmission Problems, Moscow

ABSTRACT

Effective computational methods for solving the vocal tract equation relative to resonance frequencies, amplitudes and dampings were developed. Effects of the yielding walls were described with the equivalent sound velocity which enabled an accurate frequencies-domain parameters calculation. A nasal cavity coupling was also computed. Validity of the acoustic model was tested in synthesis experiment.

INTRODUCTION

In an articulatory-formant speech synthesizer acoustic processes are presented by a set of frequency-domain parameters - formants, dampings and amplitudes coefficients. There was tried on two approaches for these parameters search. In [1] formant parameters were calculated by converting transfer function derived through the transmission line analog representation of the vocal tract. In method of [2] formants was obtained directly from the vocal tract equation. Potentially the last approach takes advantage of more accurate and numerically economical description of the vocal tract acoustics because of applying direct analytical consideration of the vocal tract equation.

Following second approach we attempted to develop algorithms for formant parameters calculation from the vocal tract area- function. Algorithms are based on frequency-domain analysis of the vocal tract equation and designed to include walls vibrating, distributed losses and nasality. The paper is organized as follows. First, algorithms for vocal tract computation are presented. Then effect of nasal coupling will be described. Then practical realization of acoustic model will be discussed. Finally, obtained results are summarized.

[1]Currently with SR Telecom Inc., Montreal, Canada

SEARCH OF FORMANTS

As usual we assume that wave propagation in lossy acoustic tube with soft walls is described as it was done in [3][4]. For formants search, first of all, the appropriate expressions to account for wall yielding are to be found. To account for yielding walls we used the method of an effective sound velocity [5]. For small perturbations of area function $A_0(x,t)$ after wall displacement in the radial direction we get equivalent sound velocity as

$$c_e(t) = \left(\rho_0\beta_0 + \frac{S_0\xi}{pA_0}\right)^{-0.5}$$

where $\beta_0 = 1/\rho_0 c_0^2$.

In case of quasi steady-state deformations of the vocal tract frequency decomposition of Webester's equation gives us the following equation for spatial mode

$$(A_0\varphi')' + \frac{\omega_k^2 A_0}{c_e^2}\varphi = 0 \qquad (1)$$

where $\varphi(x)$ is the spatial component of the pressure, respectively, " $'$ " and " $''$ " denote first derivative on x. For a cylindrical tube with vibrating walls

the equivalent sound velocity c_e becomes

$$c_e^2(x) = \frac{c_0^2}{1 - 2Y_w/\mathrm{j}\omega_k a\beta_0} \qquad (2)$$

where $a(x)$ is radius of the tube, $Y_w(\omega_k, x)$ is conductivity of walls.

The equivalent sound velocity appears to be a generalization of effective sound velocity of [6] for the case of arbitrary dependence of the wall impedance from vocal tract geometry. Besides, Eq. (1) takes into account the active resistance ignored in [6].

The boundary condition at the lips end following model of [3] is $\varphi'(L) + 3\pi\varphi(L)/8a(L) = 0$. The boundary condition at the glottis for the closed vocal slit is $\varphi'(0) = 0$.

The eigenfrequencies of Eq. (1) can be found by solving an equivalent initial value problem (Kauchy problem) by means of shooting method. However, the original boundary problem represented as the equivalent Kauchy problem has unpleasant error propagation properties when abruptness occurs in the area function $A_0(x)$. To provide robustness the method of a phase function [7] was implied. This method does not superimpose explicit constraints on the spatial derivative of $A_0(x)$ and, besides, it offers more economical computation.

Table 1 represents three first formant frequencies calculated for the area functions of Russian vowels taken from [8](column E). Column A represents absolutely rigid walls. The cross-section shape was described as ellipse with the axis ratio 2, the wall impedance was $Z_w = 800 + \mathrm{j}1.3\omega$ g/cm·s . The formant frequencies represented in the column B were taken from [6], calculated by the effective sound velocity method. As one can see, essential errors can be observed for the vowels /a/, /i/ and /i/ in relation to the measured frequencies. Meanwhile, the frequencies calculated with Eq. (1) (the column D) are close to those calculated by the transmission line analog (the column C) and both are better fit to the experimental data (the column E).

F1	A	B	C	D	E
a	636	668	677	679	700
e	424	459	459	442	440
o	495	538	538	536	535
u	228	302	291	286	300
i	231	295	285	257	240
ɨ	289	345	-	312	300
F2					
a	1098	1101	1088	1105	1080
e	1980	1981	1998	2003	1800
o	848	887	869	875	780
u	606	629	631	605	625
i	2245	2284	2338	2329	2250
ɨ	1531	1541	-	1562	1480
F3					
a	2400	2477	2412	2488	2400
e	2741	2821	2748	2887	2550
o	2380	2398	2432	2386	2500
u	2385	2391	2450	2385	2500
i	2888	3102	2898	3116	3200
ɨ	2414	2421	-	2366	2230

Table 1: Resonance frequences for Russian vowels.

DAMPINGS

Acoustic losses may be presented as a resistance $R(\omega, x)$ reflecting losses arising from viscous friction and lips radiation and an admittance $Y(\omega, x)$ reflecting losses caused by the thermal conduction and wall vibrations [5]. Those losses are accounted for as an effective coefficients of damping δ_k for each k-th mode, so that the equation for temporal mode is

$$\psi'' + 2\delta_k\psi' + \omega_k^2\psi = 0.$$

Using the decision of the last equation and Eq. (1) we get after integrating

by x and averaging by $T = \omega_k/2\pi$ as follows

$$\delta_k = \int Y\varphi_k^2 dx + \int R(A_0\varphi_k/c_e)^2 dx$$

Expression for δ_k was obtained under assumption of small losses, i.e. $\delta_k/\omega_k \ll 1$.

AMPLITUDES

The amplitude of forced oscillations is determined as

$$\psi'' + 2\delta_k\psi' + \omega_k^2\psi = e_k,$$

where the coefficient e_k is calculated as a result of decomposition of the excitation function $G(x,t)$ with the eigenfunctions of Eq. (1) as follows

$$e_k = \int G(x,t)A_0\varphi_k/c_e^2(x)dx.$$

For the source of current in the vocal slit function of excitation becomes $G(0,t) = u_g'(0,t)$. So the amplitude of oscillations determined as

$$e_k(t) = u_g'(t)A_0(0)\varphi_k(0)/c_e^2(0)$$

Other sources of excitation are determined in the same way.

NASALIZATION

If the velum is lowered then the vocal tract consists of three united parts: pharyngeal, oral and nasal cavities. If eigenfunctions for all the cavities are known, then resonance frequencies can be obtained from the condition of inconsistency of flow and pressure at the velum port [9].

Amplitudes and dampings are calculated in the manner analogous to non-branched case described above.

The resonance frequencies for the vocal tract with the coupled nasal branch are represented in Table 2 for the area functions of vowels used in Table 1. The velum opening was equal to 1 cm. To search formant frequencies we used numerical algorithm developed for branched vocal tract in [10] . As it was expected, some additional resonances appear between the resonances inherent to the vocal tract itself.

	F1	F2	F3	F4	F5
a	573	912	1157	2341	2961
e	449	906	1956	2638	3020
o	495	888	2319	2775	3612
u	314	537	758	2273	2630
i	315	881	2325	2737	3146
y	347	876	1557	2353	2744

Table 2: Resonance frequences for nasalized Russian vowels.

PARALLEL SYSTEM

The vocal tract model provides computation of amplitudes and bandwidths which are presented as a vector

$$\vec{x} = \left[\vec{F},\ \vec{B},\ \vec{\varphi}_L,\ \vec{\gamma}_N,\ \vec{A}\right]^{\mathrm{T}},$$

where $\vec{F} = [F_1, ..., F_m]^{\mathrm{T}}$, m is number of formants, the rest of the $\vec{x}$ components are defined in the same way. Variations of articulatory and acoustic conditions may cause variations of quantity of the formants for given frequency band. Selectors were introduced into parallel system to provide automatic ordering of calculated formants. The selector and the block of formant filters constitute two parallel branches for independent processing of nasal and oral resonances.

The selector compares two successive sets of calculated formants and provide distribution of calculated formant parameters within fixed set of formant filters. We used eight canals for each branch to cover the 5 kHz band.

The model of the synthesizer embodies an aerodynamic model for computation function of excitation of the vocal tract. Consideration of aerodynamic model is beyond of the paper and may be found elsewhere [1].

Now, let us estimate the number of operations required for frequency parameters computation. For 5 ms interval of area function update and 0.5 cm of spatial increment (error of formant estimation less the 4%) it takes about 1.5sec of IBM PC/486/66 CPU time per 1sec of synthesized speech. Computational expenses for the formant frequency calculations are about 15% of the computational cost for the speech wave computation with sampling frequency 20 kHz.

To verify validity of the described acoustic model, we had a synthetic experiment with an articulatory-formant speech synthesizer based on the aerodynamic model of the vocal tract [9] and laboratory model of the vocal tract dynamics. Listening tests with speech material consisting of words and short phrases indicated that both intelligibility and naturalness of synthesized expressions were close to those of natural ones.

CONCLUSIONS

Combining the advantages of both analytical and numerical approaches, the effective technique for vocal tract computation has been developed. Simple and fast algorithms for calculation of the main acoustic parameters - frequencies, amplitudes and dampings of resonance oscillations - take into account such important factors as yielding walls, nasalization and distributed losses.

Algorithm for formants search provides the amount of operations about 1.5 s of IBM PC/486/66 CPU time per sec with no special signal processing equipment. The achieved accuracy and speed are acceptable for application of these computational algorithms in the tasks of articulatory speech synthesis.

References

[1] Lin Q. "Theory of speech production and articulatory speech synthesis," Ph.D. Thesis, KTH, Stockholm, 1990.

[2] Coker C.H. "A model of articulator dynamics and control," Proc. IEEE 64, 1976, 452-460

[3] Morse P.M. Vibration and sound (McGraw-Hill, New York), 1948

[4] Portnoff M. "A quasi one dimensional digital simulation for the time-varying of the vocal tract," MS Thesis, MIT, 1973.

[5] Flanagan J.L. "Speech Analysis, Synthesis and Perception (Springer Verlag Berlin, Heidelberg, New-York), 1972.

[6] Badin P., Fant G. "Notes on vocal tract computation," STL/QPSR 6, 1984, 53-108.

[7] Hall G. and Watt J.M. Modern numerical methods for ordinary differential equations (Clarendon Press, Oxford), 1976.

[8] Fant G. Acoustic Theory of Speech Production (Mouton, The Hague, The Netherlands), 1960.

[9] Sorokin V.N. Theory of speech production. Moscow, 1985. (in Russian)

[10] Miller A.H., Sorokin V.N. "Methods of shooting for the vocal tarct equation", Acoustical journal, 2, 1991, 361-367. (in Russian)

THE ARTICULATORY CORRELATES OF DESCRIPTIVE CATEGORIES FOR SUPRALARYNGEAL VOICE QUALITIES

Francis Nolan[1] and Barbara Kühnert[1,2]

[1]Department of Linguistics, University of Cambridge
[2]Institut für Phonetik und sprachliche Kommunikation, University of Munich

ABSTRACT

This paper reports a pilot experiment using EMA (electromagnetic articulography) to monitor the 'articulatory settings' hypothesised to underlie supralaryngeal components of voice quality as defined in Laver's [1] descriptive framework. The experiment also tests whether segmental articulation is modified to preserve acoustic targets under changes of articulatory setting.

1 INTRODUCTION

'Voice quality' is used here in the sense of Abercrombie [1]: '...a quasi-permanent quality running through all the sound that issues from [a speaker's] mouth' (p.91). Laver [2] proposes an analysis for voice quality based on auditorily identified components, each hypothesised to be associated with an 'articulatory setting', a long-term biasing of the vocal organs towards a given configuration. Auditory components such as *breathy*, and *palatalised*, would result from a general tendency to adduct the vocal cords less strongly, and to bias the configuration of the tongue towards [i], both relative to neutral baselines.

Laver's descriptive framework 'stands on an auditory foundation' (despite the superficially articulatory labels such as *palatalised*), but 'the auditorily-identified components all have correlates ... capable of instrumental verification – [at] the articulatory, physiological, and acoustic levels' (p.7). Although much research has been done on correlates for laryngeal components, there has been little on the correlates of supralaryngeal components. Nolan [3] carried out an acoustic analysis of suprasegmental components in Laver's framework, and found systematic shifts in formant frequencies. Esling [4] also looked at spectral correlates of components such as *velarised* and *laryngo-pharyngalised*. Neither study included articulatory measurement. The first aim of the present experiment is to test the articulatory settings claimed to underlie *palatalised* and *velarised* voice.

Although a voice quality component can be seen as resulting from a bias in articulatory activity which pervades the whole of a person's speech, segmental requirements may impose limitations. For instance, *nasalisation* may be reversed by fricatives and plosives. But a more intriguing segment-setting interaction would be if a segment's articulation adjusted to preserve an acoustic requirement despite an 'unhelpful' setting. English [ɹ] and [ʃ] provide a potential example. The low F3 and/or F4 of [ɹ] and the relatively low first major spectral prominence of [ʃ] depend on a sufficiently large cavity in front of the major constriction. This is achieved by a post-alveolar constriction, augmented for many speakers by a degree of lip rounding and protrusion. If the acoustically effective size of the cavity is reduced in size, as in smiling, the tongue constriction may compensate by moving back, as noted by Andrew Crompton (pers. comm.). The second aim of the experiment is to explore this kind of compensatory articulation under *lip-rounded* and *lip-spread* settings.

2 EXPERIMENT

The first author acted as subject. He has extensive familiarity with Laver's framework, and has attended a training workshop in its use. Two sentences were used. The first, 'When the sunlight strikes raindrops in the air they act like a prism and form a rainbow', is part of the 'Rainbow' passage used in Nolan [3]. The sentence was read three times implementing each of the components *palatalisation* and *velarisation*, and three times *neutrally* as a control. The second sentence was 'The red rooster shattered the rural quiet with three very short shrieks', designed to contain several examples of [ɹ] and [ʃ]. This was produced three times implementing each of *lip-rounding*, *lip-spreading*, and *neutral*. Only by using controlled performances by a phonetician is it possible to study the effect of a single component. In real speakers the effect of a component would be conflated with other factors, including anatomical differences. The approach adopted here is comparable to using a phonetician's Cardinal Vowels to explore the acoustic-articulatory mapping in the vowel space.

Time-aligned acoustic, EMA (Carstens AG 100), and electro-palatographic recordings were made (the latter not discussed here) in collaboration with Phil Hoole at the Institut für Phonetik und sprachliche Kommunikation, Munich. Receiver coils were attached as follows: lower lip, lower front teeth (for vertical jaw movement), tongue blade, tongue front, tongue back, and (to allow compensation for head movement) the upper front teeth and bridge of the nose. Data processing was carried out as described in Hoole [5].

3 RESULTS

Compared with a neutral tongue-body setting, the 'centre of mass' of the tongue body would be expected to be raised and slightly fronted in *palatalisation*, and raised and retracted in *velarisation* (Laver [2] p.46).

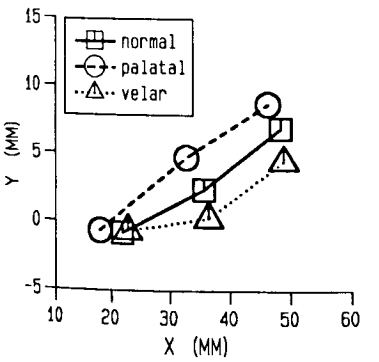


Fig 1. Mean position of EMA tongue-coils for three voice quality components.

Fig 1 shows the effect of these three settings (including neutral) on the position of the EMA coils. The x-axis shows front-back position in the occlusal plane, that is, parallel to the plane at which the teeth meet, and the y-axis shows position perpendicular to the occlusal plane. The front of the mouth is at the left. The points joined by lines are the position of the blade, front, and back coils for each setting. Each point is the mean of all analysis frames for utterances in a particular setting (approximately 2400: 4s utterance duration x 3 repetitions x 200 frames/s).

Palatalisation does, as the label suggests, involve raising and fronting of the body of the tongue. Less predictably, *velarisation* apparently brings a lowering of the tongue. What may be happening is that the tongue body is arching back and up behind the rearmost coil (placed below the join of the hard and soft palate), and the effect on the 'visible' part of the tongue is one of lowering, as the mass of the tongue retreats back. Nonetheless it does seem that the main secondary constriction is further back than the velar region, and formant data for subsets of high front and open vowels (Table 1) reveal that (particularly for the high front vowels) the values for *velarisation* in the present study are similar to those for *pharyngealisation* in the earlier recording of the same speaker.

Table 1. Formant frequencies by setting (data in italics: Nolan [3], Table 4.4).

	F1	F2	F3
	High Front Vowels		
Neutr.	396 *400*	1843 *1850*	2502 *2480*
Pal.	409 *390*	1927 *1950*	2650 *2650*
Vel.	448 *405*	1664 *1825*	2489 *2525*
Pharyn.	*465*	*1675*	2430
	Open Vowels		
Neutr.	669 *690*	1255 *1210*	2502 *2480*
Pal.	667 *670*	1562 *1430*	2650 *2650*
Vel.	649 *620*	1209 *1195*	2489 *2525*
Pharyn.	*685*	*1170*	*2460*

Laver (pp.55–6) notes there is difficulty in defining *velarisation*. Whether an ambiguity in the framework or inaccurate performance explains the

present result might be determined by having listeners trained in the framework assess the present recordings.

Fig. 2 shows EMA data for the second sentence, which was realised with *neutral*, *lip-rounded*, and *lip-spread* components. For these components the lower lip coil (not shown) varies as expected, while the contour of the tongue is essentially the same. However it is noticeable that for *lip-spread*, each of the tongue coils is slightly more retracted, and the coil on the tongue front is slightly lowered. This would not be predicted from the definition of the components, which imply purely labial settings.

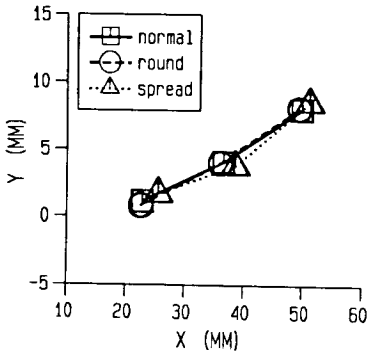


Fig 2. Mean position of EMA tongue-coils for labial voice quality components.

It is appropriate then to examine the segments for which compensatory articulation was predicted. Fig. 3 shows the average position of the tongue coils for the acoustic midpoint of [ɹ] in 'red', 'rooster', and 'very'. Fig 4 shows [ʃ] from 'shattered' and 'short'. For both sounds, the tongue is generally retracted (compare with the average coil positions for the whole utterance in Fig. 2); but it is relatively less retracted in lip-rounded, and more retracted in lip-spread; and in [ʃ] the blade is also lowered in lip-spread, suggesting retroflexion. The extra retraction (and retroflexion) would help to maintain the acoustic lowering effect otherwise reduced by the lip-spreading. In lip-rounded, the tongue may be taking advantage of the extra rounding, and retracting less than normal for these sounds. This supports the notion that there may be compensation in the articulation of a segment to maintain its target acoustic characteristics in the face of different settings.

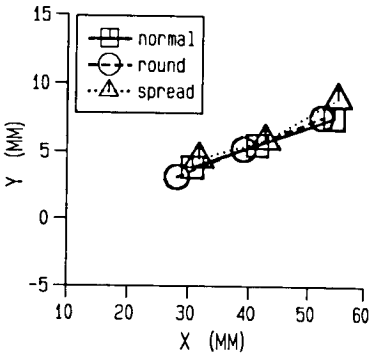


Fig 3. Mean position of EMA tongue-coils for [ɹ].

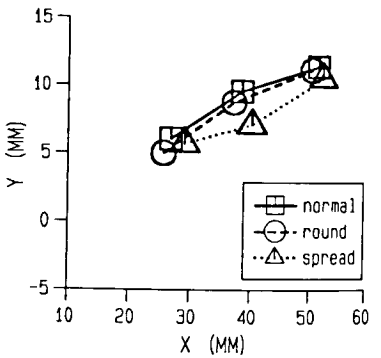


Fig 4. Mean position of EMA tongue-coils for [ʃ].

To test whether acoustic invariance is achieved, formant values were measured for [ɹ], and average spectra derived for [ʃ]. For [ɹ] (Table 2) it appears that F2 and F3 are relatively constant between *neutral* and *lip-rounded*; the articulatory retraction has not, however, fully compensated the effect of *lip-spreading*. The resultant spectra for [ʃ] in *neutral*, *lip-rounded*, and *lip-spread* are overlaid in Fig. 5. Although there are differences in the overall shape of the spectra, it is noticeable that the first major peak, which may contribute to the lower energy cutoff

important in the perceptual distinction between [ʃ] and [s], is virtually constant at around 1600 Hz.

Table 2. Formant frequencies for [ɹ]

	F1	F2	F3
Neutr.	347	1358	1766
Rnd.	307	1357	1711
Spread	376	1711	1918

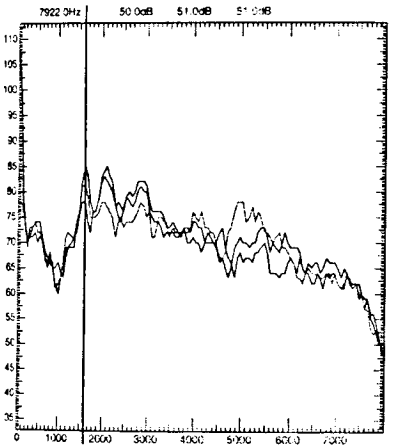


Fig 5. Overlaid average spectra for [ʃ] under three labial conditions.

4 DISCUSSION

This experiment is only a small beginning to the articulatory validation of the settings hypothesised to underlie perceptual components of voice quality. It has, however, shown the applicability of EMA as a techniques of analysis, and the results have borne out the assumptions made about *palatalisation* and raised questions about *velarisation*. Testing the mapping of terms in the framework onto physical dimensions is not a trivial exercise. Just as potentially the descriptive dimensions 'height' and 'frontness' for vowels may not closely correspond to articulatory fact for a set of vowels, so descriptive labels such as *palatalised* (voice) may not be accurate. Indeed it seems that the auditory quality labelled *velarised* may involve a rather more backward bias of the tongue body than the name implies.

The experiment also demonstrates the complex interaction between settings and segments: [ɹ] and [ʃ], whose criterial acoustic properties are sensitive to the pre-apical cavity, compensate for lip settings by fine adjustments in the coronal articulation. It is tempting to argue from this to a model of speech production which takes auditory goals as primary, and derives articulatory configurations by predictive modelling; but such an argument lies outside the scope of the present paper.

REFERENCES

[1] Abercrombie, D. (1967) *Elements of General Phonetics*. Edinburgh: EUP

[2] Laver, J. (1980) *The Phonetic Description of Voice Quality*. Cambridge: CUP.

[3] Nolan, F. (1983) *The Phonetic Bases of Speaker Recognition*. Cambridge: CUP.

[4] Esling, J.H. (1987) Vowel shift and long term average spectra in the Survey of Vancouver English. In *Proc. 11th ICPhS* Vol 4, 243-6. Tallinn: Academy of Sciences of the Estonian SSR.

[5] Hoole, P. (1993) Methodological considerations in the use of EMA in phonetic research. *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation* 31, 43-64.

Simulation of tongue shape variations in the sagittal plane based on a control by the Equilibrium-Point hypothesis.

Payan Yohan, Perrier Pascal & Laboissière Rafaël
Institut de la Communication Parlée, URA CNRS 368, INPG & Université Stendhal
46, avenue Félix Viallet, 38031 GRENOBLE Cedex 01, France
E-mail : payan@icp.grenet.fr

ABSTRACT

A 2D biomechanical Finite Element model of the tongue was developed. It integrates four extrinsic muscles and three intrinsic ones, and is controlled by the Equilibrium Point Hypothesis. The ability of this model to generate realistic VV transitions is evaluated and implications for the control of speech are discussed.

INTRODUCTION

Very early on, speech scientists were convinced of the need for models at different levels of speech production process, in order to understand the link between physical aspects of speech and the linguistic task. By modelling the relationship between vocal-tract geometry and the acoustic signal, Stevens and House [1] and lates Fant [2] emphasised the importance of vocal-tract constriction in vowel production; Wood [3] proposed a correspondance between vowel classes and constriction location. With an articulatory model Maeda [4] underlined tongue and jaw coordinations used for the production of the same vowel in different contexts. More recently, Maeda et al [5] proposed a model to account for muscle synergies in the control of position and shape of the tongue.

A speech production model should of course include such synergies. In this perspective, a first approach could involve the learning process [6]: synergies would emerge from the search for the most efficient motor strategies suitable for the production of given articulatory patterns. Another interesting approach consists in moving a step closer to the Central Nervous System (CNS), and studying how the motor control signals to the articulators are organised. From this point of view, Feldman's Equilibrium Point hypothesis [7] is very appealing: in this theory, muscles are not controlled individually, but in respect to global control variables. Muscle coordination is thus implicitly described in the model.

Laboissière et al. [8], using a physiological jaw/hyoid bone model [9] have already emphasised some of the contributions this hypothesis offers for the understanding of the control of speech production.

Our purpose in this paper is to show how this hypothesis sheds light on the control of the tongue, in VV sequences.

The TONGUE MODEL

State of the art.

The tongue is a non-rigid body, capable of bearing large deformations, in order to precisely shape the oral cavities for an accurate production of sounds. By contrast to the jaw (which is an undeformable bone, moved by the action of external muscles), the tongue has embedded muscles, whose shape will be modified under their own activity.

In order to understand the respective contribution of each muscle in the tongue's shaping, Perkell [10] showed the interest of a physiological model of the tongue, which takes into account the intrinsic physiological and anatomical structure which underlines tongue movements. The elastic properties of the tissues were represented by distributed, individually controlled, second order systems. A better analytical description of the continuous elastic structure of the tongue is proposed by the Finite Element Method (FEM). However, up to 1993 Finite Element tongue models were only able to produce static tongue shapes for given EMG inputs [11], [12], [13]. Important progress has been made by Wilhelms-Tricarico [14], who elaborated a 3D biomechanical Finite Element model integrating inertia and force generation properties. This model is able to solve the equations of motion.

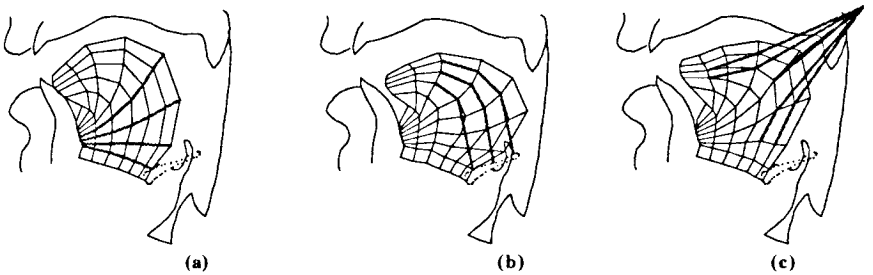


Figure 1: Tongue muscles action : (a) Posterior Genioglossus; (b) Hyoglossus; (c) Styloglossus. For a better representation, the tongue model is embedded in a drawing of the sagittal outlines of the vocal tract.

Such a model is most suitable for evaluating motor control theories. However, attention should be paid to the high degree of computational complexity, which could involve enormous simulation times. In order to overcome this possible drawback and to permit an extensive evaluation of the EP hypothesis for the control of speech movements, we propose to first limit the description of the tongue to the sagittal plane. Even with this limitation, important progress can be made by working on the numerous articulatory data available in this plane (X-ray, microbeams, electromagnetometer data).

A 2D Finite Element model of the tongue

In brief, FEM is a classical technique, in which interpolation functions are developed in order to integrate, across regions of interest, continuous material properties, such as mass, stiffness and deformation capabilities. For this aim, the non-rigid body is described in terms of *discrete nodal* points. The global deformation of the body is analytically calculated from the displacements of these nodes.

The proposed model is described by 63 nodes. The tongue structure is thus defined by 48 isoparametric elements, each of them being delimited by four nodes. This structure satisfies two main criteria : ① the distribution of the nodes, should reflect the internal muscle structure in the sagittal plane, but ② the number of nodes should be compatible with sufficiently low computational costs. The model includes extrinsic muscles —Genioglossus posterior (GGp) and anterior (GGa), Hyoglossus (HG) and Styloglossus (SG)— and intrinsic ones —superior (SL) and inferior (IL) Longitudinalis, Verticalis (V)—. Each muscle is defined by specific sets of elements. Force distribution within each element is determined by the direction of the corresponding muscle fibres. Each element can be shared between many muscles and is able to account for fibres interdigitating.

Figure 1 shows the action of three extrinsic muscles: GGp, HG and SG, respectively mostly recruited for the production of the extreme cardinal vowels [i], [a] and [u].

Finally, the Runge Kutta method is used to solve the equation of motion.

CONTROL OF THE MODEL

Input commands to the physiological models of the tongue are for the most part EMG signals [10], [12], [14]. These models are thus essentially controlled in terms of force level inputs.

Such simulations are very interesting as long as they are only intended to describe the influence of each muscle on the tongue shape. However they seem to be unsuitable for a correct description of tongue control. First of all, as mentioned in the introduction, individual commands to muscles require an additional control layer in order to specify the synergies between muscles (about 20 muscles act together during tongue movement). Moreover, EMG activation, and hence muscular force levels, are the consequences of an interaction between central activation from the CNS to the motoneuron pool, and reflex activation, related in particular to muscle lengths,

from the muscle to the motoneuron pool [15]. EMG activation can therefore not be directly controlled by the CNS.

From this point of view, the Feldman EP Hypothesis is very appealing.

Motor innervation to muscles arises from α MNs which innervate the main body of the muscle and from γ MNs which contribute to α MN excitation through reflexes. The basis of the model is the suggestion that movement arises from changes to neural control variables which shift the equilibrium point of the motor system. The essential control variables are independent changes in the membrane potentials of α and γ motoneurons (MNs) which establish a threshold muscle length (λ) at which the recruitment of MNs begins. As the system changes λ, muscle activation, and hence force, vary in relation to the difference between the actual and threshold muscle lengths. Moreover, due to reflex damping, this activation also depends on the rate of muscle length changes. Thus, by shifting λ by changes to the central facilitation of MNs, the system can produce a movement to a new equilibrium position.

The dependence of active muscle force on muscle activation is approximated by an exponential function, estimated from empirical force-length relations for cat gastrocnemius muscle [16].

The equation of motion for the complete system has the following form :

$$M\ddot{U} + f\dot{U} + K(U,\dot{U},\lambda)U = F(U,\dot{U},\lambda) + P$$

where U, $\dot{U}$ et $\ddot{U}$ are displacement, velocity and acceleration nodes.

M is a global mass matrix and f the global passive damping matrix.

F represents the active muscle force and P the external force, corresponding here to gravity.

K is the stiffness matrix, which determines the inner forces of the FE structure; this matrix accounts for the distribution stiffness amongst elements. For active muscles this stiffness depends on node displacements and velocity as given by the slope of the exponential Force/Activation relationship.

SIMULATIONS

Our aim here is to show that tongue movements measured for the transition between two vowels, can be accounted for, by linear shifts of the control variables of the system, between equilibrium positions close to the actual vowel tongue shapes. To accomplish this, we worked on cineradiographic data showing sagittal vocal tract outlines for a French subject [17].

Usually, normal vowel transitions require both jaw and tongue movements. In order to take into account the influence of jaw displacements on the tongue movements, a simple command specifying the mandible position is introduced. For this, the model is embedded in a complete vocal tract model, where jaw position is specified via the inferior incisive. The jaw command is directly obtained from the data by measuring the position of this incisive.

In order to evaluate the Equilibrium Point Hypothesis applied to the model, we worked on [i-a] transition for which tongue movement, as observed in the sagittal plane, is simple and relevant. Moreover, muscle activity during this transition is essentially due to GGp and HG : GGp was supposed to be the only active muscle for the production of the vowel [i], while HG accounted for the production of the vowel [a].

Starting from the rest position, [i] is obtained by a single shift of the posterior GGp lambda command. The amount of force involved to maintain [i] tongue posture is just sufficient to counteract gravity and inner tongue forces: in this condition, GGp lambda is close to the actual length of the muscle. The same strategy is used for HG lambda in the [a] configuration.

Movement from [i] to [a] is generated by linear shifts of the lambda command for the two muscles mentioned above. The command shift rate is the same for both lambdas and is chosen in order to correctly fit tongue shape variations over time.

Figure 2 plots the movement simulated for the whole sequence. The real [i-a] contours measured on the cineradiographic database are superposed to the simulated ones. The adequacy between simulations and data is quite good. Further analysis are in course in order to compare more precisely the kinematics properties of specific points of the tongue .

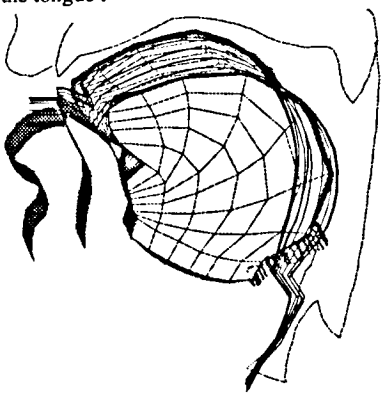


Figure 2 : simulation of [i-a] transition (dotted) and measured contours (dark).

CONCLUSIONS

By elaborating a 2D physiological model of the tongue, integrating basic elasticity and force generation principles, it was possible to propose preliminary tests of Feldman's theory applied to speech production control. This theory allows, with simple commands, a generation of realistic movement in a VV sequence. Moreover, the intended equilibrium positions are consistent with the shape effectively reached for each of the vowels. This feature is interesting to understand the relationships between vocal tract geometry and motor control space.

ACKNOWLEDGEMENT

The authors acknowledge David Ostry for many discussions. This research was supported by the European Union (ESPRIT-BR Project n° 6975), by the Cooperation France-Québec .

REFERENCE

[1] **Stevens K.N. & House A.S. (1955).** Development of a Quantitative Description of Vowel Articulation. *J. Acoust. Soc. Am.*, 27, 484-493.

[2] **Fant G. (1960).** Acoustic Theory of Speech Production., Mouton, The Hague.

[3] **Wood S. (1979).** A Radiographic Analysis of Constriction Locations for Vowels. *J. of Phonetics*, 7, 24-44.

[4] **Maeda, S. (1990).** Compensatory Articulation during Speech: Evidence from the Analysis and Synthesis of Vocal-Tract Shapes Using an Articulatory Model. In W.J. Hardcastle and A. Marchal (Eds.), *Speech Production and Speech Modeling* (pp. 131-149). Dordrecht, The Netherlands: Kluwer Academic Publishers.

[5] **Maeda, S., Honda K., and Kusawaka N. (1993).** From EMG to Vowel Formant Patterns. *3rd Seminar on Speech Production: Models and Data.* Old Saybrook, CT., 11-13 May.

[6] **Jordan, M.I. (1990).** Motor Learning and the Degrees of Freedom Problem. In M. Jeannerod (Ed.), *Attention and Performance* (Chapter XIII). Hillsdale, NJ: Erlbaum.

[7] **Feldman, A.G. (1966).** Functional Tuning of The Nervous System with Control of Movement or Maintenance of a Steady Posture – II Controllable Parameters of the Muscles. *Biophysics, 11*, 565-578.

[8] **Laboissière R., Ostry D.J, and Perrrier P. (1995).** *In these Proceedings.*

[9] **Laboissière R., Ostry D.J, and Feldman A.G.** (submitted). The Control of a Multi-Muscle Systems: Human Jaw and Hyoid Movements *J. of Neurophysiology.*

[10] **Perkell, J.S. (1974).** *A Physiologically-Oriented Model of Tongue Activity in Speech Production.* Ph.D. Thesis. Boston, Ma.: Massachussetts Institute of Technology.

[11] **Kiritani S., Miyawaki, K., and Fujimura, O. (1976).** A Computational Model of the Tongue. *Annual Bulletin, 10* (pp. 243-252). Tokyo: Research Institute of Logopedics and Phoniatrics, University of Tokyo.

[12] **Kakita, Y., Fujimura, O., and Honda, K. (1985).** Computation of Mapping from Muscular Contraction Patterns to Formant Patterns in Vowel Space. In V.A. Fromkin (Ed.), *Phonetic Linguistics* (pp. 133-144). Orlanda, Florida: Academic Press.

[13] **Hashimoto K. and Suga S. (1986)** Estimation of the muscular tensions of the human tongue by using a three-dimensional model of the tongue. *Journal of Acoustic Society of Japon* (E) 7,1,39-46.

[14] **Wilhelms-TricarioR.**(in press) Physiological Modeling of Speech Production: Methods for Modeling Soft-Tissues Articulators. *J. Acoust. Soc. Am.*.

[15] **Feldman A. (1986).** Once more on the Equilibrium-Point Hypothesis for motor control", *Journal of Motor Behavior,Vol. 18,n°1.*

[16] **Feldman A. and Orlovsky G.N. (1972).** The influence of different descending systems on the tonic reflex in the cat. *Experimental Neurology* 37:481-494.

[17] **Badin P., Gabioud B., Beautemps D., Lallouache T.M., Bailly G., Maeda S., Zerling J.P.& Brock G.(1995).** Cineradiography of VCV sequences : articulatory-acoustic data for a speech production model. *15th International Congress on Acoustics,* Trondheim, Norway.

MODAL ANALYSIS OF ACOUSTIC WAVE PROPAGATION IN THE VOCAL TRACT USING A FINITE-DIFFERENCE SIMULATION

Gordon Ramsay and Li Deng
Dept. of Electrical & Computer Engineering, University of Waterloo, Canada.

ABSTRACT

A time-domain simulation of wave propagation in the vocal tract is outlined, using finite-differences to map a system of partial differential equations onto a state-space recursion. It is shown that the eigenvalues and eigenvectors of the time-varying system matrix can be used to determine the resonant modes of the vocal tract at any desired time instant. The effect of glottal vibration on formant structure is examined using this technique.

INTRODUCTION

The correspondance between the formants of the speech signal and the resonant modes of the vocal tract is well known, as is the role of the spatial pressure and volume-velocity distributions associated with each mode in determining the response of the vocal tract to an acoustic source located at any point within the tract. Previous attempts to determine the spatial and temporal modes using acoustic models [1] [2] have relied largely on frequency-domain methods which are limited to static tract shapes, and which do not account for coupling between the glottis and the supra-glottal cavities. In this paper, a time-domain simulation method is described which allows the resonant modes of the modelled vocal tract to be determined at every sample point. The method relies on the use of a finite-difference approximation to convert a system of acoustic partial differential equations into state-space form. The eigenstructure of the resulting matrix recursion determines the spatial and temporal modes of the vocal tract, and can be calculated at each time instant. Since the model is valid for time-varying glottal and supra-glottal area functions, it can be used to examine the effect of fast variations in glottal shape on the formants. This is difficult to investigate using spectral analysis techniques due to the short time scales involved, as reported in [3][4].

THE ACOUSTIC MODEL

The acoustic model adopted here is based on a previous model due to Maeda [5]. Define a bounded region $\Omega \subset \mathcal{R}^2$ to represent all points (x,t) in the vocal tract at distance x from the trachea entrance at time t. Ω is divided into three sub-domains Ω_1, Ω_2, Ω_3 representing the trachea, glottis, and oral cavity respectively, which are separated and enclosed by boundaries $\partial\Omega_{t0}$, $\partial\Omega_{tT}$, $\partial\Omega_{x0}$, $\partial\Omega_{xL}$, $\partial\Omega_{12}$, $\partial\Omega_{23}$.

$$
\begin{aligned}
&\Omega = \{(x,t) : 0 \le x \le L(t), 0 \le t \le T\},\\
&\Omega_1 = \{0 < x < X_T\},\\
&\Omega_2 = \{X_T < x < X_G\},\\
&\Omega_3 = \{X_G < x < L(t)\},\\
&\partial\Omega_{t0} = \{t = 0\},\ \partial\Omega_{tT} = \{t = T\},\\
&\partial\Omega_{x0} = \{x = 0\},\ \partial\Omega_{xL} = \{x = L(t)\},\\
&\partial\Omega_{12} = \{x = X_T\},\ \partial\Omega_{23} = \{x = X_G\},
\end{aligned}
$$

$L(t)$ is the time-varying tract length; X_T and X_G are the distances from the trachea entrance to the entrance and exit of the glottis, assumed constant.

Suppose that the vocal tract may be approximated by a non-uniform time-varying elastic tube of equivalent circular cross-section $A(x,t)$ and circumference $S(x,t)$. Under the usual assumptions that the processes governing fluid flow are laminar and isentropic, the state of the air in the vocal tract can adequately be described by the pressure $p(x,t)$ and volume velocity $u(x,t)$ at all points along the mid-line, and the system of linearized acoustic equations governing one-dimensional planar wave propagation can then be derived from

conservation of mass and momentum. Additional losses must be included to account for viscous friction and wall vibration. Assuming that the wall surfaces are locally-reacting, the displacement $y(x,t)$ from an equilibrium radius can be conveniently modelled as the response of a second-order linear mechanical system to local pressure variations. The equation for Poiseuille flow in a cylindrical duct can be used to provide the viscous loss term. The system of partial differential equations for p,u,y in Ω_1 and Ω_3 is then given by:

$$
\begin{aligned}
\frac{\partial p}{\partial x} + \frac{\partial}{\partial t}\left(\frac{\rho u}{A}\right) + \frac{8\pi\mu}{A^2}u &= 0,\\
\frac{\partial u}{\partial x} + \frac{\partial}{\partial t}\left(\frac{Ap}{\rho c^2}\right) + \frac{\partial}{\partial t}(A + Sy) &= 0,\\
m\frac{\partial^2 y}{\partial t^2} + b\frac{\partial y}{\partial t} + ky - Sp &= 0.
\end{aligned}
$$

Flow within the glottis is assumed incompressible, and wall vibration may be neglected, leading to the following modified system of equations for Ω_2,

$$
\begin{aligned}
\frac{\partial p}{\partial x} + \frac{\partial}{\partial t}\left(\frac{\rho u}{A}\right) + \frac{12\mu l_g^2}{A^3}u &= 0,\\
\frac{\partial u}{\partial x} &= 0,
\end{aligned}
$$

where l_g is the glottal length and ρ, μ the density and viscosity of air.

The boundary condition at the trachea entrance $\partial\Omega_{x0}$ is provided by the lung pressure $P_{lung}(t)$.

The boundary condition for $\partial\Omega_{xL}$ at the lips is supplied by the radiation impedance, which can be modelled using Flanagan's approximation,

$$
\frac{\partial u}{\partial t} - \frac{9\pi^2}{128\rho c}\frac{\partial}{\partial t}(Ap) - \frac{3\pi\sqrt{\pi A}}{8\rho}p = 0.
$$

Continuity is assumed across the internal boundaries $\partial\Omega_{12}$ and $\partial\Omega_{23}$, and all quantities assume their equilibrium values initially along $\partial\Omega_{t0}$.

The mixed initial/boundary-value problem must now be solved on Ω using numerical methods. Applying the finite-difference technique, the continuous domain Ω is sampled on a grid of points $\{(x_j,t_k) : j = 1 \ldots M, k = 1 \ldots N\}$, which need not be uniform. The partial derivatives in the original continuous-domain equations are replaced by difference operators, yielding simultaneous linear algebraic equations linking the quantities p, u, y sampled at neighbouring grid points over several time steps. In this way, the continuous-domain equations are translated into a system of linear difference equations defining a recursion on a finite-dimensional state-space, which can be solved to yield an approximation to the true solution. The solution of the discretized system will converge to a solution of the original continuous-domain system if it can be shown that the discretization is *consistent* and the recursion is *stable*. A two-level implicit difference scheme has been carefully constructed from the above equations to guarantee convergence for a slowly time-varying grid defined on Ω. The details are omitted here due to lack of space.

Denoting by Z_k the vector of values for $p(x_j,t_k)$, $u(x_j,t_k)$, $y(x_j,t_k)$, $y'(x_j,t_k)$ on all grid points at time t_k, and taking $Z_0 = 0$, the resulting implicit recursion may be written as

$$
P_k Z_{k+1} = Q_k Z_k + F_k
$$

where P_k, Q_k are sparse banded matrices whose coefficients are functions of $A(x,t)$ determined by the difference scheme, and F_k is a vector driving function derived from the boundary condition on $\partial\Omega_{x0}$.

This is a standard generalized eigenvalue problem, and the properties of the recursion are clearly entirely determined by the eigenstructure of the matrices $P_k^{-1}Q_k$. In particular, at any time t_k, the solution Z_k of the recursion can be expressed as a modal sum involving the eigenvalues λ_k^i and eigenvectors ϕ_k^i of $P_k^{-1}Q_k$.

Although the original equations do not possess a well-defined system of eigenfunctions due to the time-varying tract length, the finite-difference scheme can be shown to approximate the original PDEs in the

limit as the grid dimensions tend to zero, and the changing eigenstructure of the corresponding matrix recursion can be taken to represent a "local" approximation to the evolution of the vocal tract resonant modes. The eigenvalues of the system matrix represent the time-varying poles of the vocal tract, and can be used to calculate the formants and their bandwidths, while the eigenvectors represent the changing spatial distributions of pressure and volume velocity associated with each formant.

By examining the modal structure of matrices $P_k^{-1}Q_k$, calculated for any particular time-varying area function $A(x,t)$ on Ω, it is therefore possible to arrive at a complete characterization of the behaviour of the modelled vocal tract for every discrete time sample.

EXPERIMENTAL RESULTS

As a useful application of this technique, consider the dynamic changes in glottal shape which occur during phonation. It is well known that formant motion occurs during the glottal cycle, but previous investigations [3][4] have found difficulties in determining the exact nature of this effect using conventional spectral analysis techniques.

The modal analysis method circumvents some of these problems. Figure 1 shows the variation in frequency and bandwidth for the first three formants, calculated every 0.2ms during two glottal cycles over a period of 20ms for the vowel /a/, together with the associated normalized modal pressure and volume velocity distributions. The area function was generated from an articulatory model; 6 grid points were used for the trachea, 7 in the glottis, and 79 for the oral tract, and the glottal section areas were assumed to execute co-phasic sinusoidal oscillations about a slightly-abducted rest position.

During each glottal period, the vocal tract poles were found to execute "teardrop-shaped" movements in the complex plane, with the path shape depending on the area function. The formant frequencies appear to increase monotonically with glottal area, as found in [3][4], but the effect on the bandwidths is more complicated, with positive and negative excursions. The modal distributions when the glottis is closed are similar to those calculated in [1][2], and the distortion that occurs due to glottal opening is most pronounced in the lower formants.

CONCLUSIONS

A time-domain technique for simulating the time-varying resonant modes of the vocal tract has been described, and used to examine the effect of rapid changes in glottal shape on modelled formant positions. Results derived by previous authors [3][4] using spectral analysis have been largely confirmed, but the calculation method applied here is somewhat more reliable, and clarifies the precise movement of the tract modes during simulation. It remains to be verified, however, whether the linear acoustic model is indeed a valid approximation to real speech production.

REFERENCES

[1] Fant G., Pauli S., (1975) "Spatial characteristics of vocal tract resonance modes," *Proc. Speech Communication Seminar, Stockholm 1974*, pp. 121–132.

[2] Mrayati M., Carré R., (1976) "Relations entre la forme du conduit vocal et les caractéristiques acoustiques des voyelles françaises," *Phonetica 33* pp. 285–306.

[3] Cranen B., Boves L., (1987) "Spectral consequences of a time-varying glottal impedance," *Proc. International Congress of Phonetic Sciences, 1987*, pp.361–366.

[4] Meyer P., Strube H.W., (1984) "Calculations on the time-varying vocal tract," *Speech Communication 3* pp. 109–122.

[5] Maeda S., (1982) "A digital simulation method of the vocal-tract system," *Speech Communication 1* pp. 199–229.

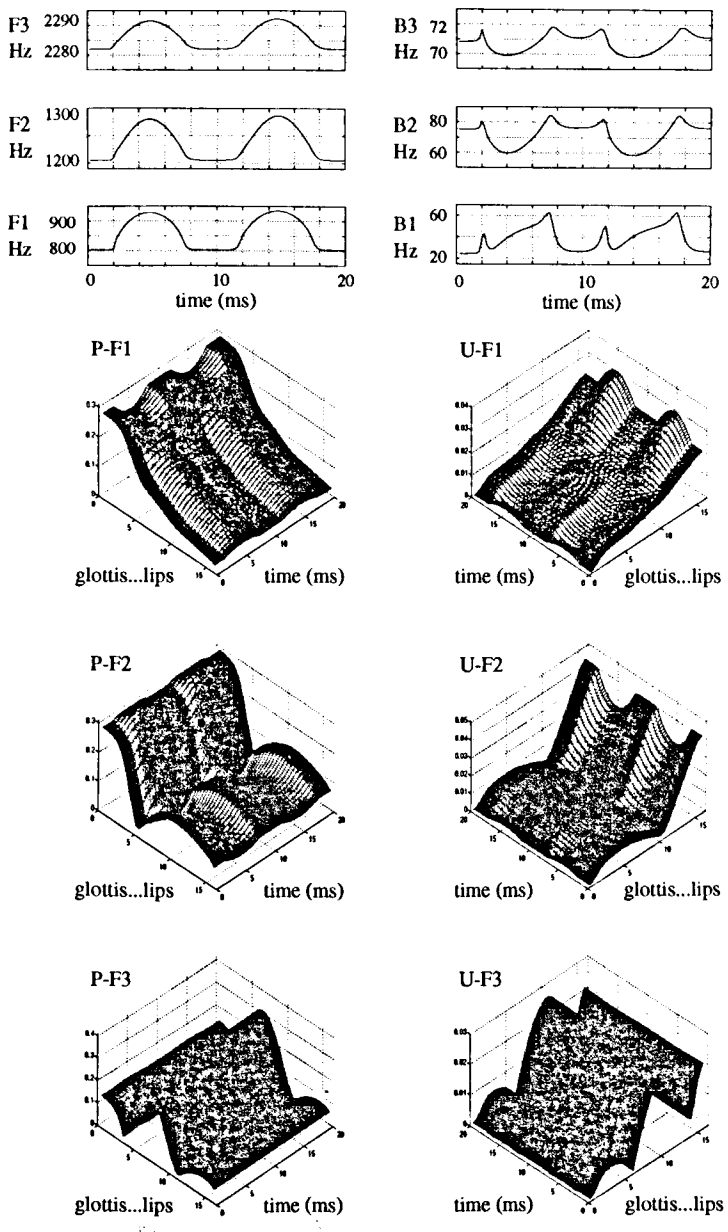


Figure 1 : Formant frequencies, bandwidths, and pressure/volume-velocity distributions for /a/.

ANALYSIS AND ARTICULATORY SYNTHESIS OF DIFFERENT VOICING TYPES

C. Scully, K. Stromberg & D. Horton (1), P. Monahan, A. Ni Chasaide & C. Gobl (2)
(1) University of Leeds, Leeds, England
(2) Trinity College, Dublin, Ireland

ABSTRACT

Results of inverse filtering from both airflow and sound pressure signals for a real speaker have been combined to develop the Leeds phenomenological model of voicing. Three controlling physiological variables are mapped onto each of three voice source waveform parameters for nine voicing types, with the aim of simulating variations of the voice source with phonetic context.

INTRODUCTION

The Leeds phenomenological model of voicing relates voice source waveshapes to the physiological conditions which generate it, for a particular speaker. A three-parameter representation of the voice source developed by Fant is used [1]. The approach has been described elsewhere [2]. Previously, voice source waveforms for natural speech were obtained by inverse filtering the total volume flowrate of air through the mouth and nose combined, Ut. This method has the advantage of giving an estimate for the dc flow through a glottal chink; but has the disadvantage that one of the acoustic source parameters, the asymmetry factor K, is difficult to estimate.

Here the same kind of real speech data were obtained but for the inverse filtering two signals were used: simultaneously recorded airflow Ut and sound pressure Sp. Robust features from each method were combined, with the aim of improving the reliability of the voice waveshape description.

Three physiological controlling parameters are used, closely related to those for the two-mass model of voicing [3]. They are the low frequency component of the glottal area AG, the pressure drop across the glottis PDIFF and the mass-tension factor for the vocal folds Q.

METHODS

Subject

The main male speaker for the SPEECH MAPS project, PB, who has a general French accent without Southern French features, provided the speech data.

Signals

An undivided Rothenberg mask and an orally inserted pressure tube with a Gaeltec pressure transducer were used, combined with a B & K condenser microphone outside the mask and a laryngograph. Four channels of data were recorded at Leeds directly onto a pc with a sampling frequency of 10 kHz: sound pressure Sp, laryngograph signal Lx, total output airflow Ut and intra-oral air pressure Po.

Speech Material

Corpus G of the SPEECH MAPS project is central to the Leeds voicing model. It contains multiple repetitions on a single expiratory breath of [pɑ] for nine voicing types. These are three levels: medium, soft, loud; three pitches: mid, high, low; three phonation types: normal, breathy pressed. The aim is to describe a range approximating to the normal limits of the voice source for speech. The mid-open, central quality vowel is chosen with the aim of achieving an approximation to the following conditions: [ɑ] has a fairly high F1 frequency which is suitable for inverse filtering; the jaw movement down from that for [p] to that for the vowel is not large, so that large changes in vocal tract volume are avoided. The vowel is made rather long, so that a quasi-static jaw and tongue configuration can be assumed near mid-vocoid, where the analyses are performed. It is hoped that, as a result, the errors introduced by assuming that output airflow is a good approximation to transglottal airflow when estimating glottal area and when performing inverse filtering from airflow are minimised.

The [p-p] context is needed for estimates of subglottal pressure as described below. The speaker tried to avoid pitch changes and other phonetic exponents of stress, with the aim of keeping subglottal pressure changes as small as possible. The multiple repetitions, about 10 to 11 for each voicing type, permit multiple analyses, with only lung volume changing from one repetition to the next.

Voiced fricative sequences from multiple repetitions on one expiratory breath of [paCa] were used also, from Corpus 3 of the SPEECH MAPS project.

ANALYSES

Four repetitions of each of the nine voicing types were analysed in several ways at the same mid-vocoid time point. The four channel data files were converted to ESPS "Waves" format. Signals were displayed using this software and the ESPS programs were modified at Leeds according to the analyses required. The signals Ut and Po were low-pass filtered for the aerodynamic analyses described in this section.

Estimation of the controlling parameters

Subglottal pressure PSG was assumed to equal intra-oral air pressure Po at the end of the initial rapid rise in Po during the [p] closures. Linear interpolation gave an estimate for PSG at mid-vocoid.The pressure drop across the glottis PDIFF was calculated as (PSG - Po).

Glottal area AG was estimated at mid-vocoid by assuming that the transglottal flow UG was equal to Ut here. The orifice equation [4] was used:

$$AG = k.UG/(\sqrt{PDIFF}) \quad (1)$$

(where k = 0.00076 with AG in cm^2, UG in cm^3/s, PDIFF in cm H_2O).

Complexity of fundamental frequency F0 patterning is derived by assuming that a myoelastic component Q and an aerodynamic component PDIFF together control F0 from:

$$F0 = Q + KF.PDIFF \quad (2)$$

KF is an empirical constant to be determined for a given speaker. For some of the voiced fricatives of Corpus 3, PSG and PDIFF were estimated as described above. KF values were calculated as dF0/dPDIFF, with a mean value of 4.1 Hz/cmH_2O for speaker PB. Q was obtained from F0 by equation 2.

Estimation of the voice source parameters

Inverse filtering was done using methods and software developed at Trinity College Dublin [5]. Three successive cycles of voicing were used, at or very close to the predefined mid-vocoid time point. Total output airflow Ut, unfiltered, was used to estimate total volume flowrate of air through the glottis Ut,g with its acoustic component Ug, and dc flow also. Simultaneously recorded Sp gave the differentiated glottal flow Ug' (dUg/dt in Figure 1). The number of pairs of conjugate complex poles the program was to find was set at 6. For a sampling frequency of 10 kHz the number is normally 5 for a male speaker; however, an extra pair was added, on the basis that the flow frequency response of the Rothenberg mask needed to be taken into account (see [6]).

The Leeds voice waveshape, based on Ug, is not mathematically equivalent to the LF model of voicing [7], based on Ug'. But comparisons of pairs of

waveforms showed that some time points could be aligned; features of the Ug′ wave could be used to enhance the analysis of the Ug wave. Figure 1 shows both waveshapes. The LF model has four parameters in addition to T0 while the Leeds model has only three; it lacks a return phase after excitation. Three robust parameters were used here; they are the same three shown to characterise a speaker's voice using the LF model [8]. In our nomenclature they are VOIA, EE and T0. The three acoustic parameters required to define the voice source waveform in the Leeds model are: VOIA, the amplitude of the acoustic component of flow; TCR, the ratio of closed time over total periodic time (TCR = 1 - open quotient) and K, an asymmetry factor. The three parameters were obtained as follows :

1. The dc flow level was drawn as an averaged value where Ug was low. This is the 'closed' phase, which does not necessarily mean complete closure; indeed usually it does not.
2. VOIA was defined from the dc flow level to the peak flow. Any ripples remaining in the Ug trace were smoothed by eye.
3. For TCR, the start and end of the closed phase were both defined from Ug′. The time point for EE (the maximum negative value of Ug′) defined the start of the closed phase. The start of the rise of Ug′ up from its zero line defined the end of the closed phase. This time point was difficult to locate, whether Ug or Ug′ was used. T0 was obtained from EE points in successive cycles of Ug′.
4. In our previous work, the asymmetry factor K was measured from the gradient of Ug half way up the rising portion to give TB; the gradient at closure as seen on Ug gave TD. K was calculated [1] from the equation :

$$K = 0.5 + 0.125(TB/TD)^2 \quad (3)$$

This was very difficult to do with any degree of confidence. Different measurements were made here, to approximate to this formula for K, as follows:
On Ug′, EI is the maximum positive value of the gradient of Ug.

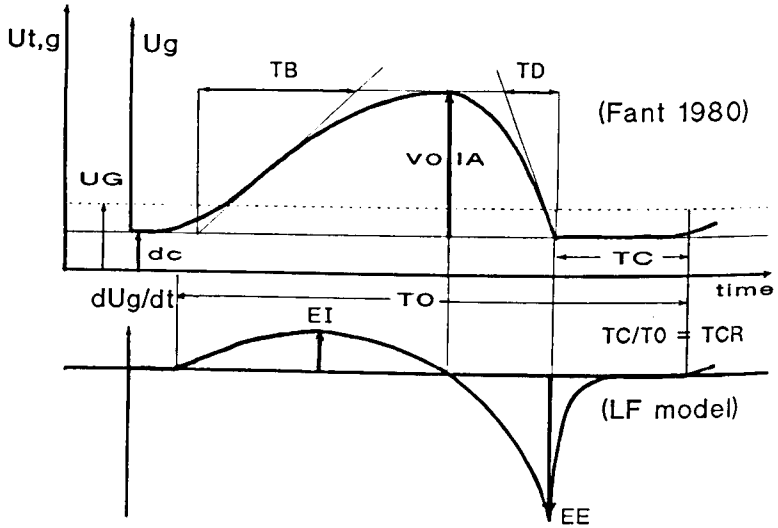


Figure 1. Waveshape representations of the voice source: flow Ug above, differential of flow Ug′ below.

It was found to be located in time quite near half way up the rising portion of Ug, so approximately:

$$TB = VOIA/EI \quad (4)$$

Similarly, on Ug′, the absolute value of EE gave an approximation to the (negative) gradient for Ug near 'closure' in the Leeds model, so approximately:

$$TD = VOIA/|EE| \quad (5)$$

So $$K = 0.5 + 0.125(|EE|/EI)^2 \quad (6)$$

CONSTRUCTION OF THE VOICING MODEL

Stepwise multivariate regression on the four repetitions of the nine voicing types was used to obtain the relationship between each voice waveshape parameter, the dependent variable, and the three physiological, independent controlling variables. The three equations obtained were :

$$VOIA = -197.00 + 68.80\ PDIFF + 982.00\ AG + 1.49\ Q \quad (7)$$

$$TCR = +0.33 + 0.03\ PDIFF - 0.65\ AG \quad (8)$$

$$K = +4.08 - 0.10\ PDIFF - 2.49\ AG - 0.02\ Q \quad (9)$$

In addition to these mapping equations, upper and lower limits were set for the voice source parameters, based on the data for speaker PB.

SIMULATIONS OF THE SPEAKER

The Leeds composite forward model of speech production is being used to simulate speaker PB's productions. Comparisons between the model and the natural speech are made for articulatory paths, aerodynamics, acoustic sources and output speech signal, real or synthetic. As a first step, the adequacy of the voicing model can be assessed by simulating vowels produced with different voicing types. The power of the voicing model to go beyond the vowel data on which it is based is being investigated with sequences containing voiced fricatives.

Values for the dc flow through a glottal chink, obtained here for the nine voicing types, have been related to UG and the voice source waveform parameters [2]; the current analyses confirm the findings.

ACKNOWLEDGEMENTS

This work was partially funded by an ESPRIT basic research project, no. 6675, SPEECH MAPS. We thank the speaker, our colleague Pierre Badin.

REFERENCES

[1] Fant, G. (1980), "Voice source dynamics", *STL-QPSR*, Stockholm 2-3, pp.17-37.
[2] Scully, C. and Stromberg, K. (1992), "Physiologically-controlled voice source models for different speakers", *Proc. Inst. of Acoustics*, 14, pp.463-471.
[3] Ishizaka, K and Flanagan, J. L. (1972), "Synthesis of voiced sounds from a two-mass model of the vocal folds", *Bell Syst Tech*, 51, pp.1233-1268.
[4] Scully, C. (1986), "Speech production simulated with a functional model of the larynx and the vocal tract", *J Phonetics*, 14, pp. 407-414.
[5] Ni Chasaide, A., Gobl, C. and Monahan, P. (1992), " A technique for analysing voice quality in pathological and normal speech", *J Clinical Speech and Language Studies*, Vol. 2, pp. 1-16.
[6] Hertegard, S. and Gauffin, J. (1992), "Acoustic properties of the Rothenberg mask", *STL-QPSR* 2-3, pp. 9-18.
[7] Fant, G. and Liljencrants, J. and Lin, Q. (1985), "A four parameter model of glottal flow" *STL-QPSR* Stockholm, 4, pp. 1-13.
[8] Fant, G., Kruckenberg, A., Liljencrants, J. and Båvegard, M. (1994), "Voice source parameters in continuous speech: transformation of LF parameters", *Proc. ICSLP-94*, Yokohama, Vol. 3, pp. 1451-1454.

A REGRESSION ANALYSIS OF THE RELATION BETWEEN PHYSIOLOGICAL SIGNALS AND F_0

Helmer Strik

University of Nijmegen, Dept. of Language and Speech
P.O. Box 9103, 6500 HD Nijmegen, The Netherlands
E-mail: strik@let.kun.nl

ABSTRACT

Measurements were obtained of several physiological mechanisms which are known to be important in the control of fundamental frequency (F_0). The data were analysed by means of a multiple regression analysis in which F_0 is the criterion and the physiological signals are the predictors. Separate analyses were carried out for statements and questions, and for falling and rising F_0. The results reveal no considerable differences in the control of F_0 for the various datasets.

1. INTRODUCTION

In the literature different views are expressed regarding the relation between F_0 and the underlying physiological signals. The goal of the present research was to clarify this relation. Research on the relation between F_0 and the physiological processes is very complicated, if only because F_0 is dependent on a large number of physiological mechanisms [1]. Moreover, direct measurements of laryngeal physiology are by necessity invasive. Since these measurements are difficult to make, only a small amount of data is usually available.

To study the relation between F_0 and the physiological signals, it seems advisable to use a quantitative analysis method. However, in most of the studies on this topic a kind of qualitative analysis is used. Two notable exceptions are [2] and [3]. Due to space limitations, it is not possible to go into the details of these two studies. Therefore, only the most important drawbacks of these studies are briefly presented here.

Both in [2] and [3] the total number of samples for which the quantitative analysis is done, is very small (i.e. 568 and 106, respectively). In these two studies analyses were also performed for subdivisions of the data. In these cases the number of data is even smaller. Another drawback of [2] is that only correlation coefficients were calculated, and no regression equations. The reason why this is a drawback will be explained below. In [3] regression equations are presented, but in this study sustained phonation was used. It is not unlikely that the relations between F_0 and the physiological signals in sustained phonation are different from the relations in running speech, as was already suggested in [3]; especially, because the F_0 values found in [3] are very high (i.e. much higher than F_0 values which are usually found in running speech).

In the current study measurements of physiological signals were made while subjects produced meaningful Dutch sentences. Our intention was to obtain a large amount of data, in order to have sufficient samples for the regression analysis.

2. MATERIAL AND METHOD

For two Dutch male subjects (LB and HB) recordings were made of the audio signal, electroglottogram, lung volume, subglottal pressure (P_{sb}), and the electromyographic activity of two laryngeal muscles: sternohyoid (SH) and vocalis (VOC). In addition to these signals, the activity of the cricothyroid (CT) muscle was also measured for subject LB, and oral pressure (P_{or}) for subject HB. The measurements were made while the subjects produced meaningful Dutch sentences with different intonation patterns. Each sentence was repeated 5 to 8 times. The signals of these repetitions were used to calculate average signals for every sentence. A more elaborate description of the experiments, and figures of the measured signals can be found in [1]. Here only those aspects are mentioned which are most relevant to the present article.

All signals were sampled at a 200 Hz rate, and were then smoothed. The muscle signals were shifted forward in time by their mean response time (as described in [2]). Only the voiced frames of the utterances were used in a stepwise multiple regression analysis (SMRA). In the SMRA the dependent variable (the criterion) is F_0, and the measured physiological signals are the independent variables (the predictors):

$F_{0,est} = C_0 + C_1*P_{sb} + C_2*SH + C_3*VOC [+ C_4*X_4]$.

The fourth term in the regression equation (X_4) is only used once for each subject (see section 3.1). In that case X_4 is different for the two subjects, i.e. P_{or} for HB and CT for LB.

For different datasets correlation coefficients and regression equations were calculated. Furthermore, for each regression coefficient the standard error and the t-value were also computed. The t-values were used to check the statistical significance of the regression coefficients, while the standard errors were used to round off the regression coefficients to their last significant digit.

Table 1. Results of SMRA for all data of subjects LB and HB. Shown are, from left to right, the regression coefficients C_i, the multiple correlation coefficient (MR), the number of datapoints (N), the identification of the regression equation (subject + number), and a brief description of the data.

C_0	C_1	C_2	C_3	C_4	MR	N	id.	description
67.9	2.9	0.061	0.378		0.886	2319	HB1	all data
67.1	3.2	0.063	0.376	-1.1	0.887	2319	HB2	all data
68.5	3.5	-0.21	0.444		0.836	2254	LB1	all data
70.7	2.7	-0.16	-0.01	0.51	0.896	2254	LB2	all data

3. RESULTS

3.1. All data

First of all, regression equations were calculated for all data of both subjects. The results are given in Table 1, and the correlation coefficients in Table 2. A comparison of the regression equations HB1 and LB1 reveals that C_0 (the constant term), C_1 (the F_0-P_{sb} ratio) and C_3 (the F_0-VOC ratio) do not differ much between these subjects. However, their C_2's (the F_0-SH ratio) are different. In most studies (see the references given in [1]) a negative relation between F_0 and SH is found. The results of subject LB are in line with this general finding, but the results for HB are not.

Apart from P_{sb}, SH and VOC, other physiological signals were measured for these subjects. For subject HB oral pressure (P_{or}) was also measured. The correlations of P_{or} with F_0 are very small (for all 2319 voiced frames of HB the correlation is 0.011). Consequently, adding P_{or} to the regression equation does not have much influence. The resulting regression equation HB2 is almost equal to the regression equation HB1.

For subject LB the activity of the cricothyroid muscle (CT) was also measured. The correlations of CT with F_0 are very high (for all 2254 voiced frames of subject LB it is 0.859). In fact, the correlation of CT with F_0 is larger than any of the other correlations with F_0 (see Table 2, row LB1). This is in accordance with what is usually found (see e.g. [2, 3]). The correlation between CT and VOC is 0.900 for all 2254 voiced frames. A high correlation between CT and VOC was also found by [2, 3]. Therefore, it seems that VOC acts in synergy with CT in the control of F_0.

For subject LB the CT was added to the regression equation, and the result is equa-

Table 2. Correlations of F_0 with Psb, SH and VOC for different subsets of the data.

Psb	SH	VOC	id.	description
0.333	0.351	0.872	HB1	all data
0.452	-0.404	0.760	LB1	all data
0.501	0.424	0.846	HB3	statements
-0.167	0.178	0.921	HB4	questions
0.594	-0.423	0.705	LB3	statements
0.167	-0.351	0.863	LB4	questions
0.191	0.404	0.834	HB5	falls
0.307	0.448	0.872	HB6	rises
0.601	-0.450	0.686	LB5	falls
0.320	-0.364	0.825	LB6	rises

Table 3. Results of SMRA for different subsets of the data. For explanation see Table 1.

C_0	C_1	C_2	C_3	MR	N	id.	description
67.8	3.5	0.079	0.334	0.871	1624	HB3	statements
64.7	3.0	0.04	0.434	0.925	695	HB4	questions
70.9	3.7	-0.26	0.42	0.808	1542	LB3	statements
58.7	3.7	-0.07	0.477	0.901	712	LB4	questions
78.1	1.7	0.07	0.325	0.848	586	HB5	falls
77.6	1.5	0.11	0.360	0.880	623	HB6	rises
61.8	4.8	-0.22	0.38	0.825	1101	LB5	falls
75.1	2.6	-0.21	0.46	0.873	484	LB6	rises

tion LB2 in Table 1. By comparing equation LB1 and LB2 it becomes clear that adding the CT has an enormous influence on the resulting regression equation. First of all, the multiple correlation increases substantially. Second, and more important, the magnitude of all regression coefficients changes. The reason why the changes are so considerable, is that the different variables are not orthogonal. This is certainly the case for CT and VOC. Consequently, a large part of the variance of F_0 that is explained by the VOC in equation LB1, will be explained by the CT in equation LB2. In equation LB2 C_3 (the F_0-VOC ratio) even becomes negative, while it is clear that F_0 and VOC are positively related.

This is an obvious disadvantage of regression equations. If the variables are not orthogonal, which is usually the case for physiological signals, the results of regression equations should be interpreted with caution.

Since P_{sb}, SH and VOC are the signals which were measured for both subjects, only these variables will be used in the rest of this article. Adding an extra variable (especially CT) does increase the amount of explained variance, but makes it impossible to compare the data between subjects. Because CT and VOC have similar effects on F_0, it is not so important which of the two variables is chosen.

After having calculated regression coefficients for all the data of both subjects, regression coefficients were computed for different subdivisions of the data: statements vs. questions, and falling vs. rising F_0. Similar subdivisions were made in [2], which makes it possible to compare the results of [2] with those of this study.

3.2. Statements and questions

In [2] the most striking differences between statements and questions were observed for the correlation of F_0 and P_{sb}. In statements it was positive, while in questions it was negative. The same effect can be observed for subject HB (see Table 2, compare rows HB3 and HB4). For subject LB the correlation of F_0 and P_{sb} is still positive for the questions, but it is much smaller than that for the statements (see Table 2, compare rows LB3 and LB4).

Although there are substantial differences between the correlations of statements and questions (also for the other variables, see Table 2), it can be observed in Table 3 (compare HB3 with HB4, and LB3 with LB4) that the differences between the regression coefficients are not so large. In other words, the regression equations reveal that the relations between F_0 and the physiological signals for statements and questions do not differ much. This is an example of an advantage of regression analysis compared to simple correlation analysis. Even if the relations among the variables are almost the same (i.e. the regression coefficients are almost the same), the correlation coefficients can have very different values depending on the kind of data used (e.g. statements vs. questions).

3.3. Falling and rising F_0

In the section above, the data were divided into statements and questions. In this section the data will be subdivided in terms of falling and rising F_0. Samples with a negative derivative are classified as falls, and samples with a positive derivative as rises. The same method was also used in [2], which makes it possible to compare the results.

For subject HB the correlation between F_0 and P_{sb} is different for falls and rises (see Table 2, compare rows HB5 and HB6), whereas the other correlations and the regression coefficients are very similar (see Table 3, compare rows HB5 and HB6). For subject LB larger differences between falling and rising F_0 can be observed, both for the correlations and the regression coefficients (see Tables 2 and 3, compare rows LB5 and LB6). In [2] the largest difference between rises and falls was also found for the correlation between F_0 and P_{sb}, as in our data. For the other correlations no substantial differences were observed in [2] (except for a difference for the correlation of F_0 and the lateral crico-arytenoid muscle).

In short, differences between falls and rises are observed in the correlations of F_0 and P_{sb} for all subjects, and in the regression coefficients of subject LB. In the latter case those differences are particularly evident for C_1 (the F_0-P_{sb} ratio).

4. DISCUSSION

In this article I have presented the results of a quantitative analysis of the relation between F_0 and some physiological mechanisms that are known to be important in the control of F_0. First of all, it is important to note that (apart for the coefficients for P_{or}) all correlation and regression coefficients are highly significant, reflecting the consistent relations among the variables. This was also found in [2] and [3].

The analysis results for all data show that the effect of the SH on F_0 was different for the two subjects, but for the other variables no major differences were found. The variables showing the highest correlations with F_0 were CT and VOC. The correlations of F_0 with P_{sb} and SH were always smaller.

Substantial differences were found between the correlation coefficients calculated for statements and questions, but the differences in the regression coefficients were not very large. Comparing the analysis results for falls and rises revealed that there were differences in the correlations of F_0 and P_{sb} for all subjects, as well as in the regression coefficients of subject LB (especially for C_1, the F_0-P_{sb} ratio). Whether these differences should be interpreted as large, remains questionable. More research is needed to give a definite answer to this question. For the time being, my interpretation of the results is that the relation between F_0 and the physiological signals in statements and questions, and in falls and rises is not very different.

The advantages of the present study, compared to [2] and [3] are, that the number of samples is much larger, that the measurements were obtained for running speech, and that besides correlation coefficients also regression equations were calculated. As mentioned above, regression coefficients are sometimes preferable to correlation coefficients. The reason is that for some subsets of the data the correlations are very different, while the regression coefficients (and therefore probably the underlying relations) are very similar. However, when the variables are not orthogonal, one should also be careful in interpreting the results of regression equations.

In the regression analyses carried out in this study, the physiological signals were used as independent variables (the predictors). Given that no explicit model was used, the implicit assumption made by using this analysis method is that the relation between F_0 and the physiological signals is linear. However, it is almost certain that this relation is not linear. For a more realistic modelling of the relation between F_0 and the physiological processes a production model is needed in which not only the vocal tract but also the voice source is modelled in a physiologically meaningful way. At the moment, a model of this kind does not exist. More research is needed to develop and test such models.

REFERENCES

[1] Strik, H. (1994) *Physiological control and behaviour of the voice source in the production of prosody.* Ph.D. thesis, University of Nijmegen.

[2] Atkinson, J.E. (1978) Correlation analysis of the physiological features controlling fundamental voice frequency. *Journal of the Acoustical Society of America*, 63, pp. 211-222.

[3] Shipp, T., Doherty, E.T. & Morrissey, P. (1979) Predicting vocal frequency from selected physiologic measures. *Journal of the Acoustical Society of America*, 66, pp. 678-684.

TONGUE STRUCTURAL MODEL: INTEGRATING MRI DATA AND ANATOMICAL STRUCTURE INTO A FINITE ELEMENT MODEL OF THE TONGUE

Chao-Min Wu

Biomed. Eng., Ohio State Univ.
Columbus, OH 43210, USA,
and ATR HIP Res. Labs.,Kyoto, Japan

Reiner Wilhelms-Tricarico

Speech Commun. Group,
Res. Lab. of Electronics, MIT
Cambridge, MA 02139, USA

ABSTRACT

A method is presented that has the potential to supplement MRI data with anatomical structural information from other resources that are not available in the MRI data. In the MRI data landmarks are specified which are then matched with landmarks in drawn specimens. Images or other geometric representations of specimens (such as fiber direction fields) are warped by using a thin-plate spline mapping in either two or three dimensions.

1 INTRODUCTION

The goal of this research is to obtain a 3-dimensional (3-D) tongue structural model as a preparation for dynamic simulations of the tongue during speech articulation. The tongue structural model, which is a new finite element representation of the human tongue, adapted to each individual's anatomy, will be developed by integrating tongue shape information from MRI scans and anatomical drawings to form an anatomical model, and incorporating into it the finite element structural framework.

We want to emulate closely the morphology and to some extent the biomechanics of a particular speaker, in a computational simulation of the tongue and other vocal tract structures. In particular, for the tongue, a finite element model is needed and the geometry of this model needs to be specified. The finite element model of the tongue will contain the intrinsic structural information that closely matches the speaker's anatomy (cf. [1]).

Some of the anatomical information can be extracted from magnetic resonance images. However, many important details, such as the directions of the muscle fibers, can only be obtained to a very limited extent from the MRI data. This information has to be supplemented from anatomical knowledge, which exists in the form of detailed and accurate anatomical drawings of specially studied specimens of tongue (see Miyawaki, [2]). These drawings have previously been used as reference to construct finite element models of the tongue (cf. [3] and [4]). This paper presents the basic methods that we use to combine MRI data with anatomical drawings.

2 MRI AND ANATOMICAL DATA

The MRI data consist of two stacks of transversal sections of the oropharyngeal region. Table 1 shows the specifications of the MRI data. These data were from a Shimadzu MRI machine, SMT-100GUX, which has a static magnetic field density of 1.0 T.

The data were obtained from two different speakers. During the MRI assessment, the two speakers articulated a constant vowel, Japanese /a/ in one case, and English /i:/ in the other case. From the data, areas of a size of 110 x 110 pixels surrounding the tongue were extracted. For the purpose of a larger and smoother display, they were interpolated and magnified, so that for further processing a stack of MRI images was used that consisted of 256×256 pixel images with a pixel size of (0.4196mm x 0.4196mm).

Table 1: *Specifications of the MRI data.*

Axial Imaging Parameters		
	Subj. KH (Japanese)	Subj. DO (English)
TR (ms)	800	800
TE (ms)	18	18
Images	26	26
Thickness	0.5cm	0.5cm
Interscan skip	0.0 cm	0.0 cm
View field	25 cm	25 cm
Matrix Size	256 x 256	256 x 256
Scan Time	4:05 min	4:05 min

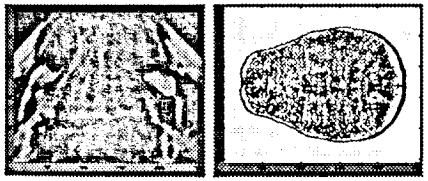


Figure 1: *(a) Enlarged oral portion of transverse transection at a level of the third vertebra from Japanese vowel /a/. (b) Full tongue sketch of transverse section at a compatible level of figures 1(a).*

Figure 1(a) shows one example of extracted regions around the tongue. It is from the Japanese subject.

Miyawaki's drawings ([2]) contain three different planes in which tongues were sliced. This makes it possible in principle to approximately reconstruct the three-dimensional structure of the tongue, including the fiber directions. Miyawaki's drawings were scanned and digitized in the computer and the transversal sections were used together with their mirror images. *Figure 1(b)* shows an example of the processed result.

Neglecting some deformations due to theslicing techniques used, Miyawaki's drawings can be aligned and stacked to obtain roughly the shape of the originally used tongue specimen. For the MRI data, the images are aligned properly. The stack of MR images can be used as a data base which allows the extraction of some landmark points in three dimensions.

Geometrically the mapping we are looking for is an interpolation mapping. We are given two sets of points: One set consists of landmarks in the stack of drawings of the tongue. The other set consists of the corresponding landmarks in the stack of MRI slices. The interpolation mapping has to map the first set of landmark points one-to-one onto the other set of landmark points. This can be illustrated easier in two dimensions.

Figure 2 shows in panel A a typical drawn tongue section and in B an outline of a tongue section as it may be obtained from MRI. In *Fig.* 2-C specified landmark pairs are shown. D shows the result of the mapping: The muscle fiber directions that are visible in A are warped and fill out the MRI contour. Panels E and F show the warping of a rectangular grid by the mapping. In the example of Fig. 2, the thin-plate spline mapping (*cf.* Bookstein,([5]) has been used to warp the figures. The thin-plate spline mapping is used for two- and three-dimensional interpolation because it has the advantage over other mappings that – among all possible interpolation mappings – it minimizes the bending energy of the interpolating function.

In the following, a short summary of the thin-plate spline mapping is given, starting at a definition of the thin-plate spline interpolation of a scalar field in two dimensions.

Let $\mathbf{P} = (x, y)$ be any point in the plane. Let $\{\mathbf{P}_i = (x_i, y_i)\}, i = 1, \ldots, N$ be a set of specific points in the plane and h_i a scalar associated with point $\mathbf{P}_i$.

The thin-plate spline interpolation of the field h_i is the function $f(x, y)$ which fulfills the following requirements:

1. $f(x_i, y_i) = h_i$ for all i.
2. The function f minimizes the following functional (bending energy):

$$I_f = \iint_{\mathbf{R}^2} (|\frac{\partial^2 f}{\partial x^2}|^2 + 2|\frac{\partial^2 f}{\partial x \partial y}|^2 + |\frac{\partial^2 f}{\partial y^2}|^2) dx dy.$$

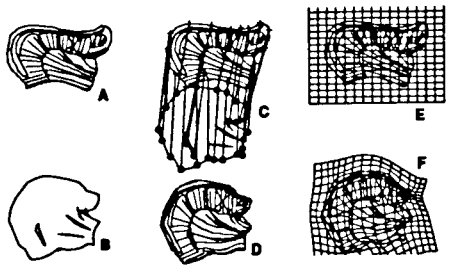


Figure 2: *Demonstration of the two-dimensional thin-plate spline mapping. (A) A typical hand-drawn sketch of a tongue section as in Miyawaki's tongue drawings. (B) Outline of tongue as could be obtained from MRI (hand-drawn). (C) Specification of corresponding landmarks (here 27 landmark pairs). (D) Panel A mapped onto panel B with thin-plate spline. (E) - (F) Demonstration of the mapping by warping a grid.*

It can be seen that in two dimensions the function f is based on the fundamental solution, U, of the biharmonic equation, $\Delta^2 U = \delta(x, y)$, where δ is the Kronecker delta function.

In two dimensions this fundamental solution function is $U(r) = r^2 log\, r$, where $r^2 = x^2 + y^2$, and in three dimensions the solution function is $U(r) = |r|$. For example, in two dimensions, the function $f(\mathbf{P})$ is

$$f(\mathbf{P}) = \sum_{i=1}^{N} w_i U(|\mathbf{P} - \mathbf{P}_i|) + a_0 + a_x x + a_y y$$

In three dimensions an additional factor $a_z z$ appears in the above formula, and $U(r) = |r|$ needs to be taken. The coefficients w_i and a_x, a_y, a_z are determined from a given the set of landmark pairs. This is shown in [5].

The extension of the method to interpolation of mappings between two three-dimensional sets of homologous points is straight-forward. One interpolating function such as the above $f(x, y)$ needs to be computed for each spatial direction.

In general, finding reliable landmarks turned out to be not so easy for the MRI data that we currently have. The location of the tip of the tongue, the path of a groove on the tongue and the rough geometry of the genioglossus and hyoglossus muscles could be assessed with confidence. An imaging anatomy atlas [6] was used as guidance to interprete the MRI data. The styloglossus muscle is also visible in the MRI data. However, the quality was insufficient to extract reliable landmark points.

Figure 3 shows a first result of the computed mapping (i.e., the anatomical model). To the left is the tip of the tongue (anterior). The thick line shows the trace of a groove on the tongue surface that can be seen in the MRI data. The line shown on the right side (posteriorly) marks the midsagittal plane on the back of the tongue, obtained from interpreting the MRI data. The circles represent the landmark points after the 3-D thin-spline mapping was applied.

3 CONCLUSIONS AND PLANNED WORK

A method was presented that has the potential to supplement MRI data with structural information from other resources that are not available in the MRI data. In the MRI data landmarks are specified which are then matched with landmarks in drawn specimens. Images or other geometric representations of specimens (such as fiber direction fields) are warped by using a thin-plate spline mapping in either two or three dimensions.

For future work, a finite element grid will be enscribed into the volume and mapped as a whole onto the MRI volume data to obtain a finite element model of an individual tongue in one (arbitrary) articulatory configuration. The finite element model can be used as a reference model for dynamic simulations of tongue movements.

(This work was supported in part by ATR HIP Res. Lab., an NSF grant to O. Fujimura, and a gift from ATR ITL to O. Fujimura. We would like to thank Dr. K. Honda from ATR-HIP and Dr. J. Negulesco from OSU for helpful discussion on interpreting MRI data.)

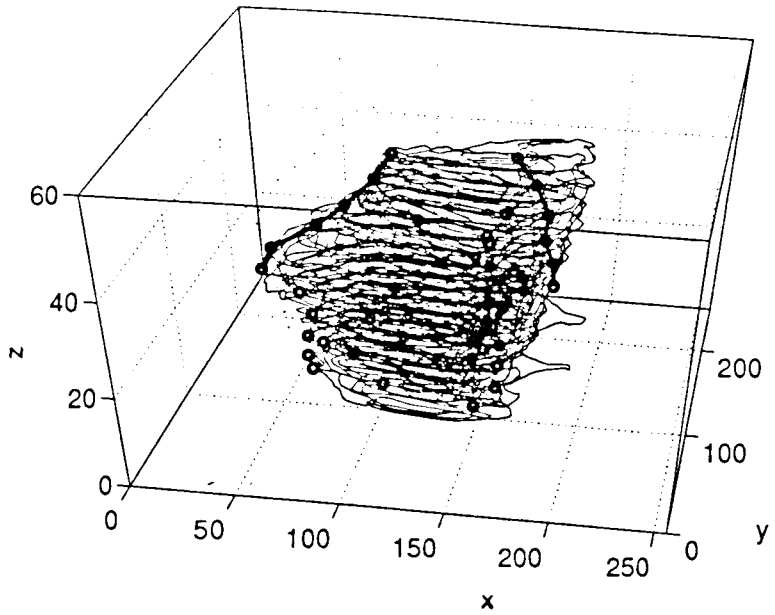


Figure 3: *The result of a 3-d mapping shown in a lateral view for data from one subject articulating the Japanese vowel /a/(in perspective projection.)*

4 REFERENCES

[1] Wilhelms-Tricarico, R. (in press), Physiological modeling of speech production: methods for modeling of soft-tissue articulators. *J. Acoust. Soc. Am.*.

[2] Miyawaki, K. (1974), A study of the musculature of the human tongue. *Ann. Bul. Research Institute of Logopedics and Phoniatrics, Univ. Tokyo*, 8:23–50.

[3] Kiritani, S., Miyawaki, K., Fujimura, O., and Miller, J.E. (1976), A computational model of the tongue. *Ann. Bul. Research Institute of Logopedics and Phoniatrics, Univ. Tokyo*, 10:243–251.

[4] Hashimoto K., and Suga, S. (1986), Estimation of the muscular tensions of the human tongue by using a three-dimensional model of the tongue. *J. Acoust. Soc. Japan*, (E) 7,1:39–46.

[5] Bookstein, F.L. (1989), Principal warps: thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-11(6):567–585.

[6] Schnitzlein, H.N., and Murtagh, F.R. (1990), *Imaging Anatomy of the Head and Spine: A photographic Color Atlas of MRI, CT, Gross and Microscopic Anatomy in Axial, Coronal, and Sagittal Planes.* Urban & Schwarzenberg, Baltimore.

HEMISPHERIC CONTRIBUTIONS TO PROCESSING AFFECTIVE AND LINGUISTIC PROSODY

T.V.Chernigovskaya, N.D.Svetozarova**, T.I.Tokareva**
**I.Sechenov Institute of Evolutionary Physiology and Biochemistry, Russian Academy of Sciences, St.Petersburg, Russia*
***State University of St.Petersburg, Russia*

ABSTRACT

The purpose of the study was to reveal cerebral hemispheric engagement in the perception of Russian prosody - affective and linguistic. Stimuli were presented monaurally. Listeners were normal adults with symmetrical hearing. The results evidenced by shorter reaction times show right-hemispheric prevalence for affective and 'idiomatic' prosody. Recognition of communicatively different prosodic types needs different kinds of asymmetry in males and females.

INTRODUCTION

For decades the left cerebral hemisphere was traditionally described as playing a major role in language functions in most individuals. The right hemisphere, however, has been demonstrated to posses considerable linguistic capabilities. It has been shown in dozens of studies carried out both in neurologic and psychiatric patients as well as in normal subjects via special techniques like dichotic, monaural and tachistoscopic stimulation. Recent neuropsychological data demonstrate contradictory character of the state of the art, partly because of the methods and models that are difficult to compare. Studies in both normal and brain-damaged subjects support with a high degree of consistency the role of the right hemisphere for emotional sphere, while hemispheric involvement in processing linguistic prosody is less clear [1-4]. Although disordered expression and recognition of emotional prosody has recently been associated with damages to the cerebral hemisphere not specialized for language itself there were notions that dysprosodic production has also been associated with left-hemisphere impairments [5-10]. On the other hand, most of the findings tend to support the idea of syncretic, holistic perception, characteristic of the right hemispheric mechanisms. Emotions - not only verbal - are proved to be the right hemispheric privilege. We also know that the right hemispheric Gestalt (cognitive and linguistic) mentality is similar to that of a child [11]. The adult listener - similar to the child in the early stage of language acquisition - starts speech processing with breaking up the spoken text into 'chunks', i.e. perceptually coherent configurations of auditory events suitable for further analysis (on the basis of stress, rhythm, etc.). This shows that the right hemisphere should be engaged into all kinds of prosodic processing, not only affective.

Our objective in this study was to reveal evidence of hemispheric specificity for the perception and understanding of different types of Russian affective and linguistic prosody in normal adults.

MATERIALS AND METHODS

Subjects

Listeners were normal adults with symmetrical hearing (thresholds of 15dB level or better for all frequences), all native Russian speakers, monolinguals and right-handers with no familial sinistrality, aged 21- 51, 7 males, 7 females.

Stimuli and Procedure

The prosody comprehension test used monaural stimulation of either the left or the right ear, contralateral ear being masked by white noise, produced by a function generator. The stimuli were natural speech utterances of 26 Russian phrases randomly ordered and played on an audiotape. Samples were those expressing emotions (surprise, politeness, anger, delight, request, etc.) as well as lexically identical but communicatively and syntactically different (declarative, interrogative, imperative, with different focal accents and syntagmatic division, some of them being well-known to listeners, 'idiomatic', some - artificially composed but grammatically correct):. e.g."He has told me." vs. "He has told me..."; "Stand there?" vs."Stand there!"; or "*John* went to Moscow yesterday"vs. "John *went* to Moscow yesterday" vs. "John went to *Moscow* yesterday' vs. "John went to Moscow *yesterday*"; or "*Drink*, not gargle" vs. "Drink not, *gargle*" etc.).

Headphone left- right orientation was switched at random. Initial orientation was alternated across subjects. Every stimulus, therefore, was presented, in the left and the right ear. Subjects were instructed to ignore the competing noise and monitor the stimulus, and to choose one of the response cards as soon as the decision was made. All instructions to subjects were recorded on the stimulus tape. The reaction time and the number and character of errors were registered.

RESULTS AND DISCUSSION

All subjects demonstrated correct recognition and understanding of the stimuli at each ear with rather few errors committed. However, the reaction time appeared to be a relevant feature, showing relative ear advantage. For each listener, the percentage of quicker recognition for each phrase and for a set of similar phrases (a prosody type) at each ear was calculated. To investigate patterns of lateralization for males and females we calculated for each of the groups the difference in performance at the two ears for each prosody type to reveal ear advantages (Fig.1 and 2).

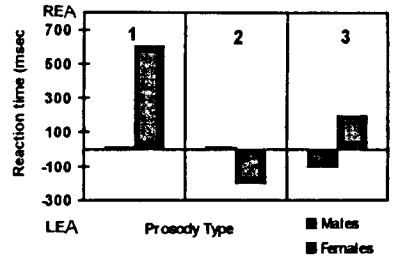


Fig.1 Ear advantages in mean perception latencies (1-'novel', 2-'idiomatic', 3-'communicative')

The overall analysis of the data showed no significant ear advantage. However, when we grouped the data in accordance with more detailed prosody types we saw evidence of more selective hemispheric involvement. Specifically, the performance was quicker when the *affective* stimuli were presented in the left ear (which is in keeping with the earlier findings), with asymmetry more evident in males. Reliably asymmetric was the processing of lexically identical but *communicatively differing* phrases in males compared to females (reliable left-

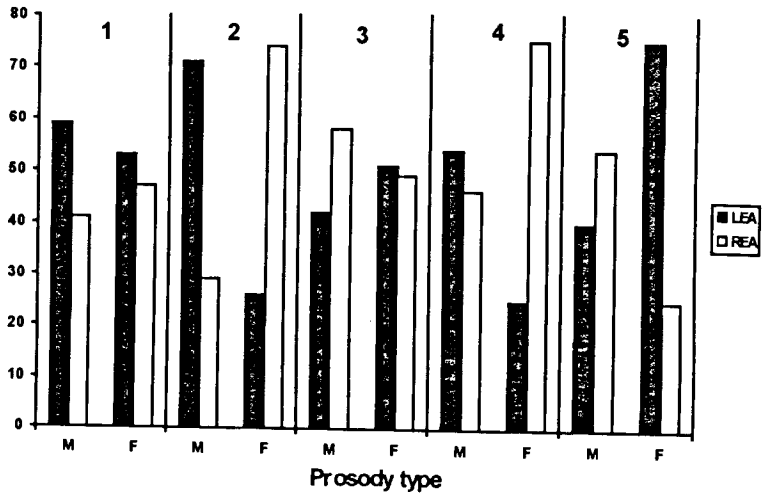


Fig.2 Percentage of phrases perceived with shorter reaction time (M- males and F- females, 1- affective, 2- communicative, 3- syntactic, 4- 'novel', 5-'idiomatic')

ear advantage in reaction time in males and right-ear advantage - though not significant - in females.). Recognition of *syntactically different* phrases showed statistically reliable right-ear advantage in males and less reliable left-ear advantage in females.In females different sentence accents, indicating difference in *syntagmatic division*, revealed right-ear advantage for '*novel*' contrary to left-ear advantage for the recognition of '*idiomatic*', '*trivial*' '*Gestalt*' samples. Less obvious was the asymmetry in males.

In summary, the results of the study offer evidence of the involvement of both cerebral hemispheres in processing prosodic information in normal subjects. Nevertheless, the data suggest that specific hemispheric prevalence in the processing is caused by several factors. Among them are *the stimuli-factors* (linguistic and cognitive - or pragmatic - features; 'novelty' acompanied by analytic functions of the left hemisphere, contrary to 'iconicity' - adequate for the right hemispheric global template recognition), and *the subject-factors* like individual psycho- physiological characteristics - age, emotional status and sex differences (resulting from cytoarchitectonic and neurochemical peculiarities). Similar to [4] our investigation indicate that prosodic processes are made up of multiple skills and therefore such functions are distributed across cerebral systems rather than strictly lateralized to a single hemisphere. We also claim that the function served by a stimulus rather than its physical nature determines laterality of processing [cf. 12].

ACKNOWLEDGEMENTS

The research described in this publication is supported in part by Grants No NVJ000 and NVJ3000 from the International Science Foundation and Russian Government and by Grant No9501-00288 from Russian Foundation for Fundamental Studies.

REFERENCES

[1] Blumstein,S.,& Cooper,W.E.(1974), Hemispheric processing of intonation contours, *Cortex*,v.10,pp.146-158.

[2] Heilman,K., Bowers,D., Speedy,L.,& Coslett, H.B.(1984),Comprehension of affective abd nonaffective prosody, *Neurology*, 34, pp.917-921.

[3] Shipley-Brown, F., Dingwall,W.O., Berlin,Ch., Yeny-Komshian, G., Gordon-Salant,S.(1988), Hemispheric processing of affective and linguistic intonation contours in normal subjects, *Brain and Language*,33, pp.16-26.

[4] Van Lancker, D.,& Sidtis, J.J. (1992) The identification of affective-prosodic stimuli by left- and right-hemisphere-damaged subjects:All errors are not created equal, *J.of Speech and Hearing Research*, v.35, pp.963-970.

[5] Monrad-Krohn, G.H. (1947), Dysprosody or altered 'melody of language', *Brain*,70, pp.405-415.

[6] Luria, A.R. (1966), *Higher cortical funcions in man.* Basic Books: New York.

[7] Whitaker,H. (1982),Levels of impairment in disorders of speech. In *Neuropsychology and cognition*, (Eds.Malatesha,R.N. & Hartlage,L.C.),v.1, The Hague: Nijhoff.

[8] Ardila, A., Rosselli,M., Ardila,O. (1988),Foreign accent: an aphasic epiphenomenon? *Aphasiology*, 2, pp.493-499.

[9]Moen, I. (1991), Functional lateralization of pitch accents and intonation in Norwegian: Monrad-Krohn's study of an aphasic patient with altered 'melody of speech', *Brain and language*,41, pp.538- 554.

[10] Chernigovskaya T. (1992), Intonation processing and hemispheric mechanisms : arbitrary linguistics and universal biology. Paper presented at the *8th Annual meeting of Language Origins Society*, Cambridge, 1992.

[11] Chernigovskaya T. (1994),Cerebral lateralization for cognitive and linguistic abilities: Neuropsychological and cultural aspects. In: *Studies in Language Origins*, (Eds. J.Wind, A.Jonker, R.Allott, L.Rolfe),v.III J.Benjamins Publ.Co.:Amsterdam/ Philadelphia, pp.55-76.

[12] Jakobson,R. & L.Waugh. (1979), *The sound shape of language.* Bloomington, Ind., and London.

PERCEPTION OF HESITATIONS IN SPONTANEOUS FRENCH SPEECH

Danielle Duez
CNRS URA 261, Laboratoire Parole et Langage
Université de Provence, France

ABSTRACT

The present study deals with the perception of hesitations in spontaneous speech.. The results showed that half of the hesitations were identified by at least 75% of the listeners. Listener responses were also found to depend on hesitation type: silences were never perceived, contrary to filled pauses, lengthenings and repeats. For these hesitations, the listeners' responses tended to correlate with duration. These findings suggest that hesitations are perceived.

INTRODUCTION

The perception of hesitations in speech has not aroused much interest in speech research. Goldman-Eisler's claim [1] that hesitation pauses do not serve communication may be one of the reasons for this lack of interest. Following Goldman-Eisler's view the majority of the studies on the significance of hesitations have tested the hypothesis that hesitations may not be "heard". For example, in an experiment where decoders heard recorded utterances and attempted to reproduce them, Martin and Strange [2] observed that decoders placed relatively more of their hesitations at sentence breaks than did encoders. Moreover, instructions to reproduce hesitations increased hesitations and words but at the expense of the percent of correct words. The authors stressed the fact that decoders distribute pauses in accordance with the distributional scheme of pauses.

Later studies on pause perception [3 and 4] confirmed the influence of the distributional scheme of pauses on pause perception; however, it was also shown that pauses occurring within constituents were not detected as well as between-phrase pauses, but they were not entirely ignored. Moreover, subjective pauses (i.e., perceived pauses which do not corresponds to silent pauses) were also found to be related to significant vowel lengthening, which mostly correspond to the realization of a syntactic boundary, and/or the presence of a hesitation. This tends to prove that hesitations are perceived.

The purpose of the perceptual experiment reported here was to check whether hesitations are *actually* perceived. It was assumed that the accuracy of hesitation perception would depend on different factors such as the type, duration, and distribution of hesitations.

EXPERIMENTAL PROCEDURE

Subjects

The subjects were twenty-five native speakers of French. All were students at the University of Provence. They had no history of speaking or hearing disorders. They were paid for the task.

Materials

The stimuli used in the experiment were extracted from three hours of conversational speech produced by seven French male speakers. The consersations were digitized at a sampling rate of 16 KHz with a SUN computer. The presence of hesitations

was controlled both acoustically and perceptually. One hundred sentences were extracted from the conversations. They were selected according to the following two criteria :1) each of the utterances began and ended with a terminal contour, and 2) each sentence contained one hesitation, or no hesitation. The limits of each utterance were determined auditorily and visually from the oscillographic and spectrographic waveforms. Out of the 100 utterances, 68 contained hesitations such as filled pauses (30 instances), repetitions of syllables and grammatical words (16 instances), silent pauses (12 instances) and lengthened syllables (10 instances). In the 32 remaining utterances, there was no hesitation.

Hesitations.

Four types of hesitation were defined:

-Filled pause (FP) : any occurrence of the French hesitation interjections (euh, hum..)

-Repeat (R): any unintended repetition of a sequence of phonetic segments that is subsequently produced in its complete intended form.

Silent pause (S): any interval on the oscillographic trace where the amplitude is indistinguishable from that of the background noise. Silent hesitation pauses were restricted to interruptions located within a minor phrase, e.g. an article and a noun, a proposition and a verb...

Lengthened syllable (L): any abnormally lengthened syllable or vowel.

As false starts were always associated with another hesitation, they were not included in the corpus.

Testing procedure

The utterances were transferred to a 486 PC computer using a program written by Pavlovic et al. [5]. The subjects were told that the 100 utterances they would hear were extracted from conversations, and that each utterance contained no hesitation, or one of the following hesitations, an "euh", a lengthened syllable, a silence, or a repeat. In the latter case, they were cautioned not to mistake a hesitation repeat for a repeat which has a semantic role. They were instructed to report the presence of a hesitation in a two-alternative forced choice task by clicking the appropriate response (yes or no) as soon as possible. Utterances were presented to subjects with headphones at a sound pressure level of 63 dB in a quiet room. Each subject took the test independently, at his own pace: after presenting an utterance, the computer waited for the subject to click the appropriate response before recording it. Then when the utterance was totally played, the next utterance was presented. An inter-utterance interval of 3 s was maintained and a different random order of utterances was used for each subject. A pretest of 8 utterances was included to familiarize subjects with the task and to give them a further chance to ask questions. The session lasted about twenty minutes

Analysis

The limits of each hesitation were determined auditorily and acoustically by analyzing both the oscillographic trace of the signal and wide band spectrograms; each hesitation was measured in ms. For a lengthened syllable, the degree of vowel lengthening was estimated with respect to the mean length of non-prominent vowels for a within-phrase lengthening, and with respect to the mean length of prominent vowels for a lengthening occurring at a major break. This estimation was done for each speaker. Repeats were also expressed as the number of syllables and the number of times the same word was repeated. Filled pause duration ranged from 75 ms to 1554 ms, lengthenings from 243 ms to 854 ms, silences: from 231 ms to 776 ms, and repeats from 170 ms to 742 ms. The repeats ranged from

the repetition of the initial syllable in a word to the repetition of a bisyllabic word. The syntactic distribution of hesitations was studied as a function of their duration and their types. Three locations were considered: 1) within a minor phrase, 2) between minor and major phrases, and 3) between clauses. The number of listeners perceiving a hesitation was examined as a function of the type, duration and location of each hesitation. As subjects were given a forced binary choice, a hesitation was considered as perceived when reported by at least 19 listeners (i.e. 75% of the listeners).

RESULTS

Hesitation type

The mean and standard deviation of listener responses for each hesitation type are as follows. FP (M= 19.3, SD= 0.7); L (M=18.3, SD=1.3); R (M=19.4, SD=1) and S(M=8.8, SD=1.1). Hesitation type had a significant effect on mean number of responses (F(3, 64)=20.8, p=0.0001). A post-hoc test revealed that the mean number of responses obtained for S was significantly different from that obtained for FP, L and R (p=0.0001). The analysis of number of listeners perceiving each hesitation was consistent with this finding : 36 of the 68 hesitations were perceived by 75% of the listeners (19 out of the 30 FP, 6 out of the 10 L, and 11 out of the 16 R). No silence was perceived as a hesitation by 75% of the listeners.

Hesitation duration

The results suggest that duration is a cue to the perception of hesitations. A 75 ms FP was detected by only 10 listeners. FP's between 100 ms and 400ms were perceived by a number of listeners ranging from 12 to 18. Most of the FP's longer than 400 ms were perceived by the majority of listeners. There was only one exception: one 910-ms FP was detected by only 18 listeners. Similar tendencies were observed for L's : L's greater than 450 ms were perceived as hesitations by at least 75% of the listeners while short lengthenings (between 200 ms and 500 ms) were perceived as hesitations by a number of listeners ranging from 9 to 12 listeners. For FP's and L's, there seems to be a duration threshold zone around 500 ms. Listener responses correlate less with duration for R, suggesting that listeners are also sensitive to the number of repetitions. The correlation coefficients obtained for FP, L and R are 0.5, 0.6 and 0.1, respectively.

Hesitation location

Listener responses tended to correlate more with duration for FP's located at syntactic boundaries (r^2=0.5) than for within-phrase FP (r^2=0.3). All hesitations longer than 500 ms were perceived by the majority of the listeners when located at phrase or clause boundaries, while a within-phrase hesitation as long as 500 ms was only detected by 13 listeners. However, the low number of cases does not allow us to test the significance of these differences.

CONCLUDING REMARKS

The finding that emerges from the present study is that hesitations are "heard". This finding is rather robust since listeners only had to identifiy one hesitation at a time. In normal communication they are used to hearing a succession of hesitations in the same location as speakers tend to accumulate hesitations when expressing themselves spontaneously. Listeners' sensitivity to hesitations is not uniform; it seems to depend on hesitation type, and hesitation duration. An extansion of the present study is in progress. It should allow us to relate a hesitation perception threshold to a given duration and a specific location within the utterance.

The present finding has some implications for the role of hesitations in speech communication. Studies on the significance of pauses [7 and 8] have shown that hesitations hinder effective communication. Other studies have suggested that hesitations serve as an opportunity for the hearer to review and integrate what is to follow [9]. The role of hesitations in speech communication is probably complex. Hesitations may disturb components of speech, especially when they are long, but at the same time they may provide potentially useful information about the characteristics of the speaker's style and communication situation. A closer investigation of hesitations should bring us a better understanding of the function of paralinguistic and extralinguistic information in speech communication.

ACKNOWLEDGEMENT

Acknowledgements to M. Brousseau and R. Espesser for their technical assistance

REFERENCES

[1] Goldman-Eisler, F. (1968). *Psycholinguistics. Experiments in spontaneous speech*, Academic Press, London and New York

[2]Martin, J. G. and Strange, W. (1968), The perception of hesitation in spontaneous speech, *Perception and Psychophysics*, 3(6), pp. 427-438.

[3] Duez, D. (1985) Perception of pauses in continuous speech, *Language and Speech*, 28(4), pp.377-384.

[4]Duez, D. (1993) Acoustic correlates of subjective pauses, *Journal of Psycholinguistic Research*, 22 (1), pp. 21-39.

[5]Pavlovic, C., Brousseau, M., Howells, D., Miller, D., Hazan, V., Faulkner, A. and Fourcin, A. (1995) Analytic assessement and training in speech and hearing using a poly-lingual workstation, *EURAUD*, TIDE Congress, Paris, to appear.

[6] Duez, D. (1982) Silent and non silent pauses in three speech styles, *Language and Speech*, 25, pp.11-28.-

[7]Aaronson, D. (1968). Temporal course of perception in an immediate recall, *J. Exp. Psychol.* 76, pp.129-140

[8]Reich, S. S. (1980) Signifiance of pauses for speech perception, *Journal of Psycholinguistic Research*, 9(4), pp.379-389.

[9]O'Connell, D. C. , Kowal, S. and Hörmann, H. (1969). Semantic determinants of pauses, *Psychol. Forsch.* , 33, pp.50-67.

CUE INTERACTION IN THE PERCEPTION OF INTERVOCALIC AND SYLLABLE-INITIAL VOICELESS FRICATIVE/AFFRICATE CONTRASTS

A. Faulkner, S Rosen[1], AM Darling, and M Huckvale
Department of Phonetics and Linguistics, University College London. and
Department of Audiology, North Western University, Evanston, Illinois[1]

ABSTRACT

The interaction of frication duration, frication rise-time and pre-frication silence interval has been studied for the English voiceless fricative/affricate contrast. Cue interaction is a pervasive feature of the results, and major differences were found between syllable-initial and intervocalic stimuli. These interactions prevent the data from being fitted by models based on either acoustic, auditory, and information integration principles. The interactions are ascribed to cognitive processes.

BACKGROUND

One of the major theoretical problems in speech perception arises from the existence of substantial trading relations or interactions between multiple acoustic cues to the identity of speech sounds. Traditional accounts of cue trading appeal to articulatory representations of speech [e.g., 1]. Others [e.g., 2] have approached cue trading at the level of information integration. We have been concerned with the contribution that auditory transformations of the acoustic speech signal may make in speech perception.

Amongst known auditory transformations are a number that may in themselves account for cue trading. For example, Delgutte [3] has proposed that the rapid adaptation that occurs in hair-cell mechanical to neural transduction results in an interaction in the neural coding of frication noise between frication rise time and the duration of the silence that precedes frication. This could account for perceptual interactions in the perception of the voiceless fricative/affricate contrast [4]. Other properties of the peripheral auditory system, such as the frequency dependence of the temporal resolution and group delay of cochlear filtering, may account for further interactions. For example, the effect of temporal voice onset time cues in the plosive voicing contrast depends on place of articulation. Since the frequency region in which temporal voice-onset time information is present depends on place of articulation, the auditory frequency channels by which this temporal information is processed will also depend on place of articulation. In consequence, the properties of different auditory frequency channels could influence the processing of such temporal information.

This study concerns the voiceless fricative/affricate contrast. The perception of the contrast between the voiceless fricative /ʃ/ and the voiceless affricate /tʃ/ (the initial consonants in the words "ship" and "chip") is generally supposed to be based on at least two perceptual cues. Both cues are associated with acoustic differences that are typical of natural speech tokens. One is the duration of the noise-excited frication part of the consonant. In /ʃ/ this is relatively long, whilst in /tʃ/ it is shorter. The second cue arises from the amplitude envelope of the frication, which has a gradual onset for /ʃ/, but a more rapid onset for /tʃ/, where there is often also a distinct brief initial burst of frication. When these consonants are preceded by a vowel, the duration of silence between the vowel and the onset of frication has been identified as a third cue; the silent interval is typically absent before /ʃ/ but present before /tʃ/.

While there has been debate about the relative importance of duration and time-amplitude envelope cues [5], there is clear evidence that these cues show trading relations [4,6].

EXPERIMENTAL STUDY

Existing published data are limited, and come from studies which differ in many respects. In order to model the interactions between these cues, it was necessary to collect a substantial body of empirical data from human listeners. The

stimuli for the experiment described here were all based on a natural /aʃa/ token produced by a female British English talker. Both syllable-initial (/ʃa/ /tʃa/) and intervocalic (/aʃa/ /atʃa/) contexts were investigated. The stimuli represented factorial combinations of total frication duration (120 to 220 ms), frication rise time (0 to 100 ms), and in the intervocalic case, the duration of the silent interval between the initial vowel and the onset of frication (0 to 80 ms). Stimulus manipulations were performed by digitally modifying the duration and amplitude of the frication noise from the natural /aʃa/ model and adding silence where necessary.

Nine listeners responded to a total of 6 repetitions of each of 193 stimuli. They were asked to label each stimulus as containing either of the consonants "sh" or "ch".

Major differences were found between the syllable-initial and intervocalic stimuli, which have not been compared in previous studies. A full report of the data appears in [7]. Selected data are shown in fig. 1. A logistic regression was used to establish significant main effects and cue interactions. The most striking interaction was that between rise time and the presence or absence of a preceding vowel. Shorter rise times increased the frequency of affricate labelling for a syllable-initial consonant (fig 1, panel d), while this effect was absent in the intervocalic case and indeed reversed where the silent interval between the initial vowel and consonant was short (fig 1, panels a and b). The data show other, more expected, properties of fricative/affricate perception in that shorter silence duration and longer frication duration both led to increased frequency of fricative labelling. The most pervasive feature of the data is that of interaction between cues. Interaction was found between frication duration and rise

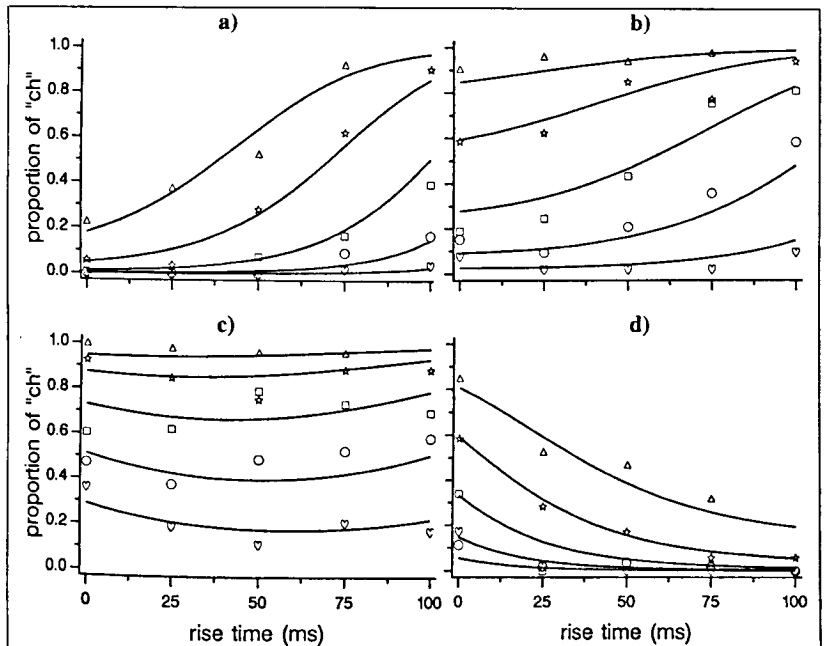


Figure 1. Proportion of affricate ("ch") responses as a function of frication rise time. Data are shown for the intervocalic stimuli with pre-frication silence durations of 0, 20, and 60 ms (panels a, b and c), and for the syllable-initial stimuli (panel d). The curves shown in the figure result from a logistic regression. Each line represents results from a particular frication duration, with values of 120, 145, 170, 195 and 220 ms, from top to bottom.

time, and between frication duration and the presence of a preceding vowel. Where the initial vowel was present, frication duration and rise time also interacted with silence duration.

MODELLING

An Auditory Model

To examine Delgutte's claim that rapid adaptation can account for interactions between silence duration and frication rise-time, an auditory model was developed. This made use of a gammatone filterbank [8] and a hair-cell model [9] to simulate peripheral auditory processing. A simplified model was used here which included only one auditory filter with a centre frequency of 3000 Hz. The output of this model was the probability of hair-cell firing as a function of time. Probability of firing is equivalent to the relative firing rate over the ensemble of auditory fibres driven by an auditory filter. The decision statistic was the peak hair-cell firing probability close to the onset of frication, relative to the subsequent probability of firing to the quasi-steady state frication. The greater the relative peak height, the stronger is the presumed evidence for an affricate consonant. A model of this sort has not previously been examined with truly speech-like inputs, and we were able to verify that the rapid adaptation exhibited by the hair-cell simulation does predict an interaction between silence duration and frication rise time. Longer silences following a vowel allow the hair-cell to recover from its response to the preceding vowel, and hence to exhibit a greater probability of firing at the onset of frication. However, the predicted interactions do not correspond to those observed. In particular, in the intervocalic context ,we observed that longer rise time leads to increased frequency of affricate responses at shorter silence durations (fig 1, panels a and b), where the model predicts the reverse effect of rise time.

An acoustic model based on rate of rise of frication

An acoustic model due to Weigelt et al. [10] was also examined. Here, the decision statistic was the peak rate of rise of the physical intensity of frication, where a high peak rate is assumed to be associated with affricate responses. Because the measurement of rate of rise necessarily entails the integration of information over time, the extracted rate of rise is affected not only by the rise time, but also by the silence interval. The rate of rise parameter is also affected by frication duration, but only for very short rise times (less than 25 ms). Hence, this model, like the auditory model above, can predict interactions between cues. However, the predictions do not account for the empirical data. As with the auditory model above, the rate of rise model cannot account for the finding that longer rise times in the intervocalic case lead to increased frequency of affricate responses.

It would be possible to consider an auditory model in which the decision statistic proposed by Weigelt et al. was applied to auditory firing probability rather than to the acoustic signal. It is difficult to imagine, however, that such a model would be able to account for our data.

FLMP model

While Massaro's fuzzy logical FLMP model [2] makes no assumptions about the relationship between acoustic or auditory parameters and the perceptual classification of speech sounds, it does claim to account for the integration of information from several sources and as such, can be fitted to data such as these. For our data, the FLMP model fails because it incorporates the assumption that features do not interact in the statistical sense, an assumption clearly contradicted by our data.

CONCLUSIONS

None of the models examined here can provide a satisfactory account of affricate/fricative perception. An articulatory account also appears untenable because the observed perceptual interactions in our intervocalic condition are not consistent with the interdependence of the corresponding acoustic properties found in speech production. Again, the problematic finding is that a shorter frication rise time increases the frequency of fricative labelling at the shorter silence durations. A shortening of frication noise rise times associated with fricative as opposed to affricate productions has never been observed in speech production [6,11]. Our perceptual results are not likely to arise from an artefact of the particular stimuli, as we have since replicated the same finding using different stimuli based on manipulations of a natural /ɑtʃɑ/ token.

A full account of the cue interactions in this consonant contrast cannot yet be provided. Auditory transformations in the time-domain do not (and probably cannot) account for all aspects of the data, although they may well have an important role. We have reached the same conclusion for the place of articulation dependence of the plosive voicing contrast. A related study [12], led to the conclusion that auditory processing cannot account for the effects of place of articulation on the interpretation of voice onset time. Our results appear to refute the claim made by Damper et al. [13] that properties of auditory frequency analysis play a key role in this phenomenon.

It may be necessary to consider cognitive rather than electro-physiological notions of auditory attributes relating for example to subjective duration and suddenness of onset. Further, the surrounding speech context of the consonant exhibits profound interactions with other cues. It may, for this reason, be necessary to take account of the phonetic or other linguistic properties of the context in order to fully understand these processes.

ACKNOWLEDGEMENT

Supported by JCI Cognitive Science/HCI grant SPG8920412.

REFERENCES

[1] Repp, B. H., Liberman, A. L., Eccardt, T., and Pesetsky, D (1978) Perceptual integration of acoustic cues for stop, fricative and affricate manner. J. Exp. Psychol., Human Percept. and Perf., 4, 621-637.

[2] Massaro, D.W. (1987) *Speech Perception by ear and eye: A paradigm for psychological inquiry.* Hillsdale, New Jersey: Lawrence Earlbaum Associates

[3] Delgutte, B. (1982) Some correlates of phonetic distinctions at the level of the auditory nerve. In: *The representation of speech in the peripheral auditory system,* eds. R. Carlson and B. Granstrom. Amsterdam: Elsevier, pp 131-149

[4] Dorman, M.F., Raphael, L.J. and Isenberg, D. (1980) Acoustic cues for a fricative-affricate contrast in word-final position. J. Phonetics, 8, 397-405

[5] Kluender, K.R. and Walsh, M.A. (1992) Amplitude rise time and the perception of the voiceless affricate/fricative distinction. Percept. Psychophys., 51, 328-333

[6] Howell, P., and Rosen, S. (1982) Production and perception of rise time in the voiceless affricate/fricative distinction. J. Acoust. Soc. Am, 73, 976-1984.

[7] Rosen, S., Darling, A.M., Faulkner, A. and Huckvale, M. (1993) Cue interaction in an intervocalic voiceless affricate/fricative contrast. Speech Hearing and Language, Work in Progress, Dept. Phonetics and Linguistics, UCL. Vol 7, 183-197.

[8] Darling, A.M. (1991) Implementing a Gammatone filter. Speech Hearing and Language, Work in Progress, Dept. Phonetics and Linguistics, UCL, Vol 5, 41-62.

[9] Meddis R (1988) Simulation of auditory-neural transduction: Further studies. J. Acoust. Soc. Am., 83, 1056-1063

[10] Weigelt, L.F., Sadoff, S.J. and Miller, J.D. (1990) Plosive/fricative distinction: the voiceless case. J. Acoust. Soc. Am., 87, 2729-2737.

[11] Howell, P., and Rosen, S. (1983) Closure and frication measurements and perceptual integration of temporal cues for the voiceless affricate/fricative contrast. Speech Hearing and Language, Work in Progress, Dept. Phonetics and Linguistics, UCL. Vol. 1, 109-117.

[12] Darling, A.M., Huckvale, M.A., Rosen, S., and Faulkner, A. (1992) " Phonetic classification of the plosive voicing contrast using computational modelling" In: *Speech and Hearing; Proc. Inst. Acoust. 1992 Conference,* Vol 14., part 6.

[13] Damper, R., Pont, M., and Elenius, K (1991) Representation of initial stop consonants in a computational model of the dorsal cochlear nucleus. STL-QPSR, 4, 7-41.

INDIVIDUAL VARIABILITY IN THE PERCEPTUAL WEIGHTING OF CUES TO STOP PLACE AND VOICING CONTRASTS

V. Hazan and B. Shi
Dept of Phonetics and Linguistics, University College London, UK

ABSTRACT

Listener variability in the perceptual weighting of acoustic cues to stop contrasts has been studied in a group of 50 listeners. For the place contrast, perceptual weighting given to the burst and transition cues varied widely with listeners and vocalic contexts. More homogeneous results were obtained for the voicing contrast. It is argued that listeners vary in their perceptual strategies and that acoustic cues vary in terms of their robustness.

INTRODUCTION

Phonetic contrasts are marked by a multiplicity of acoustic cues. The relative weighting given to cues to contrasts in manner, voicing and place of articulation has been shown to vary according to vocalic context [1,2] and speaker characteristics [1]. Although the emphasis in studies of cue weighting has been on the presentation of averaged results, many studies have found that some individual listeners may show quite different cue weighting strategies than the norm.

The aim of this study was therefore to quantify the amount of variability in cue weighting seen in a large and relatively homogeneous listener population. In order to evaluate the effect of contrast type and vocalic context on individual variability, listeners were tested on two different contrasts, each presented in three vocalic contexts.

METHODOLOGY

Stimuli

Natural tokens produced by a male English speaker were used as a base for copy-syntheses which were obtained using a version of the Klatt software synthesizer produced by Sensimetrics (KLSYN88). Once a close copy of the minimal pair was obtained, a set of continua was prepared by interpolating the parameters under investigation.

/g/-/d/ place contrast

The acoustic cues under investigation were the burst transient and F2/F3 transitions at vowel onset. In order to evaluate the effect of vocalic context on cue-weighting, three minimal pairs were used: GATE-DATE, GAT-DAT, GEET-DEET. The initial burst was synthesized through the parallel branch of the synthesizer, by exciting five formants with noise for 5 ms. The formant values for F2 and F5 were fixed at 1800 Hz and 4000 Hz. F3 varied from 2300 Hz at the /g/ endpoint to 4200 Hz at the /d/ endpoint. F4 varied from 2700 Hz at the /g/ endpoint to 4600 Hz at the /d/ endpoint. The amplitude of the two formants varied from 75 dB at the /g/ endpoint to 60 dB at the /d/ endpoint. The burst was identical for all three contrasts.

The vowel was synthesized through the cascade branch of the synthesizer. Formant transitions of F2 and F3, which were matched to values measured in the natural tokens and which extended over the first 50 ms of the vowel, constituted the second cue to the place contrast. A full description of the stimuli is given in [3].

In order to evaluate the relative contribution of each cue to the perception of the contrasts, three test conditions were prepared for each minimal pair. In the "Combined-cue" condition, both cues were varied together; in the "No transitions cue" condition, the formant transitions were fixed at a neutral value; in the "No burst cue" condition, the burst was removed through waveform editing.

/g/-/k/ voicing contrast

The two main cues to the contrast under investigation were voice onset time (VOT), i.e. the duration between burst release and voicing onset, and F1 onset frequency. The minimal pairs were: GATE-KATE, GAT-KAT and GEET-KEET. VOT varied from 15 ms to 115 ms following burst onset, in 10 ms steps. As VOT increased, so did the cutback in F1 relative to higher formants that were present within the aspiration portion.

The second cue under investigation was F1 onset frequency. The F1 transition took place over the first 50 ms of the vowel. At the /g/ endpoint, for the GATE-KATE contrast, F1 onset frequency was set at 327 Hz rising to 500 Hz. For the GAT-KAT contrast, F1 onset frequency was set at 420 Hz rising to 630 Hz. For the GEET-KEET contrast, F1 onset frequency changed from 310 Hz to 290 Hz.

For each minimal pair, two stimulus conditions were prepared: (1) "Combined-cue", in which both VOT and F1 onset frequency were co-varying, and (2) "No F1 transition cue", in which F1 at onset was fixed at the frequency reached at the end of the formant transition period.

Listeners

Listeners were 50 volunteers with pure tone thresholds of 20 dB HL or better from 0.25 to 8 kHz in both ears. The listeners, students at UCL, ranged in age from 18 to 32 years (mean: 21.1 years, s.d. 3.09), and were native English speakers who had no training in phonetics and little or no previous exposure to synthetic speech.

Test procedure

The identification test procedure was computer-controlled. Stimuli down sampled at 10 kHz were presented as two alternative forced choice identification tests. Stimuli were amplified to a comfortable listening level and presented through AKG 240 DF headphones. Listeners responded to each stimulus by pressing a touch sensitive response box.

For each contrast, all test conditions were randomised together in order to reduce range effects. 32 responses per stimulus were collected over four sessions for each listener.

RESULTS

A statistical approach based on Generalized Linear Models (GLMs) was used to determine the extent to which the change in deviance between the combined-cue condition and each of the single-cue conditions for a given contrast was significant. This technique, analogous to ANOVA, was used as it is especially tailored to the analysis of multi-variate data involving binary responses (for a full description, see [2]).

On the basis of previous results [2], it was hypothesized that, for the place contrast, there would be evidence of listener groups showing different usage of the burst and formant transition information. For each contrast, the percentage of listeners showing significant differences in identification function between the combined-cue and each single-cue condition was calculated (see Table 1). Some listeners were affected by the removal of either cue. However, it can also be seen that for each minimal pair in the place contrast tests, some listeners were unaffected by the removal of the burst information, while others were unaffected by the removal of the formant transition

information. A small number of listeners (12% for GATE-DATE, 6% for GAT-DAT, 8% for GEET-DEET) were unaffected by the removal of either cue and could therefore reliably label the contrast on the basis of whatever cue information was present.

Table 1. Percentage of listeners showing significant deviances in single-cue conditions of the place contrasts relative to the combined-cue condition.

	No burst cue	No trans. cue
GATE-DATE	68 %	70 %
GAT-DAT	84 %	42 %
GEET-DEET	88 %	10 %

The number of listeners significantly affected by the removal of the formant transition cue varied widely across vocalic contrasts from 10% for the GEET-DEET contrast to 70% for the GATE-DATE contrast. The percentage of listeners significantly affected by the removal of the burst cue varied between 68% (GATE-DATE contrast) and 88% (GEET-DEET contrast).

Table 2. Percentage of listeners showing significant deviances in the single-cue condition of the voicing contrasts relative to the combined-cue condition.

	F1 transition removed
GATE-KATE	2 %
GAT-KAT	34 %
GEET-KEET	4 %

For the /g/-/k/ contrast, an examination of individual results reveals less variability than for the place contrast. The removal of the F1 onset cue had little effect for the GATE-KATE and GEET-KEET contrasts. However, 34% of listeners showed a significant effect of removal of F1 onset cue for the GAT-KAT contrast.

DISCUSSION

The results of this study gives some evidence of the extent of variability in cue-weighting across contrasts and vocalic context. For the /g/-/d/ place contrast, the perceptual weighting of the burst and formant transition cues varied quite considerably according to vocalic context. For the GATE-DATE contrast, a similar number of listeners were affected by the removal of either cue. For the GEET-DEET contrast, there was a clear dominance of the burst cue over the formant transition cue. For the GAT-DAT contrast, a less extreme imbalance was obtained, with 84% of listeners affected by burst removal vs 42% by the removal of the formant transition cue.

The effect of vocalic context on cue-weighting appeared to be related to the degree of acoustic prominence of the cue. For example, the greatest effect of the removal of F1 transition for the voicing contrast was found for the vowel environment with the highest first formant and therefore the greatest transition extent. Similarly, the least effect of F2 transition removal in the place contrast was obtained in the context of /i/ in which the formant transitions are less pronounced due to the high F2 of the vowel.

Within a specific vowel environment, the evidence of clear individual differences in the perceptual effect of acoustic cue removal would suggest that listeners do differ in the use that they make of acoustic cue information contained in the speech signal. This confirms results of previous studies involving nonsense syllables [4] and identification tests for speech contrasts [2] and goes some way towards explaining some contradictory results found in the literature on the perceptual

weighting of acoustic cues to speech contrasts.

At the speech pattern processing level, the fact that the effect of cue-weighting was more variable for the place contrast than for the voicing contrast suggests that cues differ in terms of their "robustness". For the voicing contrast, VOT is clearly dominant cue for a vast majority of listeners, whereas for the place contrast, relative importance of burst and formant information varies greatly across listeners and across vocalic contexts. A better understanding of which acoustic cues are least subject to listener and contextual variability has implications for work on cue-enhancement in synthesized and degraded natural speech.

It may be hypothesized that individual variability in the use of acoustic cues is due not to audiological differences but to the development of different perceptual strategies during language acquisition, where individuals may focus on one of several redundant cues contained in the speech signal. There is ample evidence of individual differences in language and speech development (for a review, see [5]). Some evidence of individual differences in the perceptual weighting given to cues for a "bees/peas" voicing contrast was seen in a study [6] in which only 60% of 4-year old children were affected by a change in the vowel stem, which introduced conflicting spectral cues. This was similar to the percentage of adults affected by the conflicting cue.

The presence of sizable variability and of different perceptual strategies within a homogeneous population of listeners highlights the importance of considering the effect of human factors in the interpretation of the results of perceptual experiments. Indeed, the particular composition of the listener group might have a strong effect on scores obtained, especially if the listener group is small. The existence of individual differences in perceptual strategies might also go some way towards explaining the great difference in performance seen in the use of speech processing aids by deafened adults fitted with cochlear implants, for example.

ACKNOWLEDGEMENT

This work was funded by a project grant from the Science and Engineering Research Council and the Ministry of Defence (GR/F 33735).

REFERENCES

[1] Dorman, M.F., Studdert-Kennedy, M., Raphael, L.J. (1977), "Stop-consonant recognition: release bursts and formant transitions as functionally-equivalent, context-dependent cues", *Perception and Psychophysics*, vol. 22, pp. 109-122.

[2] Hazan, V. & Rosen, S. (1991), "Individual variability in the perception of cues to place contrasts in initial stops", *Percept. Psychophysics*, vol. 49, pp. 187-200.

[3] Hazan, V. and Shi, B. (1993), "Individual variability in the identification of plosive place and voicing contrasts" *Speech, Hearing and Language: UCL Work in Progress*, vol. 7, pp. 77-94.

[4] Santi, S. and Grenié, M. (1990), "Individual strategies in synthetic speech evaluation", *Proceedings of the ESCA workshop on Speech Synthesis*. Autrans, France, September 1990, pp. 265-68.

[5] Bates, E., Bretherton, I. and Snyder, L. (1988), *From first words to grammar: individual differences and dissociable mechanisms*, Cambridge: Cambridge University Press.

[6] Howell, P., Rosen, S., Lang, H. and Sackin, S. (1992), "The role of F_1 transitions in the perception of voicing in initial plosives", *Speech, Hearing and Language, Work in Progress UCL*, vol. 6, pp. 117-126.

ON THE INFLUENCE OF THE INTERNAL STRUCTURE OF A SYLLABLE ON THE P-CENTER-PERCEPTION

Peter. M. Janker
Forschungsschwerpunkt Allgemeine Sprachwissenschaft, Berlin,
Institut für Phonetik und Sprachliche Kommunikation der LMU, München, Germany

ABSTRACT

The two experiments described here investigate whether or not the internal structure of a syllable has an influence on the p-center-perception. Subjects had to perform a synchronisation task by tapping to sequences of German monosyllabic words with either different nuclei or varying complexity within the syllable shell.

INTODUCTION

In recent years various investigations [1-7] have been undertaken to gain knowledge about the parameters influencing the 'moment of occurrence' or the so called p-center [8]. Some of the models proposed suggest that the actual acoustic make-up of the phonological segments of a syllable is responsible for its p-center position hence the internal structure is a parameter which should not be neglected. Evidence for the assumptions made is mainly taken from experiments with synthesized artificial sound or speech stimuli.

EXPERIMENTAL DESIGN

To test whether or not the internal structure or complexity of the syllable has an influence on the p-center position two synchronisation experiments with naturally spoken stimulus material were carried out.

Stimulus material

For experiment one (VQ) a set of monosyllabic stimuli with phonologically identical shell but different nuclei was produced. To build the stimuli having a short vowel (abrupt cut) the German words <Still, Stell, Stall> [ʃtɪl, ʃtɛl, ʃtal], for the stimuli with a long vowel (smooth cut) the German words <Stil, Stehl, Stahl> [ʃti:l, ʃte:l, ʃta:l] were used.
The overall duration of the stimuli with smooth cut is approximately 80 ms longer than that of the stimuli with abrupt cut despite some compensational coda shortening in the case of smooth cut.
For experiment two (CS) a set of monosyllabic stimuli with increasing complexity in head and coda but phonologically identical nucleus was produced with the German words <Schal, Stahl, Strahl, Schalt, Strahlt, Schalst, Stahlst, Strahlst> [ʃa:l, ʃta:l, ʃtra:l, ʃa:lt, ʃtra:lt, ʃa:lst, ʃta:lst, ʃtra:lst].
The overall length of the stimuli varied between about 400 ms and 620 ms with a stronger tendency for compensatory shortening / lengthening with increasing / decreasing complexity in the head.
All recordings were performed in a soundproofed studio using an Electro Voice 631B microphone and a Sony DAT recorder. All words were well pronounced (explicitly demonstrated) in focus position within the frame sentence <Ich habe das Wort _____ gesagt.> (I said the word _____.)[1]. The recordings were transmitted via Digidesign AudioMedia II and then segmented and downsampled to 20 kHz using Signalyze on the Macintosh.

Method

These stimuli, as well as a control stimulus (click signal: 5 ms, 1 kHz tone burst), were presented binaural using a Sennheiser HD 250 headphone under computer control (DEC VaxStation VS 3200, Distec DA-converter, Krohn-Hite 3750 filter) with 20 kHz sample rate and lowpass filtered at 6 kHz (24 dB/oct).
30 subjects had to listen and tap in synchrony to sequences built of 15 repetitions of the same stimulus with an inter stimulus interval of 700 ms and an inter sequence interval of 1400 ms. The stimulus sequences, chosen in random order regardless of the experiment they belonged to were grouped in blocks of 10. The subject starts the presentation of the next block by pressing the return key. Each stimulus sequence was given four times with at least two different intermediate sequences. A sequence was repeatedly presented as long as the subject did not start to tap. To register the taps a 5 x 10 cm capacitive sensory field was used. Before the presentation of the target stimuli subjects were familiarized with the data acquisition procedure using the click signal, the sound [pst] and the word <Schwimmst> [ʃβɪmst] as stimulus material.
Overall 25200 taps were registered. For analysis the taps to the first three and the last two presentations within a sequence were omitted (leaving 16800).

[1] n.b. that the position in the German phrase is not sentence final.

Subjects

30 subjects (13 female, 17 male) took part in the experiments. All of them had a 10 minute introduction on using the computer and the stimulus presentation program, none of them had participated in a former experiment on rhythm perception.

RESULTS

The data showed a large intersubject variability but according to 'Duncan's multiple range test' out of the 30 subjects 21 had been able to perform the experimental task as intended showing a low intrasubject variability of the tapping positions for the respective stimulus and not having unusual values for skewness and kurtosis.
The intersubject variability can be seen in Figure 1 which shows the tapping positions for the click stimulus. The control

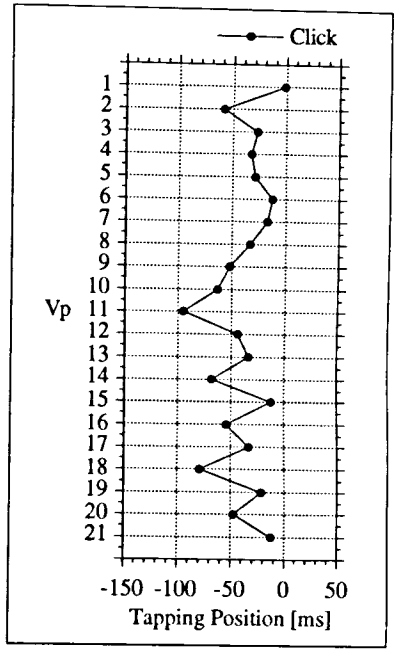


Figure 1: Tapping positions for the control stimulus (click signal) showing the amount of anticipation and the inter-subject variability.

stimulus had been presented to be able to compensate for the known effect of anticipation common in repetitive tapping task experiments. On average, the subjects tapped 39,27 ms before the physical onset of the click, which is in good agreement with findings reported elsewhere [9-12].
In Figure 2 the measured as well as the neutralized (anticipation corrected) tapping positions are given in relation to the durations of syllable head, nucleaus and coda for both experiments. An effect of the different stimuli on the location of the tapping position in relation to the stimulus onset can clearly be seen. This effect, however, may simply be caused by the different durations of the stimuli.

Experiment VQ

The stimuli for experiment one (VQ) were chosen to show whether or not the

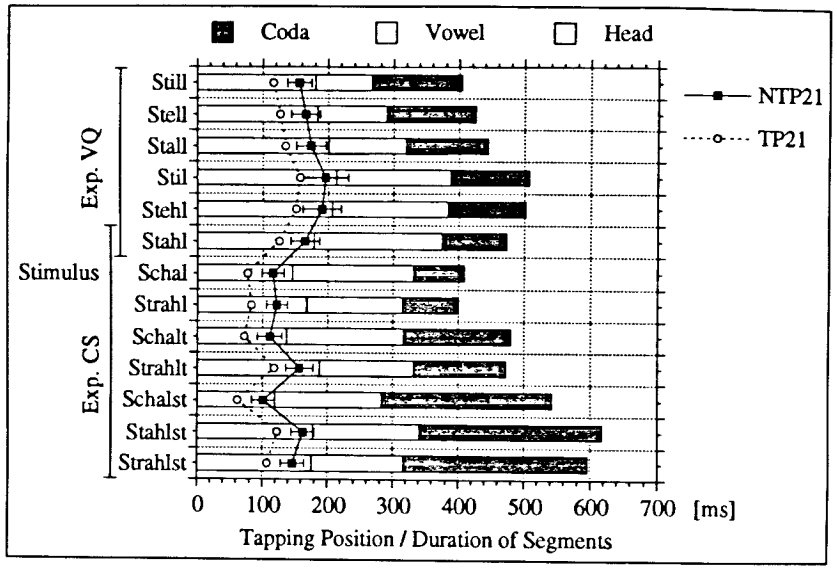


Figure 2: Mean (TP21) and neutralized (anticipation corrected) tapping position (NTP21 with SD) in relation to the stimulus onset with indicated duration of head, nucleus vowel and coda for the stimuli of experiments VQ and CS .

varying influence of different energy distributions known from experiments with synthetic stimuli can be replicated with naturally spoken stimulus material.
However, there seem to be no evidence that the differences in the abruptness of the vowel ending (i.e. [ʃtal] vs. [ʃtɑ:l]) or vowel quality (order: [ʃte:l] -> [ʃti:l] vs. [ʃtɪl] -> [ʃtɛl], second with later tapping position) are responsible for any differences in the measured tapping positions, although there is a slight tendency for the smooth cut stimuli - which are also longer - to show later ones.

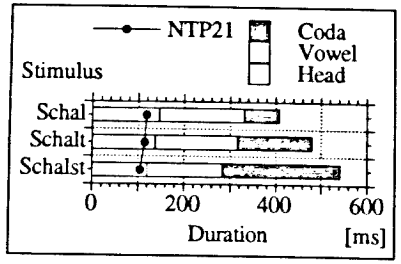


Figure 3: Variation of the coda showing no influence on the tapping position

Therefore this experiment with naturally spoken stimuli with identical shell does not replicate the findings with synthetic stimuli and holds no evidence that a different vowel quality or vowel quantity (abruptness) is in itself - in naturally spoken stimuli accompanied by compensational shortening or lengthening - of any influence on the p-center perception.

Experiment CS

The second experiment (CS) was carried out to investigate whether or not the internal structure of the head and the coda of a natural spoken syllable is of influence on the p-center-perception.
In accordance with the literature the duration of the initial consonance (head) clearly had an influence on the tapping position. However, looking at figure 2 the measured tapping positions do not indicate that the compositional structure of the head is of importance. This is also the case for the internal structure of the coda. Furthermore and in opposition to the literature the greatly varying duration of the coda in [ʃɑ:l, ʃɑ:lt, ʃɑ:lst] shows no influence at all (Figure 3).
Therefore experiment CS does not show any evidence that the complexity of the syllable shell is of importance for the p-center-perception.

CONCLUSION

Overall there is no evidence for any influence of the internal structure of a syllable on the p-center-perception. With respect to the used natural spoken stimuli the neutralized tapping positions closely follow the consonance-vowel transition, thus the duration of the initial consonance still seems to be the best reference to determine the location of the p-center.

REFERENCES

[1]Howell, P. (1988), "Prediction of p-center location from the distribution of energy in the amplitude envelope: Part I & II", *Perception & Psychophysics* , vol. 43, pp. 90-93 & 99.
[2]JANKER, P. M. (1989), "Der Einfluß von Segmentdauer- und Amplitudenmanipulation auf die P-center-Position einfacher CV-Silben", *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München,* vol. 27, pp. 71-141.
[3]JANKER, P. M., POMPINO-MARSCHALL, B. (1991), "Is the p-center influenced by 'tone'?", *Proc. 12th ICPhS* Aix-en-Provence, vol. 3, pp.290-293
[4]KÖHLMANN, M. (1984), "Rhythmische Segmentierung von Sprach- und Musiksignalen und ihre Nachbildung mit einem Funktionsschema", *Acoustica,* vol. 56, pp. 192-204.
[5]MARCUS, S. M.(1981), "Acoustic determinants of perceptual center (P-center) location", *Perception & Psychophysics,* vol. 30(3), pp. 247-256.
[6]POMPINO-MARSCHALL, B. (1989), "On the psychoacoustic nature of the P-center phenomenon", *Journal of Phonetics,* vol. 17, pp. 175-192.
[7]POMPINO-MARSCHALL, B. (1990), *Die Silbenprosodie. Ein elementarer Aspekt der Wahrnehmung von Sprechrhythmus und Sprechtempo,* Tübingen: Niemeyer.
[8]MORTON, J., MARCUS, S.M., & FRANKISH. C. R. (1976), "Perceptual centres (P-centers)", *Psychological Review,* vol. 83, pp. 405-408.
[9]ASCHERSLEBEN, G., PRINZ, W. (1992), "What gets synchronized with what in sensorimotor synchronisation", *Paper 13/1992,* MPI für psychologische Forschung, München.
[10]DUNLAP, K. (1910), "Reactions on rhythmic stimuli, with attempt to synchronize.", *Psychological Review,* vol. 17, pp. 399-416.
[11]RADIL, T. et al. (1990), "Stimulus anticipation in following rhythmic acoustical patterns by tapping.", *Experimentia,* vol. 46, pp. 762-763.
[12]JANKER, P. M. (1993), *Sprechrhythmus, Silbe, Ereignis. Eine experimentalphonetische Untersuchung zu den psychoakustisch relevanten Parametern zur rhythmischen Gliederung sprechsprachlicher Äusserungen,* Dissertation, LMU München.

Integrating Voice Quality and Tongue Root Position in Perceiving Vowels

John Kingston, Laura Walsh, Rachel Thorburn, and Christine Bartels
Linguistics Department, South College, University of Massachusetts, Amherst
Neil A. Macmillan
Psychology Department, Brooklyn College of the City University of New York

Abstract

In two experiments, we examined the perceptual interactions between the acoustic correlates of covarying articulations in vowels: tongue root advancement and tense-lax voice quality. These correlates prove to integrate because they both contribute to the perceived flatness or sharpness of a vowel's spectrum, and this integration enhances the contrast between vowels.

Introduction

Vowels articulated with the tongue root advanced are often also produced with a lax or breathy voice quality, whereas retracted tongue root vowels are produced with a tense or creaky voice quality; similar covariation is observed between tongue body raising and a lax voice quality [1]. The aryepiglottal ligament and membrane connect the tongue root to the arytenoid cartilages and may cause them to slide forward slightly and/or rock slightly apart, slackening or separating the vocal folds enough to lax the voice, when the tongue root is advanced or the body raised. However, because some languages, e.g. Dinka, exhibit independent contrasts for tongue root/body position and voice quality, it appears there is no necessary physiological interaction between the lingual and laryngeal articulations.

The experiments reported here assess an alternative, perceptual explanation for their covariation: that because lax or breathy voice causes energy to fall off more rapidly with increasing frequency in the source spectrum and advancing the tongue root lowers F_1, these articulations combine to bias a vowel's spectrum toward low frequencies, i.e. make it *flatter*, whereas a tenser or creakier voice quality and a non-advanced or retracted tongue root bias the vowel's spectrum toward higher frequencies, i.e. make it *sharper*. The covariation thus enhances the vowels' contrast along the *flat:sharp* dimension [cf. 2]. This explanation does not rely on a mechanical, physiological connection between the articulations that affect tongue root position (or body height) and the state of the vocal folds, but instead allows these articulations to be independently controlled by speakers. Kingston & Diehl [3, also 4] argue that speakers exert themselves to produce multiple articulatory differences between minimally contrasting phonemes when those articulations' acoustic correlates are similar enough psychoacoustically to integrate into higher-level perceptual properties as the flatness property proposed here.

Methods

Tense-lax variation in *V(oice) Q(uality)* was achieved by manipulating the percent of the glottal cycle in which the glottis is open and the additional energy reduction at 3 kHz in the source spectrum beyond the default decay of -6 dB/octave, i.e. the *O(pen) Q(uotient)* and *S(pectral) T(ilt)* parameters in the KLSYN88 terminal analogue synthesizer [5]. Variation in tongue root position was implemented simply through manipulation of F_1. The stimuli used in the two experiments reported here had similar ranges of F_1 values (Table 1), but they differed in what part of the voice quality continuum was paired with F_1: voice quality in the Tense experiment ranged from very tense to intermediate values along tense-lax continuum, whereas in the Lax experiment this dimension ranged from (overlapping) intermediate values to very lax. The overlap allows us to combine the results of the two experiments in constructing a map of how the entire range of voice qualities interacts perceptually with a narrower range of F_1s. Table 1 shows the 4x4 stimulus arrays defined by the orthogonal combination of F_1 and *VQ* values used in the two experiments. The size of the steps along the F_1 and *VQ* dimensions were approximately a jnd (at 70-80% correct). Other synthesis parameters were set so as to create a syllable of the shape [bVb], whose vowel was mid to high back in quality.

Table 1: Steady-state parameter values and 4x4 stimulus arrays for the Tense and Lax experiments: A-D = F_1 values (horizontal axis) and 1-4 and 5-8 = Ten(se) to Int(ermediate) and Int to Lax VQ values for the Tense and Lax experiments, respectively.

VQ			F_1			
			Advanced		Retracted	
	OQ	ST	470	484	499	514
Ten	29	-3	A1	B1	C1	D1
	33	-4	A2	B2	C2	D2
	39	-6	A3	B3	C3	D3
Int	53	-11	A4	B4	C4	D4
VQ			F_1			
	OQ	ST	450	468	506	536
Int	42	-7	A5	B5	C5	D5
	54	-11	A6	B6	C6	D6
	72	-17	A7	B7	C7	D7
Lax	90	-23	A8	B8	C8	D8

Two different groups of eight, paid, well-practiced, normal-hearing listeners participated in each experiment. Just a single stimulus was presented in each trial, to which the listener gave one of 2 responses, followed by a confidence judgment, and then by feedback. In the Lax experiment (run first) 16 alternating practice trials were followed by 96 randomized test trials; performance was assessed from the last 90 test trials/task/listener. In the Tense experiment, two blocks of 12 alternating practice trails followed by 66 randomized test trials were run for each condition, one early and one late. Performance was assessed from the last 60 test trials in each block, for a total of 120 trials/task/listener, an increase by one-third over the Lax experiment.

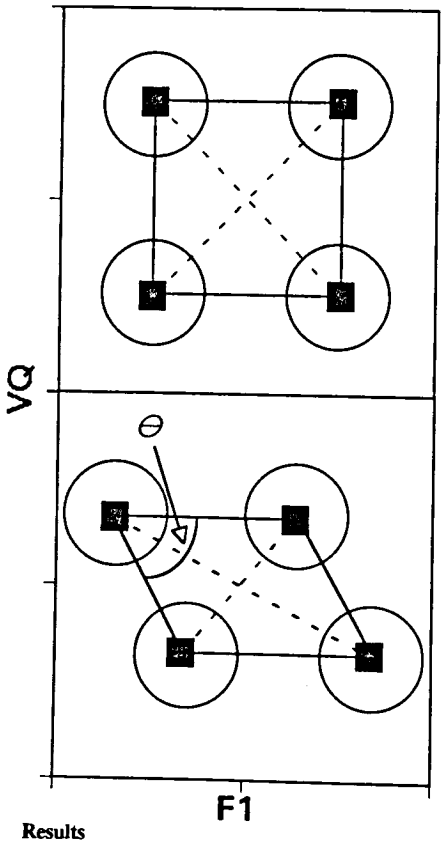
In the results reported here, listeners had to classify stimuli differing by one step in the 4x4 arrays in one of two ways: along just a single dimension or in a correlated fashion along both dimensions; these will be referred to as *single-dimension* and *correlated classification* tasks. There are 12 single-dimension classification tasks along each dimension, e.g. A1 vs A2 for *VQ* differences and A1 vs B1 for F_1 differences. In the correlated classification tasks, the correlation between F_1 and tense *VQ* could be either positive, e.g. A2 vs B1, or negative, e.g. A1 vs B2; there are 9 such tasks for each correlation polarity. Positive correlation corresponds to the natural covariation.

As in our previous work [6], we use detection theory [7] to model differences in how accurately our listeners classified the stimuli, because this theory provides the same estimate (d') of accuracy from each task. Furthermore, as d' estimates perceptual distance, it describes our listeners' performance as a mapping from the stimulus space defined by the dimensions F_1 and *VQ* to the two-dimensional *decision* space seen from above in Fig. 1, in which d's represent the distances between the mean locations of the stimuli (points in Fig. 1). The listener divides the decision space into response regions, but because various sources of noise produce trial-to-trial variability on both dimensions, the stimuli's perceptual values in the decision space form bivariate distributions of response likelihood (the circles in Fig. 1 suggest that the distributions are equal-variance, uncorrelated, bivariate normal).

We model our listeners' performance in terms of the d's obtained from 4 single-dimension and 2 correlated tasks that define each of the nine 2x2 subarrays of the larger 4x4 array in each experiment: the single-dimension d's determine the lengths of the sides of a parallelogram, the correlated d's the lengths of its diagonals. The rectangular arrangement in the upper panel of Fig. 1 is the arrangement of the stimuli when their perceptual values on one dimension do not depend on their values on the other, i.e. when the dimensions are *separable*. In the lower panel the means of the distributions are no longer arranged rectangularly, and the value of a stimulus on one dimension *does* depend on its value on the other; because the locations of the means of the distributions shifts in the space, we call the dimensions *mean-integral* in this case [6 cf. 9]. Comparing these two panels suggests that it is the (in)equality of d's obtained from the pair of correlated tasks representing the lengths of the diagonals in a 2x2 subarray that determines whether the stimulus dimensions are perceptually integral. The upper left angle (θ) of a parallelogram is a measure of mean-integrality: θ approaches 90° when the stimulus dimensions are perceptually separable but deviates from it when they're mean-integral, toward 0° when the negatively correlated task is

easier than the positively correlated vs toward 180° when the positively correlated task is easier. The θ that provided the best-fitting parallelograms to the correlated and single-dimension *d'* values was found by iteration; the fit of the correlated to the single-dimension data tests the model. If F_1 and *VQ* integrate into the property called flatness above, then listeners should be uniformly more accurate on the negatively than the positively correlated tasks, i.e. $\theta < 90°$.

Figure 1: Parallelograms for separable and mean-integral dimensions.



Results

Table 2 lists the θs obtained for each 2x2 subarray from each experiment for each 2x2 subarray. The fits are clearly much better in the Tense than the Lax Experiment, probably because the number of trials contributing to each *d'* was increased by a third and because the listeners were more thoroughly pre-tested and trained in that experiment.

Table 2. θs (in degrees) for the parallelogram models of each 2x2 subarray of the 4x4 arrays in the Tense and Lax experiments (with rms errors in d' = standard deviation units).

VQ		F_1		
		A:B	B:C	C:D
Tense	1:2	49 (.113)	34 (.148)	52 (.129)
	2:3	84 (.299)	62 (.149)	43 (.087)
	3:4	107 (.544)	120 (.340)	127 (.174)
Lax	5:6	159 (.147)	120 (1.027)	180 (.620)
	6:7	0 (.830)	15 (.758)	27 (.785)
	7:8	56 (.344)	79 (.564)	19 (.245)

Both halves of Table 2 show evidence of strong mean-integrality effects, i.e. *VQ* and F_1, do interact; however, the extent and direction of the interaction, measured by θ, varies systematically with location in the arrays. For the 2x2 subarrays from the tenser row pairs 1 vs 2 and 2 vs 3 in the 4x4 array in the Tense experiment (top), that θ is clearly less than 90° shows the stimuli in which laxness and F_1 negatively covary are easier to classify. However, θs are clearly greater than 90° for the parallelograms of the laxest pair of rows, 3 vs 4, but θs also are not much greater than 90°, indicating that the dimensions may be near separable with intermediate voice qualities. However, for the 2x2s drawn from rows 5 vs 6 in the Lax experiment (bottom), whose voice qualities overlap with those in rows 3 vs 4 in the Tense experiment, θs are clearly all much greater than 90°, indicating strong mean-integrality in the opposite direction. With yet laxer voice qualities, the direction of mean-integrality flips once more, as θs are all obviously less than 90° for 2x2s drawn from rows 6 vs 7 and 7 vs 8. Thus, at the tense and lax ends but not the middle of the voice quality continuum, we find the direction of mean-integrality expected if the acoustic correlates of tongue root advancement and voice quality enhance the contrast between vowels by integrating into the perceptual property, flatness. In the middle, these dimensions may be separable or integrate strongly in the opposite direction, i.e. into a property of the spectrum we could call *compactness* (low F_1 and tense voice being compact in contrast to the diffuse combination of high F_1 and lax voice) rather than into flatness.

Discussion

The flips observed in the direction of mean-integrality may follow from simple psychoacoustic properties of the stimuli in these experiments. Because the laxer the voice the more rapidly source-spectrum energy falls off with increasing frequency and the more advanced the tongue root the lower F_1 is, the difference in amplitude between the first harmonic and those immediately above it should vary directly with flatness. To model these effects, we fit a line to the peaks of the first four harmonics for each stimulus. The slope of this line estimates how their relative amplitudes differ as a function of both voice quality and F_1: this slope should be more shallow or even more negative the laxer the voice and the lower the F_1, i.e. the flatter the vowel's spectrum. The difference in slopes between the stimuli in each of the correlated tasks measures how much they differ in flatness, and thus estimates the psychoacoustic value of flatness for predicting differences in relative accuracy on these tasks. When these slope differences were fit to the observed mean *d'*s for the correlated tasks in a simple regression model, a positive change in *d'* of 0.37/dB difference in slope was obtained Although highly significant [$F(1,34) = 7.93$, $p = 0.008$], the proportion of variance accounted for was only 0.18. When the slope differences are fit only to the correlated data from the 2x2s at the tense and lax ends of the arrays, where θs were less than 90°, the relationship is more strongly positive, 0.53 change in *d'*/dB, and the proportion of variance accounted for jumps substantially, to 0.42 [$F(1,18) = 13.00$, $p = 0.002$]. The direction of mean-integrality at the two ends of the tense-lax may therefore arise from differences in the psychoacoustic property flatness, even if some other property is psychoacoustically more salient in the middle of this continuum.

[Work supported by NIH Grant R-29-DC01708 to first author and NSF grant DBS92-12043 to second author; earlier versions: Thorburn, R., *et al.* (1994), J. Acoust. Soc. Am. 95 (Abstract) and Walsh, L. *et al.* (1995), J Acoust. Soc. Am. 97 (Abstract).]

References

[1] Denning, K. (1989) *The diachronic development of phonological voice quality*, Ph.D. diss. Stanford.

[2] Jakobson, R., G. Fant & M. Halle. (1952), *Preliminaries to speech analysis*, Cambridge: MIT Press.

[3] Kingston, J. & R.L. Diehl. (1994), "Phonetic knowledge", *Lg.* 70: pp. 419-454.

[4] Diehl, R.L., J. Kingston, W.A. Castleman. (1995), "On the internal perceptual structure of phonological features: the [voice] distinction", *J. Acoust. Soc. Am.*, 97 (Abstract).

[5] Klatt, D.H. & L.C. Klatt. (1990), "Analysis, synthesis, and perception of voice quality variations among female and male talkers", *J. Acoustic. Soc. Am.* 87, pp. 820-857.

[6] Kingston, J. & N.A. Macmillan. (1995), "Integrality of nasalization and F_1 in vowels in isolation and before oral and nasal consonants", *J. Acoust. Soc. Am.* 97, pp. 1261-1285.

[7] Macmillan, N.A. & C.D. Creelman. (1991), *Detection theory: A user's guide*. Cambridge: Cambridge UP.

THE ILLUSION OF PERCEPTUAL RESTORATION OF MISSING PHONEME IN CHILDREN.

I.V.Korolyova, G.G.Shurgaya
Institute of Ear, Throat, Nose and Speech, St.-Petersburg, Russia
State University, St.-Petersburg, Russia

ABSTRACT

It was shown that 5-6 year old children perceptively restored a missing phoneme in a word more often than adults. Adults had the influence of the localization and the acoustic properties of the missing phoneme on the perceptive restoration but children had not. The children could seldom name the missing phoneme, but often they intuitively localized it correctly repeating the word with mispronounced phoneme in the right place.

Perceptual restoration of a missing phoneme (PRMPh) is the auditory illusion in which one hear a nonexistent part of a word [1]. In studies of the PRMPh the adults listen to the speech in which some segments are replaced by nonspeech signals and usually can not detect the missing segment, often perceiving the word with missing phoneme as intact. It was shown that with the adults the PRMPh depends on the acoustic characteristics of the phoneme and the replacing signal, phoneme localization, speech material [1,2]. The children's ability to restore the missing phoneme and the exploration of their PRMPh particularities were the purpose of our study.

Method.

Subjects. The participants were 20 children (10 boys and 10 girls from 5,2 to 6,0 years of age) and 20 adults (10 men and 10 women from 16,0 to 29,2 years of age). The subjects were native Russian speakers with no problems in hearing and speech. All children could not read but had some notion about "word" and "letter" and they knew some letters of the Russian alphabet.

Stimuli. All the test words were three-syllable ones and familiar for the children. The words were spoken clearly by a man and recorded on the audio tape. Then the words with the missing phonemes were constructed by a programmable signal analyzer (NEC, San-ei Ltd., Japan) using the procedure that had been described by Samuel [2]. The target phoneme was located visually on the word oscillogram on the analyzer display and auditorily over headphones. The determined phoneme was bracketed by two pointers and the digital root mean square amplitude (DRMSA) of its central 20 msec was computed. Then each point within the bracketed segment (10 kHz sample rate) was replaced with a random number between +DRMSA and -DMSA. The test words (intact and with a replaced phoneme) were recorded on audio tape in an occasional way. The main test consisted of 90 words, 30 of which were intact, 60 words were with a missing phoneme, the latter being: 20 - with the missing first consonant (stop or fricative), 20 - with the missing middle consonant, 20 - with the missing vowel (the second phoneme in the word, stressed or unstressed).

Procedure. The test stimuli were played to the subjects binaurally from audio tape over stereo headphones. The testing was individual. The nature of the stimuli had been explained beforehand. The adults were told that they would hear the words in some of which one of the letters (because as a rule they did not know what a phoneme is) was replaced by a noise. Their task was to repeat the word, to say whether the word was intact or not, and to name the missing letter or to determine its approximate localization in the word. That explanation was not clear for the children and they were told that they would hear "correct" and "incorrect" words with one "bad" letter. Their task was to repeat the word, to determine if it was "correct" or not, to name the "bad" letter or to say if it was at the beginning, in the middle or at the end of the word. The children mostly repeated strictly what they had heard. The subjects' responses were protocoled by the researcher. During the training session before the subjects were made to listen to each word they were told whether the word was intact or not and what phoneme had been replaced. Then the subjects listened to the words and answered without preliminary explanation. When the subject understood the task and his responses (whether the word was intact or not) were found to be not chance ones, the main test began.

Results and discussion

The children and the adults perceived the intactness of the words very well but the children had a slightly less number of correct responses (see Table). The subjects often failed to notice the absence of a phoneme in the word with a replaced phoneme and perceived it as intact, i.e. they perceptually restored the missing phoneme. With the children PRMPh was more frequent than with the adults. With the adults PRMPh was more frequent for consonants in the middle position than for initial consonants, least of all for vowels. With the children the influence of these factors were insignificant.

Samuel [2] had proposed to use the interactive schema theory by Rumelhart [3] for PRMPh explanation. In this theory the perceptual-cognitive system is viewed as a collection of active processing units. The units are data structures that are activated either by other processing units (top-down) or by incoming sensory information (bottom-up). The perceptual process is essentially an evidence-gathering task, perception occurs when a given schema has received sufficient evidence. The pieces of evidence can come from bottom-up and top-down sources. In general, the top-down information can be characterized as expectation and the bottom-up as confirmation. The PRMPh is an evidence for top-down processing of speech because it occurs more often for central phoneme where the initial part of the word helps to advance the correct hypothesis. The influence of the acoustic characteristics of the replacing

Table. The indexes of the different response types for children and adults (the ratio to the quantity of the stimuli of the accordant group in percents)

The types of responses	Children	Adults
The intact words		
1. "The word is intact"	90	96
2. "The word is not intact"	10	4
The words with a missing phoneme		
1. "The word is intact" (PRMPh) *	45	17
init. cons./mid.cons./vowel */*/*	49/54/36	21/32/1
2. "The word is not intact" *	53	83

The note: init. cons. - the indexes for the words with the replaced initial phoneme; mid. cons. - for the words with the replaced middle phoneme; vowel - for the words with the replaced vowel; * - for these indexes the differences between the children and adults are significant at level < 0.01.

sound on the PRMPh is an evidence for bottom-up processing.

However, the use of the interactive model is not enough for the explanation of our results. Actually, the greater number of PRMPh with the children must mean their greater use of "top-down" mechanism. However this does not agree with the fact that the influence of the missing phoneme localization was insignificant with the children. On the other hand it does not mean that the children have a more effective "bottom-up" mechanism because the influence of the phoneme acoustic characteristics (the replacing signals were perceived as noise and sounded rather like consonants) was insignificant with the children.

The analysis of the responses to words with the missing phoneme when the word distortion had been detected can help to explain the results. The identical responses with the adults and the children were: the correct and incorrect determination of a missing phoneme, the subject could not localize the missing phoneme in the word, the word was not understood. The children had a insignificant number (3%) of correct determinations of a missing phoneme. For both groups: the best determination was for initial consonants, the worst - for vowels. As a rule the children could not name the replaced phoneme (correctly or incorrectly) or indicate its approximate localization in the word. The adults detected a missing phoneme incorrectly in words with replaced vowels, they usually named the neighbouring (mostly preceding) consonant. The number of the words that the children did not understand was nearly the same as with the adults. As a rule the subjects did not understand the words with replaced vowels (mostly stressed). In that case the children gave another word, the adults - rather a pseudo word more often.

Moreover the children had two additional response types. In one type which was named as "the unconscious correct determination of the replaced phoneme" the children said that the word was distorted and then they repeated the word with the distortion of the corresponding phoneme. But they could not name isolated the phoneme that had been replaced. In the other type called "the word is distorted, but is understood" the children first mispronounced the word and then they pronounced the correct word . For example, one said "I heard [f'ufka] (the pseudo word), but it must be [d'evochka] (a girl)". These responses were characteristic for words with a replaced stressed vowel.

The data analysis suggested that the adults and the children have and may use various speech information. Apparently, children have a word description as a whole image which they mainly use in perception. "That description includes essential details and it is insensitive for local signal distortion. This provides for the effective perception of intact words and words with a missed phoneme (the number of non understood words was almost equal for both groups). Evidently, an intact stressed vowel is necessary for using this description, because its replacement prevented the comprehension of the word. Using word description as a whole image with significant details in perception makes it possible to detect the distortions in words. However, it is not effective for that purpose therefore the children showed the greater number of PRMPh than the adults. If it is necessary to localize the distortion, especially to name the distorted phoneme that description is not enough. Though the children could not name the replaced phoneme they actually perceived the distortion of the phoneme correctly repeating the word with the distortion of the corresponding phoneme. Apparently, the replaced phoneme detection requires the using of word description as a sequence of phonemes. 5-6 year old children do not

have it and we believe that this description develops with reading ability.

Nevertheless 5-6 year old children seemed to have some information about phonemes and its acoustic cues, because sometimes they named the replaced phoneme (usually incorrectly). Moreover, when the children said that the word was "incorrect" and then repeated it with distortion of the corresponding phoneme ("the word is distorted, but it is understood"). They imitated specific phoneme features: the replaced vowel was pronounced loudly and with a drawl, the voiced consonants - as voiceless ones. Besides the children as well as the adults perceiving the words with a replaced vowel considered that the neighbouring consonants had been replaced .

Adults have a word description as a whole image and as a sequence of phonemes. The latter are not used at normal speech perception because it requires much time. We suppose that adults and children use the word description as a whole image in processing in normal perception. If perception requires the use of phoneme description (in the test with PRMPh or another task requiring the phoneme analysis of speech) the adults use it together with the main whole image description singling out and identifing the word phonemes. But that analysis does not keep up with main processing based on whole image of word description which provides for the fast achievement of the principal purpose - word comprehension. This causes the effect of the PRMPh. The speech processing is interactive: the upper levels of the system advance a hypothesis about the incoming signal and require the necessary information from the bottom levels. This information is used for the hypothesis verification and correction as well as for the creation of new hypotheses. The final decision is taken when the hypothesis and the results of the analysis coincide, usually on the base of the particular information (sufficient in this situation). The interaction of top-down and bottom-up mechanisms causes the influence of the replaced phoneme localization and acoustic characteristics on the PRMPh. Apparently, the insignificant influence of these factors on the PRMPh and the pecularietes of children's responses mean that 5-6 years old children have not yet developed both these mechanisms and their interaction.

References

[1] Warren, R.M. (1970), "Perceptual restoration of missing speech sounds. ",Science, N167. pp.392-393. .

2] Samuel, A.G. (1981), "Phoneme restoration: insights from a new methodology. "J. of Exp. Psychol.: Gen., vol.110, pp. 474-494.

[3] Rumelhart, D.E.(1977),"Toward an interactive model of reading. "In: Dornic S.(Ed.).Attention and performance. VI.Hillsdale. N.-Y.:Erebaum.

GENERAL AUDITORY PROCESSES MAY ACCOUNT FOR THE EFFECT OF PRECEDING LIQUID ON PERCEPTION OF PLACE OF ARTICULATION.

Andrew J. Lotto and Keith R. Kluender
Dept. of Psychology, University of Wisconsin-Madison, USA

ABSTRACT

The perception of syllable-initial /d/ and /g/ can be affected by the composition of the preceding syllable [1]. It has been suggested that this result demonstrates the existence of a mechanism which compensates for coarticulation. In a series of perceptual experiments, including the use of an avian species, this effect is shown to be due to general auditory processes which may be described as frequency contrast.

I. INTRODUCTION

Immediate context plays an enormous role in perception of acoustic information specifying phonemes. Many context effects in speech perception have the appearance of serving as compensation for effects of coarticulation. For example, perception of syllable-initial /d/ and /g/ can be affected by the composition of preceding acoustic information such that, for a series of synthesized consonant-vowel syllables (CVs) varying in onset characteristics of the third formant (F3) and varying perceptually from /da/ to /ga/, subjects are more likely to perceive /da/ when preceded by the syllable /ar/, and to perceive /ga/ when preceded by /al/ [1]. This effect has been found for speakers of Japanese who cannot distinguish between /l/ and /r/ [2] and for prelinguistic infants [3]. The received interpretation of these findings has been that listeners are somehow sensitive to articulatory implementation. The following experiments assess the degree to which these perceptual effects are specific to qualities of human articulatory sources.

II. EXPERIMENT 1

Others have concluded that the perceptual effect described above results from a mechanism specialized to compensate for vocal tract constraints through the use of tacit knowledge of articulatory dynamics [1], or that the effect is due to the recovery of vocal tract actions [3]. The first experiment was performed to examine the extent to which perceptual effects of coarticulation are dependent upon maintaining a unitary source. Presumably a mechanism serving to accommodate coarticulation would be specific to the speech stream of a single speaker. After all, the perception of a single talker's speech often must take place in the presence of other acoustic sources including other talkers.

Stimuli.

A 10-step series of /da-ga/ syllables varying in F3 onset frequency was synthesized with endpoint stimuli based on the natural productions of a male talker. For these CVs, the onset frequency of F3 varied from 1800 to 2700 Hz in 100 Hz steps and changed linearly over 80 msec to a steady state value of 2450 Hz. All other parameters remained unchanged between members of the series. The first formant rose from 300 to 750 Hz and F2 decreased from 1650 to 1200 Hz over 80 msec. Fundamental frequency was 110 Hz from onset until decreasing to 95 Hz over the last 50 msec. Total stimulus duration of these synthesized syllables was 250 ms. These stimuli were preceded by two separate natural speech versions of preceding /ar/ and /al/. One version was produced by a relatively tall adult male talker (190 cm in height) after whom the /da-ga/ series was modelled with an average *f*0 of 110 Hz. The other was produced by a relatively short female talker (157 cm) with an average *f*0 of 210 Hz. Formant frequency values averaged about 12% higher for the female VCs.

Procedure.

Twelve English-speaking subjects participated in a two-choice forced identification task. All stimuli were passed through a 16-bit D/A at a sampling rate of 10 kHz, low-pass filtered at 4.8 kHz, amplified, and randomly presented under the control of a microcomputer over headphones at 75 dB SPL. Listeners responded after each disyllable by pressing either of two buttons labelled 'd' and 'g'.

Results.

Identification functions depicting results are displayed in Figure 1. In addition to replicating the original finding for disyllabic stimuli for which both syllables are modeled after the same talker [1], we found that the effect extends to disyllabic stimuli for which there is a clear mismatch between sources for the first (♀) and second (♂) syllable. It appears that the context effect of preceding /al/ and /ar/ is not critically sensitive to the entire stimulus complex being produced by a single talker or even by a modestly similar vocal tract.

III. EXPERIMENT 2

Because results from the first experiment indicated that precise matching of articulatory/acoustic characteristics for the initial VC and following CV was not essential for the effect, Experiment 2 tested whether simple schematized nonspeech versions of F3 information alone could affect labeling of following CVs.

Stimuli.

Two sine-wave glides were synthesized. Each was matched in frequency characteristics to the F3 transitions of the /al/ or /ar/ syllables of Experiment 1 and were matched in amplitude to the energy within a critical band of the center frequency of F3. The same series of synthetic /da-ga/ syllables used in the Experiment 1 followed the sine-wave glides.

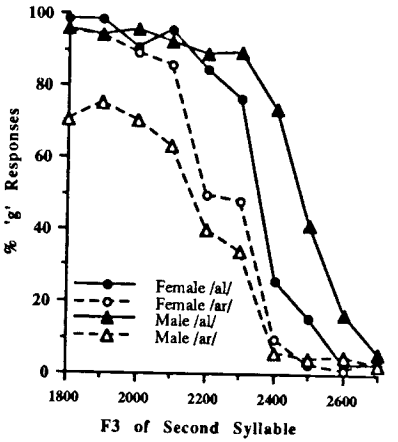


Figure 1. Identification functions from Experiment 1

Procedure.

The procedure was identical to that of Experiment 1. Thirteen subjects identified the consonant as 'd' or 'g'.

Results.

Identification functions depicting results are displayed in Figure 2. Although the context effect of sine-wave glides was not as great as the effect for full spectrum speech /al/ and /ar/, there is a significant effect of glide type as reflected in responses to the synthesized /da-ga/ series. Even a preceding stimulus that consists of only a sine-wave caricature of a portion of rich full-spectrum speech (F3) is adequate to give rise to the context effect on perception of the following CV as /da/ or /ga/.

IV. EXPERIMENT 3

Conceivably, the sine-wave glides in the second experiment may have resembled speech sufficiently to activate

some special speech mechanism. To address this possibility, simpler constant-frequency tones were used as precursor stimuli for this experiment.

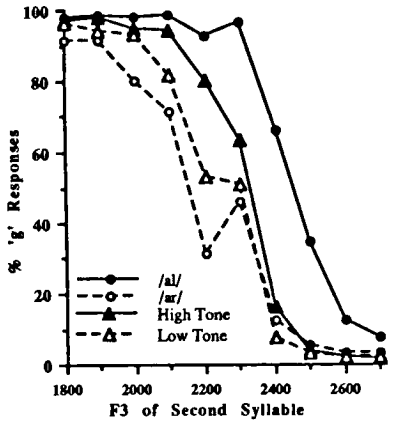


Figure 2. Identification functions for Experiment 2

Stimuli.

Two constant-frequency sine waves with frequencies equal to the offset frequency of F3 for the /al/ and /ar/ speech stimuli were used as precursor stimuli. Sine-wave amplitude was set to the RMS amplitude of the full-syllable male /al/ and /ar/ from Experiment 1. In contrast to the nonspeech stimuli of Experiment 2, these stimuli were not matched to the speech stimuli either in frequency change over time or in amplitude as a function of frequency. The same series of synthetic CV syllables used in Experiments 1 and 2 and varying perceptually from /da/ to /ga/ followed the sine-wave tones.

Procedure.

As in the preceding experiments, subjects (16) responded by pressing 'd' or 'g'.

Results.

Identification functions depicting the results are displayed in Figure 3. The difference in frequency of preceding tones resulted in a shift in identification boundaries despite the fact that the only characteristic the tones shared with /al/ and /ar/ was that they contained substantial energy in the region of F3. Although the context effect of constant-frequency sine waves was not as great as the effect for full spectrum speech /al/ and /ar/ in this experiment, the effect of these nonspeech precursors is of comparable magnitude to Mann's [1] for natural speech stimuli.

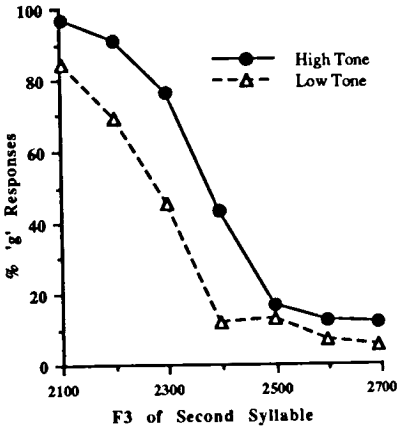


Figure 3. Identification functions for Experiment 3.

V. EXPERIMENT 4

Results from the first three experiments indicate that the original effect of preceding /al/ and /ar/ on perception of following /da/ or /ga/ generalized quite broadly. These findings suggest that the effect could potentially be general to auditory perception. If this is so, similar results should arise when nonhuman animals serve as subjects. In the last experiment, Japanese Quail (*Coturnix japonica*) were used to test the generality of this effect.

Procedure.

Two Japanese Quail were trained (by operant procedures) to peck a lighted key when presented with either the syllable /da/ or /ga/ and to refrain from pecking it when presented with the alternative syllable (/ga/ or /da/, respectively). After the birds learned to reliably peck differentially to /da/ or /ga/, they were presented with test syllable pairs consisting of the synthesized /al/ or /ar/ followed by one of the ambiguous intermediary members of the /da/-/ga/ series.

Results.

Histograms representing the results of this study are shown in Figure 4. Avian responses evidenced an effect of the preceding syllable such that 'labeling' shifted to more /ga/ responses following /al/. The most parsimonious explanation for this result is that the effect is due to general auditory processes of frequency contrast.

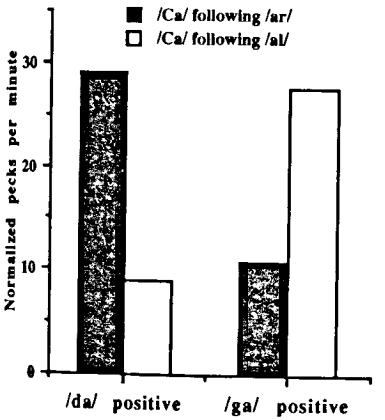


Figure 4. Histograms of Normalized Peck Rates from Experiment 4.

DISCUSSION

Results from these four experiments suggest that this context effect is not specific to speech in any conspicuous way. At least part of the original effect is likely due to some general property of the auditory system.

One potentially could describe this hypothesized general auditory effect as frequency contrast. The data from these experiments are consistent with this description. Following a high tone or glide, the identification boundary shifts toward higher F3 onsets (more are perceived as low, "ga") relative to when the preceding tone is lower.

At this time, the term is not meant to suffice as explanation or to designate a specific process. The point is simply that the effect reported first in Mann [1] and replicated here may be due, at least in part, to general auditory processes. If the effect is of a general nature, then the earlier findings with Japanese listeners [2] and with infants [3] should not be unexpected.

In general, contrast may be useful for many of the acoustic transformations arising from coarticulation. It appears that all coarticulation in speech results in frequency assimilation. Consequently, contrastive perceptual effects between contiguous speech sounds should generally be adaptive.

V. REFERENCES

[1] Mann, V. A. (1980), "Influence of preceding liquid on stop consonant perception", *Perception and Psychophysics*, vol. 28, pp. 211-235.

[2] Mann, V. A. (1986), "Distinguishing universal and language-dependent levels of speech perception: Evidence from Japanese listeners' perception of English 'l' and 'r'", *Cognition*, vol. 24, pp. 169-196.

[3] Fowler, C. A., Best, C. T., & McRoberts, G. W. (1990), "Young infants' perception of liquid coarticulatory influences on following stop consonants", *Perception & Psychophysics*, vol. 48, pp. 559-570.

VI. ACKNOWLEDGMENTS

This work was supported by NIDCD Grant DC-00719 and NSF Young Investigator Award DBS-9258482 to the second author.

PERCEPTUAL MAPPING AND VOWEL NORMALISATION

Robert H. Mannell
Speech, Hearing and Language Research Centre
Macquarie University, Sydney, Australia

ABSTRACT

A large number of vowel tokens in an /h_d/ frame were synthesised using a formant synthesiser. Each experimental condition consisted of a simulated "speaker" for whom three parameters, vowel space size, F0 range and higher formant frequencies, were characteristic of a male, female or "neutral" voice. These three parameters were either matched or mis-matched with each other in terms of their target vocal gender. For each "speaker" the tokens were evenly spaced on the F1/F2 plane. For each condition 20 speakers were asked to identify the English vowel that each token most sounded like. The resulting vowel phoneme perceptual maps were compared to examine the interacting effects of each of the parameters on vowel normalisation.

INTRODUCTION

Mannell [1] examined the perceptual mapping of Australian English long and short monophthongs. In that experiment there were two conditions which simulated a male and a female speaker. The male speaker's characteristics were based on measurements of 19 Australian English vowels spoken in /h_d/ context by 172 male speakers [2][3]. From these measurements it was possible to determine the limits of the male Australian English vowel acoustic space (henceforth referred to as the male frame) on the F1/F2 plane as well as appropriate F3 values for each F2 value. From these measurements a vowel plane was derived in F1/F2/F3 space and two sets of vowel tokens were derived on this plane representing long (300 ms) and short (150 ms) monophthongs in an /h_d/ frame. Each set of "male" vowels consisted of all possible vowel qualities on the F1/F2 plane separated by 100 Hz in both dimensions and within the constraints imposed by the defined male frame. F4 and F5 were fixed at 3500 Hz and 4500 Hz respectively. The female frame was derived from the male frame by multiplying the F1 and F2 maxima and the higher formant values by 1.2 to produce a larger frame size. The resulting frame size was compared with measurements obtained from 12 female speakers of Australian English to confirm that the derived vowel space was a valid representation of actual female data. The F0 contour was held constant for all tokens (male and female) at an average "gender neutral" value of 160 Hz. All of the tokens were generated by a parallel formant synthesiser [4] using specialised synthesis-by-rule software written especially for this experiment.

Perceptual contours (25%, 50% and 75% identification) were derived for every vowel phoneme and the resulting contour maps for the male and female long and short vowel spaces were compared. Particular attention was paid to the 50% identification contours or "predominance boundaries" [5] within which the identification of a particular phoneme predominated (ie. ≥ 50%). The male and female perceptual spaces were very similar in shape, differing mainly in the size of the spaces. The female spaces were shown to closely match the male spaces when both the long and short vowel female spaces were uniformly divided by a factor of 1.2. The match was even closer when a -50 Hz correction was made for the normalised female F1 values (possibly correcting for differences in male and female oral/pharyngeal tract lengths). Normalisation to a particular vocal type (in this case, simulated male and female voices) had clearly occurred as there were numerous vowel pairs that were identical in terms of their F1/F2 values but which resulted in consistently different vowel phoneme identifications for the male and female voices. For example, the "long" (300 ms) vowel with an F1 of 600 Hz and an F2 of 1900 Hz lies within the female /ɜ:/ predominance boundary and in the male /æ/ predominance boundary. Clearly some vocal factor or combination of factors has triggered different normalisation strategies for these two vowels based on the listeners' perception of differences between the two "speakers". Since F0 has been held constant for this experiment the trigger for the different normalisation strategies must depend upon one or both of the only two parameters which differentiate the two "speakers" in this experiment, vowel frame size and higher formant frequencies.

To examine whether the vowel frame size or the higher formant values had the stronger effect on vowel normalisation, Mannell [1] presented a series of "male" vowels representing a selection of vowels across the entire male vowel space. This was then followed by a series of 33 vowels which had "female" higher formant values but which had F1/F2 values which wholly fit within the male vowel frame. These 33 vowels were then followed by the female version of the vowel referred to above, which had an F1 of 600 Hz and an F2 of 1900 Hz and typically female higher formant values. It was assumed that 33 preceding vowels would be sufficient to alert the listeners to the new voice and to familiarise them with the "female" voice. This familiarisation would not, however, be based on the vowel frame size information as the test vowel would not be preceded by any vowels with F1/F2 values outside the male frame and thus exclusive to the larger female frame. If the listeners normalised fully to the female voice then this vowel should be heard as /ɜ:/, if the normalisation was based on the preceding male voice then the vowel should be heard as /æ/, and if the higher formants were responsible for partial shifting of the normalisation strategy towards that for the female voice then a mixture of /ɜ:/ and /æ/ responses should occur. The result was that 17 out of 20 subjects perceived /æ/ and there were no /ɜ:/ responses (the remaining three subjects heard /e/). The insertion of one high F2 (non-male-frame) vowel in the list of vowels preceding the test vowel reversed this effect, with more than 50% of the subjects perceiving the vowel as /ɜ:/ as would be appropriate for a female voice (this effect will be examined in more detail in future experiments). What seemed clear from is result was that normalisation appears to be strongly influenced by the listener's determination of the vowel frame size. Further, only one high F2 front vowel appears to be necessary to establish appropriate normalisation procedures. This last observation is consistent with the point normalisation hypothesis of Nearey [5].

These experiments, whilst pointing out the importance of vowel frame size in the normalisation of vowels, did not examine the effect of F0 on normalisation, nor did they examine the ways in which vowel-frame, F0 and higher-formant parameters interact during the process of vowel normalisation.

METHOD

In the present experiment the same procedure was followed as outlined on the first page of this paper, but with the following differences. Firstly, whilst the points on the male spaces were still separated by 100 Hz, on the female spaces the individual points were separated by 120 Hz, resulting in similar numbers of tokens for the male and the female spaces. Secondly, and most importantly, there was a much larger number of "male" and "female" conditions. The conditions varied with respect to F0, vowel-frame-size and higher formant values. The F0

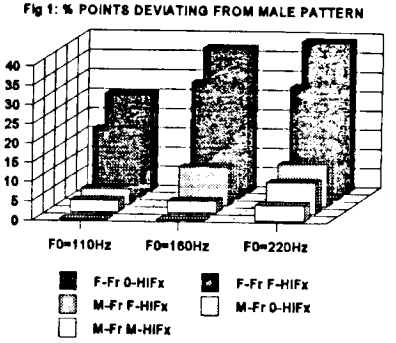
parameter was one of three pitch contours with mean F0 values of 110Hz ("male"), 160Hz ("neutral") and 220Hz ("female"). The vowel frame size parameter was either a male frame, or a female frame (as described above). The higher formants (F3/F4/F5) were either typically "male" or "female" or entirely absent (ie. F1/F2 two formant synthetic vowels). The conditions tested are summarised in table 1.

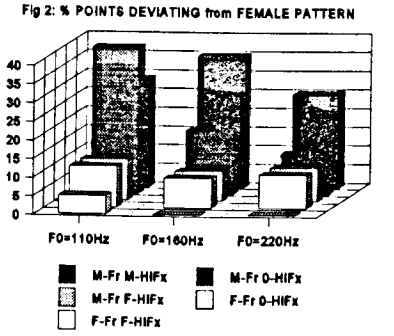
#	Frame	F0	Hi Fx
1	M	110	M
2	M	110	absent
3	M	160	M
4	M	160	absent
5	M	220	M
6	M	220	absent
7	F	110	F
8	F	110	absent
9	F	160	F
10	F	160	absent
11	F	220	F
12	F	220	absent
13	M	110	F
14	M	160	F
15	M	220	F

Table 1. Summary of experimental conditions.

300 listening subjects (phonetically naive, native speakers of Australian English, 20 subjects per test condition) were asked to identify the each token orthographically from a closed set of possible responses. The tokens were presented individually via headphones in a sound treated room. Predominance boundaries were determined for each vowel monophthong phoneme for each condition to produce a set of 30 perceptual maps (15 conditions, long vs short vowels). Data-point-by-data-point χ^2 comparisons of points within the area common to the male and female spaces were made for each relevant pair of conditions. Differences between conditions were determined as the number of data-points significantly different at p=0.01 and are expressed below as the percentage of total data points.

RESULTS





In the above figures "M-Fr" and "F-Fr" refer to male and female frames respectively, whilst "M-HiFx", "F-HiFx" and "0-HiFx" refer to male, female and missing F3/F4/F5 respectively.

DISCUSSION

The vowel frame size has the greatest effect on normalisation and thus vowel perception. This is most clearly seen when female frame data is measured relative to the most male condition (110Hz, male frame, male higher formants: see figure 1). This effect is also strong when male frame data is measured relative to the most female condition (220Hz, female frame, female higher formants: see figure 2) but in this case higher F0 values pull the percept for male-frame data much more strongly in the direction of the female pattern than low F0 pulls the percept for female-frame data in the direction of the male pattern. This can presumably be explained by the fact that all male-frame F1/F2 values could also be female values whilst the extreme front and low vowels in female-frame data cannot be perceived as male and so strongly mark the speaker as female.

An F0 of 110Hz tends to pull the perceptual pattern in the direction of the male pattern. Conversely an F0 of 220Hz tends to pull the perceptual pattern in the direction of the female pattern. This effect is strongest when F0 is reinforced by matching frame-size or higher formant values.

Appropriate male higher formant values reinforce the male perceptual pattern, missing higher formants weakens that pattern somewhat whilst inappropriately female higher formants for male-frame tokens has a strong effect on the perceptual space, pulling it in the direction of the female pattern. The male-frame tokens with female higher formants result in a perceptual pattern intermediate between the male and the female pattern (the perceptual patterns are as distant from the male pattern as they are from the female pattern). On the other hand, missing higher formants appears to consistently enhance the perceived femaleness of female-frame tokens relative to tokens with "female" higher formant values. This may be because the female high formant model utilised in this experiment is not a good model of female vowel productions and so confuse the listening subjects.

The maximum deviations between male and female perceptual patterns of no more than 40% is due to the overlap of the vowel spaces of central and back vowels. The deviations tend to occur at front and low vowel boundaries and result in the shifting of the boundaries to higher (female) or lower (male) frequencies.

CONCLUSION

All three parameters have some effect on normalisation processes during the perception of vowels. The frame-size parameter has the strongest effect but generally requires the support of at least one other factor (F0 or high formants) to produce the strongest male or female patterns. The effect of F0 on vowel perception is greatest when the vowel is otherwise ambiguous.

REFERENCES

[1] Mannell, R.H. (1988), "Perceptual space of male and female Australian English vowels", *Proc. 2nd. Australian International Conf. Speech Science and Technology*, Sydney, Nov. 1988, 22-27.

[2] Bernard, J.B. (1970), "Toward the acoustic specification of Australian English", *Zeitschrift fur Phonetik, Sprachwissenschaft und Kommunikations-forschung*, Band 23, Heft 2/3.

[3] Bernard, J.B. and Mannell, R.H. (1986), "A study of /h_d/ words in Australian English", *Working Papers*, SHLRC, Macquarie University.

[4] Clark, J.E., Summerfield, C.D. and Mannell, R.H. (1986), "A high performance digital hardware synthesiser", *Proc. 1st. Australian Conf. Speech Science and Technology*, Canberra, November 1986, 342-347.

[5] Nearey, T.M. (1977), *Phonetic feature systems for vowels*, PhD dissertation, Univ. of Connecticut.

CONTEXTUAL AND LEXICAL EFFECTS IN THE IDENTIFICATION OF FRICATIVES

Noël Nguyen

*Laboratory of Psycholinguistics, University of Geneva, Switzerland**

ABSTRACT

This research studies the respective roles of the phonetic context and of the lexicon in the identification of [s] and [š]. In a first experiment, subjects had to identify fricatives combined with different vowels. Results showed that vowel effects occurred at a post-perceptual stage of processing only. In a second experiment, the lexical status of the carrier string and the adjacent vowel were both manipulated. Lexical and vowel effects appeared to be non-additively combined in the identification of fricatives.

INTRODUCTION

Phonetic segments are not produced independently of one another. Coarticulation is in fact considered to be a source of great acoustic variability. However, listeners do not seem to have major problems in identifying segments in the speech chain. It is assumed that a processing mechanism of some kind enables them to factor out any influence that segments have on each other. This is confirmed by experimental evidence which shows that listeners do indeed make compensatory adjustments in the identification of a segment as a function of its phonetic context.

In this domain, a great deal of attention has been devoted to context-dependent variations in the identification of coronal fricatives. In an /s/+V sequence in particular, the fricative acoustic shape is known to be sensitive to the rounded/unrounded character of the subsequent vowel. When combined with a rounded vowel, /s/ is most often produced with an anticipatory rounding of the lips, which results in a lengthening of the front cavity, and therefore in a lowering of the noise frequency in the fricative spectrum. Perceptual studies have shown that such acoustic variations appear to be compensated for in an identification task, as a fricative on a [s]-[š] continuum is more frequently identified as /s/ in the vicinity of a rounded vowel [4, 10].

This research studies the respective roles of the phonetic context and of the lexicon in the identification of /s/ and /š/ in French. Two major issues have been adressed. First, I have attempted to determine the level of processing at which the phonetic context comes into play in a fricative identification task. While some theories of speech perception (e.g. [2]) postulate that one segment can have a direct influence on how an adjacent segment is perceived, it may be also hypothesized that the phonetic context is only taken into account at a post-perceptual stage of processing, as a decision bias [7]. Second, I have examined how the influence of adjacent segments may interact with contextual effects of another nature, namely lexical effects. Previous work has shown that the /s/-/š/ categorical boundary can shift as a function of the lexical status of the carrier string [3]. However, to my knowledge, there is still no study on how informations from both the phonetic context and the lexicon are combined, especially when these two types of information provide conflicting cues concerning the identity of the fricative.

*e-mail: nnguyen@fapse.unige.ch

EXPERIMENT 1

The goal of Experiment 1 was to examine whether vowel effects in [s] and [š] identification can also be observed in French. In addition, an attempt was made to characterize the stage of processing at which such contextual effects are likely to occur.

Material and Method

An 11-step [s]-[š] continuum was created from one natural [s] and one natural [š], using the procedure described in [3]. Each of the fricative stimuli was then combined with either [a], [i] or [u] (natural tokens). In total, the material was made up of 33 CV syllables which were presented to 30 listeners in a fricative identification task (forced choice).

It was assumed that the identification scores would vary as a function of the vowel degree of lip rounding (more "s" responses for [u]) as well as of the vowel place of articulation (less "s" responses for [i], see [10]).

Results

The proportions of "s" responses for each fricative stimulus and each adjacent vowel are presented in Figure 1 (the [s] and [š] endpoint stimuli are on the left and on the right, respectively). This figure shows that the vowel effects on fricative identification described in previous work were replicated here. As predicted, the [s]-[š] categorical boundary moved toward the [š] endpoint in the context of a back rounded vowel ([u]), and toward the [s] endpoint in the context of a front unrounded one ([i]). Differences in the mean percentage of "s" responses (averaged over the 11 fricative stimuli), as a function of the adjacent vowel, were significant ($F(2,58) = 13.93, p < .001$).

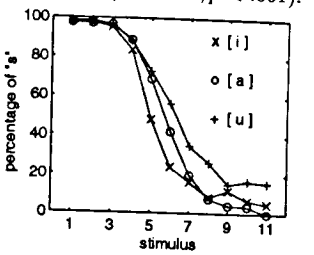


Fig. 1. Effect of vocalic context on fricative identification.

To determine at which processing stage the vowel came into play in the identification of fricatives, two models making different predictions on this point were tested against the results. The first model was formally identical to Massaro's Fuzzy Logical Model of Perception (FLMP) [5, 6]. In this model, the fricative and the adjacent vowel were considered as two independently-evaluated sources of information. It was hypothesized that the vowel had an effect on the fricative categorization, but did not have any influence on how the fricative was internally represented in terms of features. The fricative acoustic structure was represented by one continuous-valued feature (F_1), ranging from 0 ([š]) to 1 ([s]). The vowel acoustic structure was also represented by one feature (F_2), ranging from 0 ([i]) to 1 ([u]). The degree of match with the prototypes /s/ and /š/ was determined by means of multiplicative combination rules.

The second model was itself similar to the featural modifier model (FMM), also presented in [6]. In this model, the context was represented by one parameter, C_j, which was combined with the fricative stimulus information at the prototype matching stage by means of the following rules:

$$/s/ = F_1^{C_j}$$
$$/š/ = (1 - F_1)^{C_j}$$

This model postulates that the vocalic context has a direct influence on

the fricative stimulus-to-feature mapping ([6]; see also [7, p. 369]). The root mean square deviation (RMSD) between the observed and predicted values was calculated for each model and each subject. It appeared that the FLMP made more accurate predictions than the FMM for a majority of subjects (24 out of 30). The average RMSD was significantly higher for the FMM (.1) than for the FLMP (.05; $t(29) = 4.87, p < .001$).

In a second step, the observed results were submitted to a sensitivity analysis. It was assumed that, if the adjacent vowel directly affected sensitivity[1] in the fricative identification task, than the discriminability between adjacent stimuli along the [s]-[š] continuum would differ depending on the vowel category [5, 9]. The fricative discriminability was determined on the basis of the d' measure. A d' value was computed for each pair of adjacent fricative stimuli and each vowel from the proportions of "s" responses [5]. The cumulative sum of d' averaged over the 30 subjects for each vowel is presented in Table 1.

Table 1: Mean cumulative d'

V2	N	mean	SD
a	30	5.23	1.16
i	30	5.45	1.31
u	30	4.91	1.42

Although differences in the mean cumulative d' were observed between vowels, a one-factor ANOVA showed that these differences were not significant. Thus, the d' analysis and the model assessment gave convergent results. They both tended to indicate that the vowel played a significant role at the decision stage only, in the identification of fricatives.

EXPERIMENT 2

The goal of this experiment was to characterize the way in which a potential influence of the lexicon on fricative categorization, would be combined with that of the phonetic context. The main issue was whether one of these two effects would dominate the other, when they push the subject to make opposite predictions about the fricative category.

Material and Method

The material was composed of 16 11-step [s]-[š] continua embedded in different carrier sequences. For one half of the continua, the [s] endpoint formed a word and the [š] endpoint a nonword (ex.: *soulier-choulier*). For the other half, the [š] endpoint formed a word and the [s] endpoint a nonword (ex.: *sapeau-chapeau*). The lexical status of the carrier string was orthogonally combined with two other variables, namely *a)* the identity of the vowel following the fricative ([a], [u]) and *b)* the position of the fricative within the carrier string (initial, median). The stimuli were presented to 26 subjects in a fricative identification task.

Results

Figure 2 shows the mean percentage of "s" responses for the two adjacent vowels, when the [s] endpoint formed a word (upper line), and when it formed a nonword (lower line). There was a main effect of the adjacent vowel ($F(1,25) = 11.54, p < 0.005$), a main effect of the lexicon ($F(1,25) = 28.17, p < 0.001$), and a significant vowel × lexicon interaction ($F(1,25) = 11.87, p < 0.005$). Thus, there was an effect upon the fricative categorization both of the following vowel and of the lexical status of the carrier string. Moreover, these two effects did not appear to be statistically independent of each other.

An attempt was made to characterize the origin of this interaction, by testing two different models against the results, as in Experiment 1. The first model was derived from the FLMP already presented above. The only difference was that a third feature (F_3), representing the degree of "s-lexicality" was introduced.

[1] And not only bias.

The values of F_3 ranged from 0 ([s] endpoint forms a nonword) to 1 ([s] endpoint forms a word).

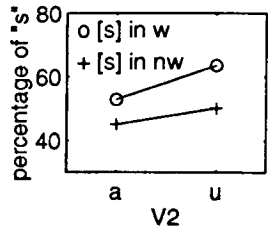


Fig. 2. Effect of vocalic context and of lexicon on fricative identification

The second model was identical to the first one, except that the features were combined with each other using an additive rather than a multiplicative rule. This model predicted that the vowel effect and the lexical effect would not interact with each other. The results showed that the multiplicative rule accounted for the observed results more accurately than the additive rule for 24 out of 26 subjects. The average RMSD was significantly lower for the first model (.08) than for the second one (.11; $t(25) = 6.26, p < .001$).

CONCLUSION

Our results tend to indicate that while the quality of the adjacent vowel may modify the respective probabilities of alveolar and post-alveolar responses in an [s]/[š] identification task, it does not have any influence on sensitivity. Such results are in good agreement with segmental models of speech perception in which the phonetic context comes into play at a post-perceptual level only, to bias responses in one direction or another [1, 7]. When the lexical status of the carrier string was manipulated as well as the adjacent vowel, both factors had an influence upon the fricative categorization. Although the interaction that was observed between these two factors might be simply due to a floor effect (the percentage of alveolar responses could not come down below a certain threshold, whatever the phonetic and lexical context), it may also indicate that vocalic and lexical cues to the fricative category are non-additively combined in a fricative identification task.

ACKNOWLEDGEMENTS

Work supported by the Swiss Federal Office for Education and Science (project # 93.0351). I am grateful to Uli Frauenfelder for helpful comments.

References

[1] Elman, J.L., and McClelland, J.L. (1988). "Cognitive penetration of the mechanisms of perception: compensation for coarticulation of lexically restored phonemes", *J. Mem. Lang.* **27**, 143–165.

[2] Fowler, C.A. (1984). "Segmentation of coarticulated speech in perception", *Percept. Psychophys.* **36**, 359–368.

[3] McQueen, J.M. (1991). "The influence of the lexicon on phonetic categorization: stimulus quality in word-final ambiguity", *J. Exp. Psychol.: Hum. Percep. Perf.* **17**, 433–443.

[4] Mann, V.A., & Repp, B.H. (1980). "Influence of vocalic context on perception of the [š]-[s] distinction", *Percept. Psychophys.* **28**, 213–228.

[5] Massaro, D.W. (1989). "Testing between the TRACE model and the Fuzzy Logical Model of Perception", *Cog. Psychol.* **21**, 398–421.

[6] Massaro, D.W., & Cohen, M.M. (1983). "Phonological context in speech perception", *Percept. Psychophys.* **34**, 338–348.

[7] Nearey, T. (1990). "The segment as a unit of speech perception", *J. Phonetics* **18**, 347–373.

[8] Nguyen, N., Hoole, P., & Marchal, A. (1994). "Regenerating the spectral shape of [s] and [š] from a limited set of articulatory parameters", *J. Acoust. Soc. Am.* **96**, 33–39.

[9] Repp, B.H. (1981). "Two strategies in fricative discrimination", *Percept. Psychophys.* **30**, 217–227.

[10] Whalen, D.H. (1981). "Effects of vocalic formant transitions and vowel quality on the English [s]-[š] boundary", *J. Acoust. Soc. Am.* **69**, 275–282.

DO TONGUE TWISTERS OCCUR NATURALLY?

M.O'Kane, P.E.Kenne and H.Pearcy
The University of Adelaide, Adelaide, Australia

ABSTRACT

We define a generalized tongue twister which encompasses both the 'traditional' multi-word tongue twisters as well as introducing the notion of an intra-word tongue twister. We examine the occurrence of generalized tongue twisters in spontaneous speech occurring in court cases and conclude (not surprisingly) that generalized tongue twisters occur infrequently.

INTRODUCTION

A tongue twister is a phrase which is difficult to say because of the repetition of a certain letter or certain similar sounds. Considerable research has been carried out on the problems of reading and producing tongue twisters [1,2,3], much of this concerned with the way in which speech (silent or read) slows down at a tongue twister. Table 1 below lists a number of tongue twisters (some of these appear in [1,2,3]). While all of the examples are syntactically and semantically well formed, they appear to be unlikely to occur in verbal discourse (unless it's discourse about tongue twisters). Given this, do tongue twisters occur naturally in verbal discourse, particularly if speakers are under stress? We hypothesized that tongue twisters would generally be avoided.

We took verbatim court transcripts and analyzed them for the occurrence of tongue twisters using the operational definition that "a tongue twister is a sequence of sounds in which the same or a closely-related consonant phoneme occurs more than twice within a short phoneme distance." Under this definition "the tall toddler talked timidly to Tom" is a tongue twister, while "the small toddler spoke bravely to her sister" is not. The definition also allows for what might be called "intra word tongue twisters"; an example would be the word "phenomenon."

Table 1. Examples of multi-word tongue twisters

1	The bootblack brought the black book back
2	Which witches wished wicked wishes
3	Five French friars fanned the fainting flea
4	The Swiss wristwatch strap shop shuts soon
5	The wild wind whipped Whit from the wharf
6	A dreary don droned on dully
7	Peter picked a pick of pickled peppers
8	She sells sea shells by the sea shore
9	The spacious zoo sits beside a sandy seashore
10	Barbara burned the brown bread badly
11	Francis Forbe's father fries five flounders
12	Nasty Noreen noticed the neat note
13	The puzzled priest processed the perplexing paper
14	The tired dentist dozed but he drilled dutifully
15	Naughty Nan's knitting knotted nighties
16	The press published the poem and promised to pay for permission
17	Pack a pair of purple pampers
18	The talented teenager took the trophy in the tournament
19	The sparrow snatched the spider swiftly off the ceiling
20	The taxis delivered the tourists directly to the tavern

The transcripts of two court cases (c1 and c2) were used as data. Sizes of these cases are shown in table 2. Each transcript was divided into a number of phrase blocks, as defined by the punctuation, and each of the phrase blocks which contained at least five words was considered a candidate to be a tongue twister.

Table 2. Case details

Case	Words	Phrase blocks	Candidate phrase blocks
c1	202000	29368	15630
c2	500000	49760	18276

The majority of the tongue twisters in table 1 have the property that a single phoneme occurs in the word initial position for at least 50% of the words, for example, numbers 1, 7 and 10 in table 1. (Note that not all tongue twisters have this property, for example, numbers 8 and 19 in table 1.) A tongue twister such as number 8 is characterized by the alternation (or close distance between) a number of minimal pairs (she/sea; sells/shells); number 1 also exhibits this property (back/black). We searched the transcripts for phrases with either of these properties. Both tests did not did not have requirement of a word-initial consonant, and the minimal pair test was also relaxed to test for phrases with at least two minimal pairs, which also includes repeated words. The results for th*m*iese tests are summarized in table 3.

Table 3. Phoneme and nimal pair results

Case	Majority phoneme	Minimal pair
c1	22	888
c2	100	1894

Of the 22 examples of potential multi-word tongue twisters found for case c1, the 'twister sound' was the diphthong /aI/ in seven cases and the semi-vowel /w/ in six cases, with the remainder being small numbers of various consonants. (The results for case c2 are similar - approximately 85% of the twister sounds are accounted for by /aI/ and /w/.) When the examples were examined individually, none was considered to be particularly difficult to say; the most difficult probably being "when we went to organize a workshop." Given that vowels and semi-vowels generally are not sounds over which people stutter [4], it is not surprising that these examples were not seen as difficult. Tables 4 and 5 respectively give examples of phrases with a majority twister sound, and phrases having a number of minimal pairs.

Table 4. Majority twister sound examples

1	it is pretty clear is not it
2	is that it is not in Turkey
3	when we went to organize a workshop
4	some would say surprisingly enough
5	the date that this accord
6	someone had particular stainless steel skills
7	how did he represent himself

Table 5. Minimal pair examples

1	it is pretty clear is not it
2	is that it is not in Turkey
3	that is about as high as you can put it
4	they may be but that would have to be demonstrated
5	is that not what is suggested by the words
6	in interpreting the results in the area
7	it was on the basis of the geochemistry

We also examined the transcripts for the occurrence of intra-word tongue twisters. We looked for the occurrence of patterns of the form NxNxN, where N is a nasal consonant, and x is anything other than a nasal. We also looked for patterns such as NxxNxN and NxNxxN as well as PxPxP, where P is any plosive consonant. Table 6 lists all the patterns which were considered. Table 6 also shows the number of words containing a given pattern, and the number of distinct words containing that pattern; for example, the entry 66/20 for NxxNxN for case c2 indicates that there are 66 occurrences of words containing the pattern NxxNxN, but there are only 20 distinct

words in this set of 66. Examples of intra word tongue twisters are given in table 7.

Table 6. Intra-word tongue twisters

Pattern	c1	c2
NxNxN	192/26	306/29
NxxNxN	38/16	66/20
NxNxxN	12/7	14/9
NxxNxxN	6/6	7/5
PxPxP	210/39	349/63
PxxPxP	223/55	382/70
PxPxxP	134/45	215/53
PxxPxxP	156/44	250/64

Table 7. Examples of intra word tongue twisters

Pattern	Examples
NxNxN	prominent
NxxNxN	convenient
NxNxxN	mentioning
NxxNxxN	inconvenienced
PxPxP	typical
PxxPxP	probably
PxPxxP	independent
PxxPxxP	complex

What characterizes a tongue twister, and why are they difficult to produce? We asked subjects to rank the tongue twisters in table 1 by order of difficulty of reading them aloud. Table 1 is ordered by decreasing order of perceived difficulty of production, for example, (1) "The bootblack brought the black book back" is thought to be more difficult to read than (2) "Which witches wished wicked wishes." We did not test whether subjects perceptions corresponded with difficulty of production.

Haber and Haber [3] in examined the production errors of a number of subjects reading tongue twisters suggest (but do not test) a number of hypotheses about what makes a tongue difficult to say:

• twisters such as the "Swiss wristwatch strap shop shuts soon" and "which witches wished wicked wishes?" in which both initial and final consonant clusters are alternated rather than initial clusters only have more substitution errors. These tongue twisters are ranked by our subjects as the two most difficult twisters to read.

• the intercession of unstressed syllables between alternated sounds decreases the twister property, for example "five French friars fanned the fainting flea" is harder than "Francis Forbe's father fries five flounders." Notice that "Nasty Noreen noticed the neat note" and "Naughty Nan's knitting knotted nighties" do not tend to support this hypothesis, however, nasal consonants do not tend to be twister sounds with great difficulty. Furhter support for this hypothesis is given by the ranking of "The sparrow snatched the spider swiftly off the ceiling" (19) relative to the other twisters involving fricatives and sibilants.

• the manipulation of reiterated words with minimal pair matches bootblack-black-back or sells-shells produces errors;

• some vowels, when juxtaposed cause difficulty in production for example, toy boat; wristwatch; strap shop;

• some syntactic patterns, probably those involving a sequence of primary stresses (black book back; five French friars fanned) invite twisted tongues;

• some sounds, namely the fricatives and sibilants above all, may be inherently difficult to manipulate rapidly. This hypothesis is supported by the rankings given in table 1.

Subjects were also separately asked to rank the phrases given in tables 4 and 5 by order of difficulty of reading aloud, and as for table 1, these tables are also listed in decreasing order of difficulty of reading.

CONCLUSION

Utterances exhibiting tongue twister properties occur very infrequently in verbal discourse, and the majority of those utterances which do exhibit one or more twister properties use a vowel as the twister sound. Intra word twisters are also infrequent.

REFERENCES

[1] Hanson, V.L., Goodell, E.W. and Perfetti, C.A. (1991),"Tongue-twister effects in the silent reading of hearing and deaf college students", *Haskins Laboratories Status Report on Speech Research* SR-107/108,July-December 1991, pp.171-180.

[2] McCutchen, D. and Perfetti, C.A. (1982), "The visual tongue-twister effect: phonological activation in silent reading", *Journal of Verbal Learning and Verbal Behaviour*, vol.21, pp.672-687.

[3] Haber, L.R. and Haber, R.N. (1982), "Does silent reading involve articulation? Evidence from tongue-twisters", *American Journal of Psychology*, vol.95, pp.409-419.

[4] Hunt, J. (1861), *Stammering and Stuttering*, republished Hafner Publishing Company New York 1967.

RATE-DEPENDENT PERCEPTION OF VOT: AUDITORY CONTRAST OR RATE NORMALISATION

Jörgen Pind
Faculty of Social Sciences, University of Iceland, Reykjavík, Iceland

ABSTRACT

Listeners are sensitivity to the temporal variability of speech. The mechanisms underlying this sensitivity are, however, unclear. One theory holds that listeners perceive temporal speech cues by "taking into account" the speech context. Another theory holds that rate-dependent perception is based on auditory contrast. An experiment on the perception of VOT in Icelandic indicates that auditory contrast is not a sufficient explanation for rate-dependent speech perception.

INTRODUCTION

A striking aspect of speech sounds is their context-dependent nature. The realisation of individual sounds is heavily dependent on other neighbouring sounds. One factor which has been shown to influence the behaviour and manifestation of speech segments is speaking rate. Speech sounds compress and expand with changes in speaking rate. Such changes in the durations of speech sounds pose a potential problem for the listener, especially as regards the perception of *temporal* speech cues, speech cues which are defined by their duration. How can the listener disentangle those durational properties of speech sounds which are phonemic, intrinsic to the phonetic message, from those which are due to extrinsic factors such as speaking rate?

Research over the past decades has shown that listeners are sensitive to the temporal structure of speech. In particular, experiments on rate-dependent perception, show that listeners' percepts are often influenced by speaking rate [1].

One unresolved issue is the nature of the underlying mechanism responsible for these rate-dependent adjustments. One theory holds that these reflect a process of "taking into account" analogous to that often posited for visual perception [2], where the perceptual system engages in a thought-like process to establish the perceptual boundaries e.g. the VOT boundary separating /g/ from /k/. In fact, it turns out that the rate-adjustments seen e.g. for VOT are typically less than such a model would imply [3,4].

Another theory put forward by Diehl and Walsh [5] claims that rate adjustments in perception are in fact not properly interpreted as speech-specific adjustments, but rather in terms of a "general auditory principle" of durational contrast. Focusing on the /ba-wa/ distinction (cued by vowel transition duration) these authors claim that the perceptual boundary separating these two syllables should move to a longer transition if followed by a longer vowel, since the long vowel would tend to make any particular transition duration appear shorter than if it were followed by a short vowel. Experiments using non-speech analogs have been taken to support to this theory.

If, indeed, rate-dependent speech perception is primarily a process involving auditory contrast the question arises as to what the domain of the contrast-inducing speech segment is. If it is to be possible to predict the extent of rate-dependent adjustment in perception it is necessary to establish how far the contrast-inducing elements extends beyond the segment of interest, e.g. word-initial VOT.

In one-syllable utterances (commonly employed in studies of rate-dependent perception) the correlation between the duration of the vowel and the perceived speech rate is high — the shorter the vowel, the faster the speech rate. This relationship does not, however, necessarily hold when we consider whole words. Consider thus a language like Icelandic which makes a distinction between phonemically long and short vowels and consonants. This distinction is cued by the duration of vowels and consonants, in many cases by the relationship of vowel and consonant durations. Research has shown that a higher-order invariant of vowel to rhyme duration will account for the perception of quantity in the face of extensive transformations of rate [4,6].

If the major contrast-inducing factor is that of the following vowel it is possible, using suitably chosen Icelandic words, to arrange for changes in vowel duration to *either* cue vowel quantity *or* speaking rate. Using such stimuli it should be possible to distinguish between the two theories of rate normalisation and auditory contrast. If the theory of rate normalisation ("taking into account") is correct only the vowel durations signifying a rate change should lead to a shift in the phoneme boundaries for VOT. If the auditory contrast theory is correct both manipulations, whether rate-specific or quantity-based, should lead to comparable changes in VOT boundaries since both involve the same contrast of VOT to the following vowel segment.

METHOD

Stimuli

This experiment made use of synthetic speech made with the Sensimetrics SenSyn™ synthesiser, a version of the Klatt cascade/parallel formant synthesiser [7]. The synthesiser was run in the cascade configuration. Six VOT stimulus continua were made containing three different vowel durations. In three of the continua this vowel duration was a cue for speaking rate, in the other three the identical vowel durations were a cue for different vowel quantities (see Figure 1). The words synthesised were tokens of the words 'gaka' [ka:ka], (nonsense word) 'gagga' [kak:a] (to cackle), 'kaka' [k^ha:ka] (cake) and 'kagga' [k^hak:a] (car, acc. sg.).

The different transformations of rate and quantity were accomplished in the manner shown in Figure 1. In the base stimulus the vowel was 260 ms long (including any word-initial VOT) and the closure was 140 ms long. The initial syllable was thus 400 ms long. In this syllable the vowel to rhyme ratio is thus 260/400 = 0.65 which is appropriate for the perception of a word with a long vowel followed by a short consonant. The initial syllable was followed by a 140 ms long [a] for the second syllable. Depending on the duration of the VOT this word would either be perceived as 'gaka' or 'kaka'.

In the Quantity series the duration of the initial syllable was kept constant at 400 ms while the vowel was shortened, first to 200 ms (in this case the closure duration was increased to 200 ms) and then to 140 ms (closure duration 260 ms). The latter stimulus has a vowel to rhyme ratio of 0.35, appropriate for words with a short vowel and following long consonant, i.e. the words 'gagga' and 'kagga'.

In the rate series the same vowel durations were used as in the Quantity series but, contrary to the Quantity series, other segments were also shortened to the same extent as the initial vowel. In this series the vowel to rhyme ratio is therefore kept constant at 0.65 in all three stimulus continua. In the two stimulus series, Quantity and Rate, the very same manipulations of vowel duration can in one case be traced to a change in phonemic make-up, in the other to a change of speaking rate.

The steady state formants of the vowel [a] had the following values. F1 was 750 Hz, F2 1280 Hz and F3 2425 Hz. Appropriate transitions for a velar place of articulation were synthesised at the beginning of both vowels and also at the end of the first vowel. The fundamental frequency of each stimulus was fixed at 100 Hz.

All six continua had variable VOT for the word-initial velar stop ranging in 5 ms steps from 15 ms to 70 ms, made by replacing the voiced excitation with aspiration and noise excitation in the region of F2 and F3 and by increasing the bandwidth of F1 from 90 Hz (the default) to 200 Hz. The shortest VOT of 15 ms consisted of a 10 ms word-initial burst followed by 5 ms of silence. The total number of stimuli in the experiment thus amounted to 3 (vowel durations) × 2 (stimulus series, quantity or rate) × 12 (VOT steps) = 72. Notice that one stimulus series, at the very top of Figure 1, is the same in both the Quantity and the Rate series.

The stimuli were recorded on two tapes with an inter-stimulus interval of 2.5 seconds. One tape contained the three Quantity series the other the three Rate series in randomised order.

Subjects

Twelve subjects took part in the experiment, two members of staff at the University of Iceland and ten undergraduate students of Psychology. All subjects reported normal hearing.

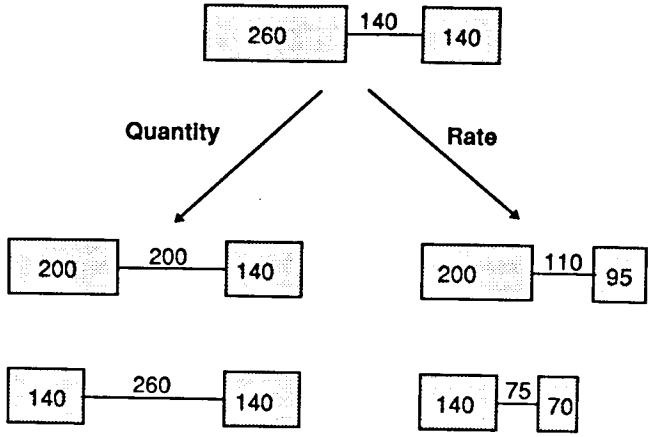


Figure 1. Schematic diagrams illustrating the structure of the stimulus continua used in the present experiment. The shaded boxes denote vowels, the lines connecting them the closures. Numbers refer to the durations of the segments in ms. Note that identical transformations of vowel durations (260 → 200 → 140 ms) signify either a change of quantity or a change of rate.

Procedure

Six subjects listened to the Quantity tape followed by the Rate tape. For the other six subjects the order was reversed. The testing took place in a quiet room. Subjects listened to the stimuli, which were played at a comfortable listening level, over Sennheiser HD-530-II circumaural headphones, and indicated their responses in forced-choice tests by marking appropriate fields on answer sheets.

RESULTS AND DISCUSSION

Pooled identification curves for 11 subjects (one being left out, showing highly irregular response patterns) are presented in Figure 2. The left-hand figure shows the responses for the Quantity series, the right-hand figure the responses for the Rate series. The figures show the percentage of /ka-/ responses as a function of the duration of VOT in the different continua. Phoneme boundaries for individual subjects were calculated using the method of probits. Table 1 shows the average location of the boundaries for 11 subjects.

A two-way repeated measures ANOVA (series × vowel duration) shows that the effect of series is not significant $F(1,10) < 1$ while that of vowel duration is, $F(2, 20) = 8.565$, $p < 0.01$. The interaction is not significant, $F(12,20) = 1.014$, $p = 0.38$. Pairwise comparisons, in all cases using the Bonferroni correction, reveal that none of the means in the Quantity series differ significantly from each other. In the Rate series the VOT boundaries are significantly different in the 260 and 140 ms continua, $F(1,10) = 14.934$, $p = 0.018$, not significantly different in the 200 and 140 ms continua, and just misses significance in the 260 and 200 ms continua, $F(1,10) = 10.317$, $p = 0.054$. Though both series, Quantity and Rate, show a shift towards shorter VOT boundaries with shorter vowel duration only in the Rate conditions do these shifts achieve statistical significance.

Table 1. Average VOT phoneme boundaries (in ms) for eleven subjects.

	Vowel duration		
Series	260 ms	200 ms	140 ms
Quantity	39.27	37.57	36.56
Rate	39.25	36.83	34.25

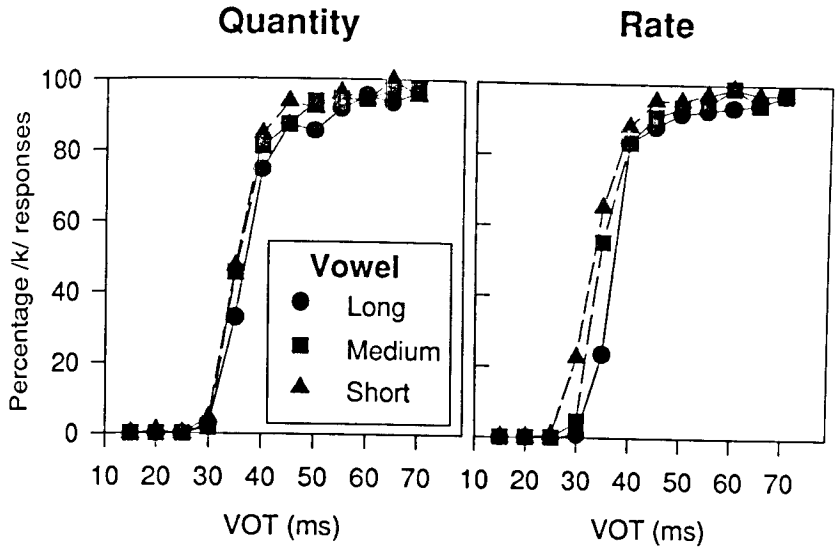


Figure 2. Pooled identification curves (11 subjects) in the present experiment. The figure on the left shows the results for the Quantity series, the figure on the right the results for the Rate series.

Table 1 reveals a shift of 2.7 ms in the location of the VOT phoneme boundary (from 39.27 to 36.56 ms of VOT) and of 5 ms in the Rate series (from 39.25 to 34.25 ms).

The results of the present experiment show that durational changes in a vowel following word-initial VOT do not all exert the same influence on the location of the VOT boundaries. If such changes in vowel duration can be traced to changes in speaking rate then the VOT boundaries show the typical effect of rate-dependent normalisation. The same changes in vowel duration, when expressing changes in vowel quantity, do not lead to significant rate-normalisation. The present results can most easily be explained by assuming that the rate-dependent perception of VOT operates by a process of "taking account of" the speaking rate.

ACKNOWLEDGEMENT

This research was supported by the Icelandic Science Foundation and the Research Fund of the University of Iceland.

REFERENCES

[1] Miller, J. L. (1987), "Rate-dependent processing in speech perception". In A. W. Ellis (Ed.), *Progress in the Psychology of Language*, Vol. III (pp. 119–157). Hove: Lawrence Erlbaum Associates.

[2] Epstein, W. (1973), "The process of 'taking-into-account' in visual perception", *Perception*, vol. 2, pp. 267–285.

[3] Miller, J. L., Green, K. P., & Reeves, A. (1986), "Speaking rate and segments: A look at the relation between speech production and speech perception for the voicing contrast", *Phonetica*, vol. 43, pp. 106–115.

[4] Pind, J. (1995), "Speaking rate, voice-onset time and quantity: The search for higher-order invariants for two Icelandic speech cues", *Perception & Psychophysics*, vol. 57(3), in press.

[5] Diehl, R. L. & Walsh, M. A. (1989), "An auditory basis for the stimulus-length effect in the perception of stops and glides", *Journal of the Acoustical Society of America*, vol. 85, pp. 2154–2164.

[6] Pind, J. (1986), "The perception of quantity in Icelandic", *Phonetica*, vol. 43, pp. 116–139.

[7] Klatt, D. H. & Klatt, L. C. (1990), "Analysis, synthesis, and perception of voice quality among female and male talkers", *Journal of the Acoustical Society of America*, vol. 87, pp. 820–857.

NEURAL MECHANISMS IN PERCEPTION OF SOUNDS WITH DIFFERENT TIME PARAMETERS

E.A.Radionova

I.M.Sechenov Institute of Evolutionary Physiology and Biochemistry and

I.P.Pavlov Institute of Physiology, St.Petersburg, Russia

ABSTRACT

Two functional systems of neurons with opposite time characteristics are distinguished in the auditory system: a system of neurons responding to sound transients and a system of neurons which reflect in their activity stationary parameters of sound signals. It is supposed that complex sounds (including speech sounds) are processed in relation to their time parameters simultaneously in two ways which are characteristic of the above two neuronal systems.

INTRODUCTION

While studying responses of the auditory system neurons to sound signals with different parameters two functional systems of neurons were distinguished which differed sharply in a number of their characteristics - both in responses to stationary signals and in responses to signals with pronounced transients [1, 2]. The difference between the two systems concerns such properties of neurons as their response pattern, temporal summation, response dependence on the sound initial phase and rise time of its amplitude, sensitivity to phase spectrum of stationary sounds, threshold and latency characteristics. The present work is devoted to description of these two neuronal systems and possible role of each of these systems in the process of the auditory analysis of acoustical signals with different time characteristics. (A system of neurons with intermediate properties is not considered here in view of a limited space of this paper).

The present description concerns mainly the lower levels of the auditory system, namely cochlear nuclei at the medullar level of the brain and inferior colliculus at the midbrain level. Just at these levels of the auditory system the properties of the above neuronal systems are most pronounced.

It is supposed that complex sounds (speech sounds included) can be processed in relation to their time parameters simultaneously in different ways which are characteristic of the above two neuronal systems. The role of the higher levels of the auditory system (medial geniculate body and the auditory cortex) in this process is considered in DISCUSSION.

MATERIAL AND METHODS

Recording technique

Responses of neurons from the cochlear nucleus and inferior colliculus to sound signals were investigated in electrophysiological experiments on anaesthetized cats with the help of isolated tungsten microelectrodes (the electrode tip of about 0.001 mm, 1-5 megohm resistance as measured at 1 kHz). Impulse activity of single neurons as well as summed neuronal activity was recorded. The latter was registered as on-, off- evoked potentials (EPs) and as sustained activity - either synchronized with the sound signal (the so called frequency following response) or non-synchronized activity (summed asynchronous response).

Stimuli

Experiments were performed in a screened sound-attenuated chamber. Sound signals were presented through the electrostatic earphone with ±4 dB frequency dependent characteristic in the range 0.07-30 kHz, under the closed field conditions (all measurements with the Bruel and Kjaer 135 type microphone). Sound signals were tone or noise bursts, as well as complex signals of 2-6 harmonic components. Only linear range of sound intensities was used. The following time characteristics of sound signals varied: 1/ sound duration (within 1-200 ms); 2/ rise time of the signal envelope (in the range 0-40 ms); 3/ initial phase (from 0 to 360 degrees, initial and the end phases being of the same value); 4/ the wave form of complex signals - at the cost of changing either frequency or phase of its harmonic components. The rate of stimulus presentation amounted to 0.3-1/s.

Data processing

Neuronal responses were accumulated over 5-20 stimulus presentations: impulse activity was recorded with the "dot" method or as poststimulus histograms, single realizations of summed activity were superimposed. The following response characteristics were estimated: response pattern and the response value - as the number of impulses (for single neurons) or as the summed response amplitude; response duration, latency, and threshold values, as well as the wave form of summed evoked potentials and of summed synchronized activity. All these characteristics were evaluated depending on the sound signal time characteristics enumerated above.

RESULTS

General characteristic of two extreme neuronal systems

Initially, two types of neuronal systems with opposite properties were distinguished basing on the value of the neuron latency augmentation following intensity decrease of the best frequency tone down to its threshold value (long- and short-latency response types). Further it was found that this property was connected with other special features of neuron response to sound stimulation, especially with those connected with signal transitory and stationary parameters. All these features in common determine the neuron response type. It should be noted that the long-latency response type determined at the best frequency tone stimulation can change for the short-latency response type at other frequencies.

At the cochlear nucleus level (which is the first central level of the auditory system receiving all projections from the ipsilateral auditory nerve) about 90% of neurons respond with a tonic discharge pattern following stimulation with a tone burst of the characteristic (best) frequency (i.e. the frequency of the lowest response threshold for a given neuron). With the sound signal duration of about 100 ms and intensity high enough (approximately 40 dB or higher above the neurons' response threshold) the neuron discharge pattern usually consists of about 10-20 impulses covering the whole time of stimulus duration. The latency of such a response pattern amounts to about 2 ms. Decrease in stimulus intensity down to its threshold value results in a great latency augmentation - up to several dozens of milliseconds. Just this phenomenon is characteristic for neuronal responses of the so called long-latency type. Of interest is also the fact of great latency fluctuation (within dozens of milliseconds) at the response threshold.

At the higher, inferior colliculus level, percent of neurons of this type is significantly lower and amounts to about 35% of all neurons investigated. Naturally, quantitative values of their functional characteristics differ essentially from those established for the cochlear nucleus neurons: number of impulses in discharge is lower (usually about 5-10 impulses), the shortest latency value is higher (most often about 5-7 ms), and the latency at the reaction threshold is generally lower (within 20 ms in half of cases) than at the cochlear nucleus level.

Unlike the long-latency neurons, neurons of the other, short-latency type usually respond to the best frequency tones with phasic on-responses of 1-2 impulses or of a short burst of impulses. Their latencies (which are usually short at high intensities of stimulation) remain nearly unchanged at threshold intensities. Neurons of this type are rather rare at the cochlear nucleus level but amount to about 50% at the level of the inferior colliculus.

Neuronal responses to sound signals of different durations

As it was mentioned above, neurons of the long-latency type usually respond to sound signals of long enough duration (of the order of dozens of milliseconds) with a tonic discharge pattern, duration of the latter corresponding to stimulus duration. In result, the sound signal wave form can be reflected in the summed sustained activity of this type neurons. With diminution of the signal duration down to 1-2 ms, the neuronal response becomes shorter and reduced to 1-2 impulses. In case the sound signal intensity remains constant the latency of

the neuron response does not depend on stimulus duration and remains constant at any duration of the sound signal. However response threshold and the range of its fluctuation depend greatly on stimulus duration. At the level of the cochlear nucleus, diminution of signal duration from 100 to 1 ms results in response threshold augmentation for dozens of decibels; the range of its fluctuation augments for several decibels. Neurons of the inferior colliculus level also show augmentation of the response threshold following decrease of stimulus duration, though to somewhat lesser extent than cochlear nucleus neurons; meanwhile the range of threshold fluctuation augments to a greater extent than at the cochlear nucleus level.

As to the short-latency type neurons, characteristics of their responses practically do not depend on the sound signal duration. These neurons, usually responding to sounds with a phasic on- or on-off discharges of 1-2 impulses (or of a short burst of impulses) show (at any stimulus duration) short latencies of minimal fluctuations and nearly constant thresholds of relatively high values (higher for about 20 dB as compared with the long-latency type neurons, at an average).

Thus, only long-latency type neurons can reflect in their activity stimulus duration, its wave form, and their changes.

Neuronal responses to sound signals with different rise time of their envelope

Augmentation of the rise time of the signal envelope evokes certain changes in the initial part of discharge of the long-latency type neurons, which results in desynchronization of their activity. Neuronal response value, its duration are decreased, and the latency value increases - due to intensity decrease over the sloping part of the signal envelope. Meanwhile the response threshold does not show any changes.

Neurons of the short-latency type, on the contrary, are characterized by practically unchanged discharge pattern (on-, off-responses) and latency values. However their responses to sounds with sloping envelope are retained only within a very narrow range of the stimulus rise time values (below 10 ms at the cochlear nucleus level), since the response threshold rises greatly. Besides, under conditions of stimulation with signals with sloping envelope, neurons of the short-latency type show a significantly lower noise tolerance than the long-latency type neurons. It should be noted that the higher is the level of the auditory system the less sensitive are neuronal responses to augmentation of the signal front rise time (as it is shown in the works with evoked potentials, EPs, which reflect summed initial responses of neurons). This fact evidences that at higher levels of the auditory system, neurons' tolerance of the desynchronizing effect of increasing rise time of the signal envelope rises.

In general, neurons of the long-latency type prove essentially more tolerant of the sound signal front sloping than neurons of the short-latency type.

Neuronal responses to initial phase changes in tonal signals

Changes of the initial phase of tonal signals are essential for responses of the short-latency type neurons and practically produce no effect on the response characteristics of the long-latency type neurons. It is important that the phase effect observed on the short-latency neurons with on- or on-off responses depends on the sound frequency relation to the neuron characteristic frequency [3]. In case the sound signal frequency (F) is lower than the neuron characteristic frequency (CF), maximal response value (i.e. maximal number of impulses in the neuron's discharge or maximal amplitude of the EP) is observable at the signal initial phase of 90 and 270 degrees. At the initial phase of 0 and 180 degrees the response value is minimal. Therefore the function relating the number of impulses to the initial phase value (from 0 to 360 deg.) has an M-like shape. However in case F>CF, maximal response values correspond to the initial phase values of 0 and 180 degrees, whereas at the initial phase of 90 and 270 deg. the neuron response value is minimal. In result the function relating the response value to the initial phase has a W-like shape. The above phase effects are pronounced the more, the greater is the difference between the CF and F. When F=CF the phase effect is minimal or absent.

Thus, only short-latency neurons show the possibility to differentiate the character of the transitory process at the moments of the signal on- and off-sets, determined by the signal initial phase value.

Neuronal responses to complex sounds of different wave form

The wave form of complex signals finds its reflection in poststimulus histograms of the long-latency neurons' activity and in the summed synchronized activity (the so called frequency following response, FFR) of neurons from the lower levels of the auditory system. At the cochlear nucleus level the wave form of acoustical oscillations is reproduced in the FFR with but slight distortions: nearly the whole oral speech is reproduced in the FFR (and can be listened to with the help of appropriate apparatus [4]). At the inferior colliculus level these distortions are more pronounced. Meanwhile a phase change for 15-30 deg. in one of the complex signal components produces a pronounced change in the wave form of the FFR recorded even at the inferior colliculus level.

Of interest is that phase change of the second (higher) harmonic of a two-tone signal results in different phase effects depending on the characteristic frequency of the long-latency type neurons [5]: functions relating response value to the phase change (from 0 to 360 deg.) show one maximum in low-frequency neurons (CF<5 kHz) but two maxima in high-frequency neurons (CF>11 kHz). Neurons with intermediate CFs demonstrate the functions of intermediate form [5-7].

It may be concluded that only long-latency type neurons can specifically respond to complex sounds and to changes of their wave form.

DISCUSSION

Thus, stationary and transient parameters of sound signals can be processed in parallel in two neuronal channels: a system of long-latency neurons (I) can perform analysis of the sound signal stationary parameters - on the basis of the neurons' ability for temporal summation, accumulating in time acoustical information, slight dependence of the response characteristics on the signal front slope, and neuronal ability to reflect the sound wave form in impulse and summed activity. A system of short-latency neurons (II) can perform analysis of signal transients - on the basis of their response pattern, short latency of practically no fluctuations at any parameters of sound stimulation, and pronounced dependence of response on the signal rise time. At higher levels of the auditory system the role of system I decreases and the role of system II augments. Besides, a system of neurons with intermediate properties manifests itself more and more, especially at the MGB and cortical levels. This may result from high convergence of impulsation from neurons of the lower systems I and II. Thus initial channelizing the responses (at the lower levels of the auditory system) according to stationary and transient properties of sound signals is changed for integration of these responses which seems necessary to form an integrated auditory image in the brain.

REFERENCES

[1] Gersuni, G. V., ed. (1971), *Sensory processes at the neuronal and behavioral levels*, New York: Academic Press, pp. 135 -180.

[2] Radionova, E. A. (1971), *Functional characteristics of the cochlear nuclei neurons and audition.* Leningrad: Nauka (in Russian, English summary).

[3] Radionova, E. (1988), "Off-responses of the auditory system", *Hearing research*, vol. 35, pp. 229-236.

[4] Radionova, E. (1987), "Speech sounds in frequency following responses", Proc. XI ICPhS, Tallinn, vol. 1, pp. 255-257.

[5] Radionova, E. (1990), "Different phase sensitivity of low- and high-frequency neurons", Hearing Research, vol. 48, pp. 221-230.

[6] Radionova, E. (1987), *Sound signal analysis in the auditory system*, Leningrad: Nauka (in Russian).

[7] Radionova, E. (1987), "Monaural phase sensitivity in neurons", *Auditory pathway. Structure and function*, New York: Plenum, pp. 213-216.

THE TEMPORARY ENERGY DISTRIBUTION MODEL (TED) OF PITCH PERCEPTION

Henning Reetz (henning.reetz@uni-konstanz.de)
Dept. of Linguistics, University of Konstanz, Germany

ABSTRACT

Pitch perception models assume that either the *place* along the basilar membrane or the firing *rate* of neurons encodes the perceived pitch. Place coding is widely accepted, because random sine phase components show no periodic peak pattern in the complex waveform [1], and two sine tones presented at different ears can cause a pitch percept [2], neither of which a peripheral peak-picking model can explain. Contrary to this argues this paper for a *rate* coding of the energy of the acoustic signal in the temporal domain as the source for pitch perception, resolving the two aforementioned problems. The model is implemented in a pitch extraction algorithm.

THEORY

Sound waves reaching the outer ear are converted by the middle ear mechanics into travelling waves on the basilar membrane in the inner ear, eventually leading to neural firing of the haircells in the membrane. The *place* of maximal elongation of the membrane, and correspondingly, the maximal firing of neurons, is a frequency-dependent gradient along the membrane. This encodes different frequencies in different neuron locations along the membrane. In some way, the basilar membrane acts as a mechanical power-spectrum analyzer and the brain is supposedly able to derive the pitch of the signal from the spectral representation of it.

At the same time, individual cells fire in synchrony with the maxima of the acoustic signal, and the firing *rate*, or more precise, the distances between neural peaks encode the periodicity of the signal in the time domain. Furthermore, not only the neurons fire at the place of maximal elongation of the membrane, but all other neurons as well, although with reduced rate. And because temporal information is available in higher regions of the brain with a precision of a few microseconds, the brain could derive the pitch from the temporal representation of the signal.

In Goldstein's optimum processor theory [3], either the *place* or *rate* of the neural representation can be the input of the central pitch-processor, removing the ground for the spectral representation claimed by [2] as the only possibility to explain their experimental findings. Thus, the strongest argument for a *place* representation of the pitch in the auditory nerve is based on the experiments by [1]. He used complex signals differing only in the phase relations of their sine wave components. These signals have the same power-spectrum but differ considerably in their waveforms. The signals were perceived with the same pitch, hence he argues that only the phase insensitive spectral representation can explain the pitch perception, because a temporal representation of pitch would change with the differing waveforms. The argument is based on a waveform representation of the signal at the inner ear.

Unfortunately, the middle and inner ear are not linear systems and the basilar membrane performs a complicated three-dimensional movement with different travelling times along its length for different frequency components. The system uses active feedback components and cannot be described in linear terms. The movement of the hair-cells is a sine-wave movement for a sine-wave signal, but for complex signals, especially for time-varying signals like the speech waveform, this analogy breaks down. The ear has to be considered as an energy transformer and is not an amplitude encoder [4]. This essential point is missed by the argument that pitch perception is phase-insensitive and therefore cannot be explained in the time domain [1]. In fact, the loss of phase information in the spectral domain is not a consequence of the Fourier Transformation, but the outcome of computing the power-spectrum from it.

While the movement of the basilar membrane has not been mathematically modelled yet, we know that the membrane and its hair-cells convert signal energy into neural firing synchronized with signal maxima (for sine waves). More generally, the firing is in synchrony with maxima of the signal's energy, with firing rates of the neurons differing in intensity along the basilar membrane. Therefore, a rate coding of energy in the temporal domain is likely to exist in the auditory nerve.

THE TED MODEL

In contrast to the 'neural firing in synchrony to signal amplitude' model, I propose a 'neural firing in synchrony to signal energy and frequency' model, where the distribution of energy in the temporal domain (Temporal Energy Distribution) encodes the pitch information. I describe now this model in the subsequent text in more detail. Numbers in the text refer to Figure 1.

The central idea of the model is the parallel representation of the acoustic signal in energy bands with different frequency responses. Taking a small window from a signal and computing its energy represents high-frequency energy, while wide windows represent low frequency energy (1). A range of windows with increasing sizes represents the energy in bands with decreasing frequencies (2). These energy bands can be understood as an instantaneous energy spectrum which is different from the classical power-spectrum. In the classical power-spectrum, the frequency distribution in a window is given under the assumption that the signal part in the window is a stationary signal. The classical power-spectrum is also usually used as data-reduction step, namely for locating harmonics in it. In opposition to this, the computation of the energy bands is made without an assumption of the type of signal, and it yields an increase in the data rate.

Next, the energy bands are converted into a representation about the maxima in it (3). In each energy band, the signal maximum leads to a 'firing' of a neuron according to the 'all-or-nothing' principe, i.e., information about the absolute value of the maximum is lost and only the information that the signal has reached a maximum at a certain time is encoded in the neural firing. The 'ignition' of the cell is linked to adaptation and refraction processes, preventing the neurons from firing at every local maximum, independent of their size and duration in relation to the neighboring signal.

This parallel concerto of firings is gathered into a temporal histogram of all neurons (4). This histogram is the energy distribution in all energy bands over time. The distances between the maxima in it reflect the periodicity in the signal.

SOME CONSEQUENCES

The TED histogram can be interpreted in terms of speech production and perception. In speech production, energy is emitted either permanently (e.g., in voiceless fricatives) or impulsively, where the impulse can be a singularity (e.g., in a plosion burst) or impulses can occur repetitively (e.g., in a voiced sound). The TED histogram shows the impulsive energy emission, which can be a singularity, or a repetitive but irregular emission (e.g., in a creaky voice), or it can be a quasi-periodic sound. The difference between these three groups of sounds is reflected in the distribution of energy as being either singular, not periodic, or quasi-periodic. Especially the capability to identify any voiced sound, may it be periodic or not, gives the TED representation more power than most other pitch detection methods.

In perceptive terms, the TED model locates any energy distribution in the signal, independent of its origin. Furthermore, the TED representation has

some unusual feature for a temporal representation: random noise leads to a more or less random firing of all neurons, resulting in a nearly flat TED histogram. Periodic signals yield periodic firing in several energy bands, resulting in periodicity in the TED histogram. Spectral phase relations, number of involved harmonics, and random noise only decrease the 'peak-to-noise' ratio in the TED histogram, but does not hinder the pitch detection. As illustration, Figure 1 displays the behaviour of the model with a sine signal covered by random noise with an S/N ratio of -12 dB.

ALGORITHM

The TED model was implemented in an algorithm whose general operation is described now with regard to Figure 1. (1) The speech signal is windowed with Hamming window sizes between 1 and 15 ms, converting the signal into parallel bands; the windows move sample-by-sample over the signal. (2) The windowed samples are squared and added up in the individual bands. (3) The local maxima within ±1 ms are selected and are represented as peaks with unitary height within each band. (4) The peaks of all bands are combined into a TED histogram. (5) Peaks with irregular distances to neighboring peaks and peaks with low amplitude are eliminated from the histogram. (6) The distances between peaks are represented as a pitch value if (i) they form a sequence of at least four peaks, and (ii) this sequence is longer than 30 ms.

The algorithm has a very simple structure but is slow on a digital general-purpose computer. Its *on-line* behaviour and its regular structure with simple computations and decisions makes it suited for realization in silicon where it could operate in real-time.

CONCLUSION

An algorithm has been presented whose design is based on principles derived from the auditory processing in the inner ear. (These principles have been further tested with perception experiments presented elsewhere [5]). The speech signal is represented by a temporal structure in parallel energy bands which are computed in the temporal domain. This representation reflects speech production and perception issues equally well. In ongoing research I investigate the possibility to eliminate the periodicity test (step 5 of the algorithm) by incorporating more details of the intensity adaption of the inner ear into the model. Tentatively, the temporal energy distribution might also be suitable for the segmental representation of speech.

REFERENCES

[1] Wightman, F.L. (1973) "Pitch and stimulus fine structure", *JASA, 54:* 397-406.

[2] Houtsma, A.J.M. and J.L. Goldstein (1972) "The central origin of the pitch of complex tones: evidence from musical interval recognition", *JASA, 51:* 520-529.

[3] Goldstein, J.L. (1973) "An optimum processor theory for the central formation of the pitch of complex tones", *JASA, 54:* 1496-1516.

[4] Duifhuis, H. (1992) "Cochlear modelling and physiology", In: *The auditory processing of speech - from sounds to words,* M.E.H. Schouten, (Ed.) Mouton: Berlin. p. 15-27.

[5] Reetz, H. (1995), *A temporal pitch perception model*, Doctoral diss. (unpublished)

[6] Houtsma, A.J.M., T.D. Rossing, and W.M. Wagenaars (1987) *Auditory demonstrations (CD)*, Institute for Perception Research (IPO) and Northern Illinois University (NIU), supported by the Acoustical Society of America: Eindhoven, The Netherlands.

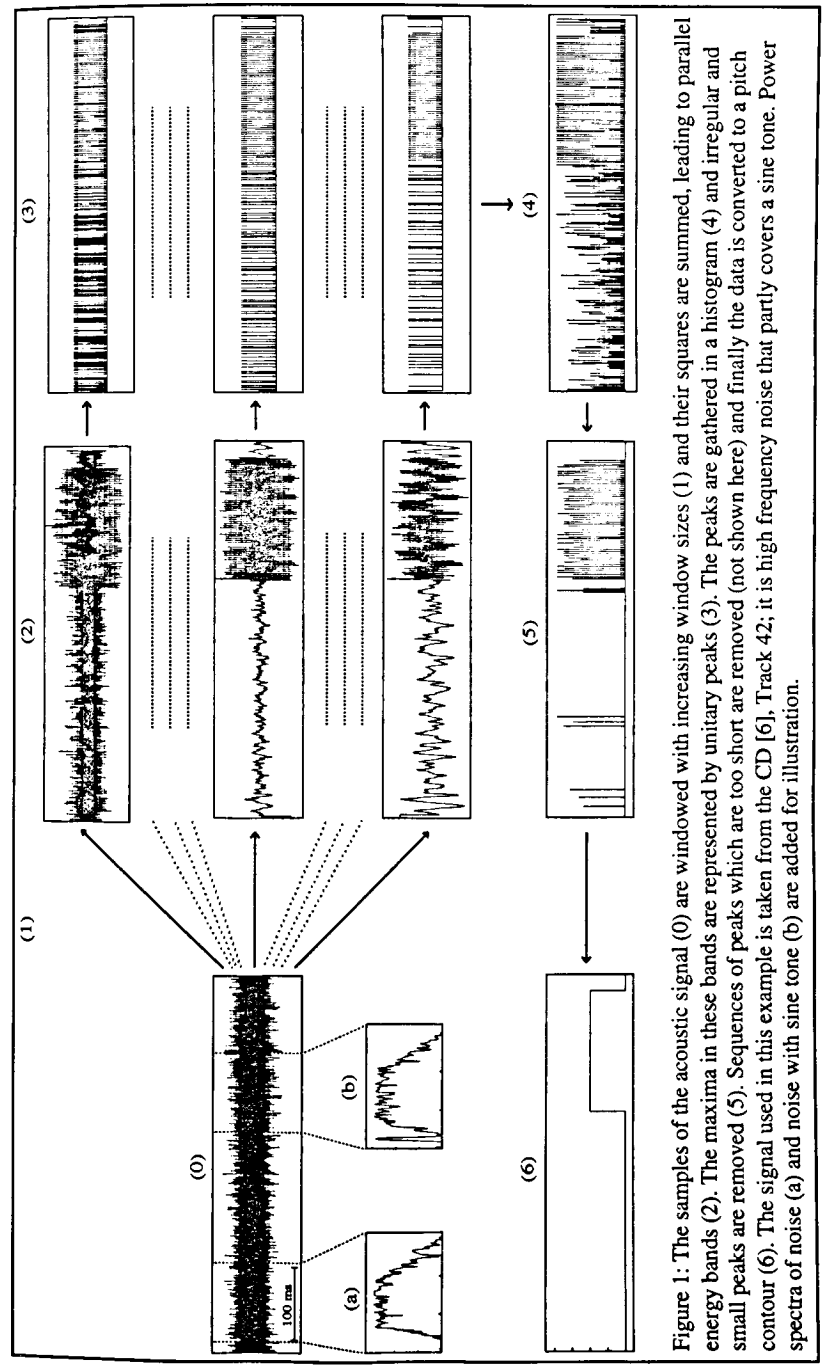


Figure 1: The samples of the acoustic signal (0) are windowed with increasing window sizes (1) and their squares are summed, leading to parallel energy bands (2). The maxima in these bands are represented by unitary peaks (3). The peaks are gathered in a histogram (4) and irregular and small peaks are removed (5). Sequences of peaks which are too short are removed (not shown here) and finally the data is converted to a pitch contour (6). The signal used in this example is taken from the CD [6], Track 42; it is high frequency noise that partly covers a sine tone. Power spectra of noise (a) and noise with sine tone (b) are added for illustration.

WORD INTELLIGIBILITY AND PLACE ASSIMILATION IN SPONTANEOUS SPEECH

C. Sotillo[1], J. McAllister[1], E. G. Bard[1], G. Doherty-Sneddon[2], and A. Newlands[2]

Human Communication Research Centre,
[1]Dept. of Linguistics, Edinburgh University, [2]Dept. of Psychology, Glasgow University

ABSTRACT

Tokens of words involving word final place of articulation assimilation are less intelligible than canonical forms when excerpted from context and presented to a group of listeners. While the process of assimilation is able to explain certain differences in intelligibility, it is clear that there are additional factors, such as the frequency of productive morphological suffixes, which influence the ease with which a word is recognised.

INTRODUCTION

One way in which tokens of the same spoken word vary is the extent to which their acoustic form supports recognition of the word in the absence of the word's natural context. The ease with which subjects are able to recognise an excised word can be taken as a measure of that token's intelligibility. This definition of intelligibility thus represents the bottom-up processes involved in lexical access, reflecting the amount of information that is available in the acoustic signal. The poorer the quality of input, the lower the intelligibility.

Sources of information other than the auditory input also aid a listener in recognising the incoming message. For example, the syntactic and semantic context of an utterance will restrict the set of appropriate lexical choices [1]. Such top-down processes have been shown to have an effect on the duration and intelligibility of word tokens. Speakers adjust their pronunciation of words in running speech to complement the information available to listeners in the remainder of the discourse. So, for example, the more predictable a word is from its sentential context, the less clear the token will be [2].

Having established that such reductions in intelligibility occur, the question arises as to how such differences in clarity might be realised. Well-known phonological processes such as place assimilation are prevalent in running speech [3], and it seems reasonable to assume that the application of these processes may have an effect on intelligibility.

Opinions vary on how assimilation should be modelled in current theories of phonetics/phonology. It is not entirely clear, for example, whether assimilation should be accounted for at the phonetic or phonological level of representation. What cannot be disputed is that the sandhi phenomena of connected speech have implications for theories of lexical access.

This paper addresses the relationship between assimilation and the ability to recognise words, as reflected in measures of intelligibility. The question is whether regular and predictable variation, like that involved in place assimilation, imposes an additional *cost*, or leaves the process of word recognition unaffected.

Cross-modal priming experiments using isolated words [4, 5] show that even very small deviations from the canonical form of a word (i.e. changes of just one phonological feature) will result in a loss of priming. From this, one would predict that assimilatory processes would have an inhibitory rather than facilitatory effect on word recognition. An assimilated token would be harder to recognise when excerpted.

However, recent work on repetition priming of words *within sentences* using place of articulation assimilation found no loss of priming for assimilated words, except when followed by an unviable context [6]. This result could be accounted for by a conservative lexical matching process which, though intolerant of mismatch, only rejects candidates when there is unambiguous mismatching information. In this case, an assimilated token excerpted from context is no more likely to be rejected than an unambiguous token.

We report a study exploring the relation between intelligibility and place of articulation assimilation and consider whether assimilation necessarily involves an increase in processing load. We discuss the implications of the result for theories of lexical access.

METHOD

Materials

Data was selected from the HCRC Map Task Corpus [7]. The 128 unscripted conversations involve pairs of participants who collaborate to replicate on one's schematic map a route drawn on the other's. Task success requires discussion of the various landmarks along the route, the names of which were carefully chosen to provide the appropriate environment for certain phonological reduction processes in English. In particular, certain names involved possible place assimilation of word-final alveolar nasal stop consonants. Both labial (e.g. *caravan park*) and velar (e.g. *Indian country*) contexts were used. After completing their set of map tasks all speakers were required to read carefully a list of landmark names to provide clear 'citation' forms against which other tokens could be compared. Materials were recorded on DAT (Sony DTC1000ES) using one Shure SM10A close-talking microphone and one DAT channel per participant.

Single word tokens from fluent first and second mentions of landmark names were then used in a series of intelligibility experiments to explore the effects on intelligibility of information status within dialogue [8]. This set of studies established an intelligibility score (in terms of percent correct recognition) for each word token when it was excerpted from context and presented in noise to a panel of listeners from the same language community as the Corpus participants.

We selected from this pool those words relating to landmark names involving possible place assimilation of word-final nasals. There were 21 usable examples of nasals preceding labial stops (e.g. *telephone* box), and 13 usable examples of nasals preceding velar stops (e.g. *lemon* grove). For each of these landmarks there were two running speech tokens: the first, introductory, mention and the second mention. In some cases the second mention was by the same speaker who introduced the word, in other cases the second mention was by a different speaker. Each running speech token had a corresponding citation form against which it could be compared. This gave us a total of 68 running speech/citation form pairs.

Procedure

Each utterance containing a required token was digitised at a rate of 16KHz using the Entropic Signal Processing System through the XWAVES speech analysis program on a Sparc station. The start and end points of each word were located using a combination of audio and visual information provided by time-amplitude waveforms and broad band spectrograms. Cuts were made at zero-crossings.

Excerpted tokens were used for two kinds of studies. For the experiments on intelligibility tokens were overlaid with noise by multiplying, sample by sample, the original speech file by a 16KHz file of random noise (where all sample values were in the range 0.5 to 1.5). For each resulting stimulus the amplitude was related to that of the original speech data file, and the data points had the same sign as the original data values they replaced. The tokens presented to subjects in the perceptual task were not masked by noise.

Intelligibility Task

In each experiment, either 48 or 60 word types were used, with four or three tokens per type depending on the experimental design. Tokens were allocated to different presentation tapes according to Latin square designs and played to groups of listeners. Only one token of any word type was presented to each subject, and each token was heard by at least 9 subjects. Subjects were asked to write down the identity of each word they heard. For the subset of word types used in the perceptual task, mean scores for correct recognition were calculated for all tokens which appeared in more than one intelligibility experiment.

Perceptual Task

The unscripted nature of our running speech material, the use of non-linguistically trained subjects, and the ratio of tokens to speakers rendered a detailed acoustic analysis (for example in terms of pole/zero decomposition [9]) impossible.

We opted therefore to explore the perceptual evidence for assimilation by presenting tokens to a group of nine phoneticians who were asked to make a set of judgements about the place of articulation of each word-final nasal

consonant. Experts rated each nasal on three scales: labial, alveolar, and velar. A rating of 0 indicated that no evidence was perceived to suggest the consonant was produced at this place, while a rating of 5 indicated that the perceptual evidence was fully consistent with an articulation at this place. The three options were not mutually exclusive, so that in principle it was possible to assign a rating of 5 on more than one scale for any one token.

RESULTS

Scores for correct recognition for this subset of intelligibility data were submitted to an ANOVA by materials (by subjects analysis was not possible since the items were gathered from a series of different experiments). Raw intelligibility scores for citation forms and spontaneous mentions can be seen in Table 1.

Citation forms are significantly more intelligible than their corresponding running speech tokens [*Form:* F_2 (1,31) = 33.74, $p < 0.0001$].

In addition, an ANOVA run on *loss of intelligibility*, that is, the difference in rate of correct identifications between citation forms and running speech tokens, revealed a significant effect of mention, with greater loss of intelligibility for second mentions [*Mention:* F_2 (1,31) = 7.27, $p = 0.01$].

Thus the repetition effects on intelligibility reported elsewhere [10, 8], hold for this subset of data.

Table 1. Intelligibility of citation and running speech tokens for introductory and repeated mentions

Form	Mention	
	First	Second
Citation	.70	.76
Running speech	.48	.41

When the experts' overall judgements of assimilation were examined, just over one third of all responses indicated no assimilation had taken place (35.2%); nearly one fifth of all tokens involved a clear assimilation (17.85%), while the remainder involved percepts of an alveolar with a a varying degree of labial or velar quality.

ANOVAs on experts' mean place judgements showed strong Form effects with citation forms being judged as significantly more [n]-like and less [m]- or [ŋ]-like than corresponding running speech tokens [*[n]:* $F(1,32) = 20.09$, $p < 0.0001$; *[ŋ]:* $F(1,32) = 8.79$, $p < 0.01$; *[m]:* $F(1,32) = 3.62$, $p = 0.066$].

The difference in [ŋ]-ness judgement between running speech tokens and citation forms was greater for second mentions than for first mentions, with second mentions being perceived as more assimilated. [*Form X Mention:* $F(1,32) = 4.22$, $p < 0.05$]. This was true regardless of following context, though [ŋ]-ness judgements preceding labials were significantly lower than those preceding velars [*Place:* $F(1,32) = 15.14$, $p < 0.0005$].

No effect of mention was found for [m]- or [n]-ness judgements of tokens preceding either labials or velars. [*Form X Mention for [n]:* $F(1,32) < 1$, n.s.; *for [m]:* $F(1,32) = 1.33$, n.s.].

It appears that we have evidence to suggest assimilation was indeed taking place, and we also have an intelligibility effect to explain. What, then, is the relation between the two?

Assimilation and intelligibility

A series of correlations showed that although judged assimilation was related to intelligibility it did not account for all of the intelligibility differences in the data.

Significant correlations between intelligibility and place judgements were found only for words preceding velar stops: the more [ŋ]-like (i.e., assimilated) were less intelligible [$r = -.409$, $p < 0.005$], the more [n]-like (un-assimilated) more intelligible [$r = .491$, $p < 0.001$]. For words preceding labial stops, however, analogous correlations were not significant [$r = -.105$, n.s., and $r = .184$, n.s.].

In addition, non-assimilatory non-target pronunciation ([m]-like character in a velar context) was also found to correspond with decreased clarity [$r = -.361$, $p < 0.009$] for words preceding velar stop consonants.

Intelligibility subjects' responses

In an attempt to account for the lack of correlation between intelligibility and judgements of assimilation for words preceding labials, we analysed the alternative responses of the original subjects in the intelligibility studies.

The alternative words offered in cases of incorrect recognition were classed according to their word-final segment, and these subjects' responses were compared with the responses of the experts.

Words judged by experts as [m]-like elicited more incorrect identifications ending in [m]. This was true both of assimilated tokens preceding labials [$r = .212$, $p = 0.05$] and of tokens preceding velars which were judged by experts as sounding (inappropriately) [m]-like [$r = .313$, $p = 0.02$].

Words preceding velars and judged by experts to have assimilated towards [ŋ] correlated with subjects' incorrect identifications ending in [ŋ] [$r = .418$, $p = 0.002$]. However, words preceding labials and judged by experts as sounding inappropriately [ŋ]-like showed no relation to subjects' responses [$r = -.076$, n.s.]. A closer examination of this set of data revealed that subjects were offering words ending in [ŋ] regardless of the experts' judgements.

We suggest that this result can be explained by the structure of the lexicon in English. The productive -ING affix leads to subjects responding with lots of [ŋ] ending words, whether or not there is auditory evidence for velar articulation.

CONCLUSIONS

The general conclusion is that there is a relation between intelligibility and assimilation: tokens of a word which are perceived to have been assimilated result in poorer recognition when excerpted from context. We infer from this, that there is indeed a cost involved in the processing of assimilated tokens. It is necessary to exert effort in recognising the context in which an assimilation occurs in order for it to be successfully recognised as an appropriate change. Without supporting context, an assimilated token is harder to recognise than its canonical counterpart. These results are in line with experiments on cross-modal priming of isolated words, where a single feature mismatch reduces the priming effect.

We must also conclude that the relation between intelligibility and assimilation is complex. Firstly, the effect of assimilation on intelligibility varies according to the place of articulation of the assimilatory environment (e.g. labial or velar). Secondly, assimilation appears to be one of several factors which make tokens harder to recognise. The failure of perceived assimilation to account for the repetition effect on intelligibility indicates that there are other factors at play. We argue that these factors include not just the phonetic and phonological, but also the lexical. The structure of the lexicon, and the frequency of occurrence of particular morphological structures need also to be considered in any full account of what makes words easy or difficult to recognise.

This work was supported by the ESRC(UK) via the HCRC. *Dr. McAllister is currently at the Department of Psychology, University of Auckland, New Zealand. This work was conducted while Dr. McAllister was in Edinburgh as a Visiting Reseach Fellow. Dr. Doherty-Sneddon is now at the Department of Psychology, University of Stirling. Address for correspondence: C. Sotillo, HCRC, University of Edinburgh, 2 Buccleuch Place, Edinburgh EH8 9LW, UK*

REFERENCES

[1] Bard, E.G., Shillcock, R.C. and Altmann, G.T.M. (1988), "The recognition of words after their acoustic offsets in spontaneous speech: effects of subsequent context", *Perception and Psychophysics*, 44, 395-408.

[2] Lieberman, P. (1963), "Some effects of semantic and grammatical context on the production and perception of speech", *Language and Speech*, 6, 172-5.

[3] Dalby, J (1984), *Phonetic structure of fast speech in American English*, Unpublished PhD thesis, University of Indiana.

[4] Marslen-Wilson, W.D. and Zwitserlood, P. (1989), "Accessing spoken words: On the importance of word onsets", *JEP:HPP*, 15, 576-585.

[5] Marslen-Wilson, W.D. and Gaskell, G. (1992), "Match and mismatch in lexical access" [abstract] *IJP*, 27, 61.

[6] Gaskell, G. and Marslen-Wilson, W.D. (in press), "Phonological variation and inference in lexical access", *JEP:HPP*.

[7] Anderson, A.H., et al. (1991), "The HCRC Map Task Corpus", *Language and Speech*, 34, 351-366.

[8] Bard, E.G., Sotillo, C., Anderson, A.H., Doherty-Sneddon, G., & Newlands, A. (forthcoming) "The control of intelligibility in running speech", *Proceedings of XIII ICPhS*, Stockholm.

[9] Yegnanarayana, B. (1981), "Speech analysis by pole-zero decomposition of short-time spectra", *Signal Processing*, 3, 5-17.

[10] Fowler, C.A. & Housum, J. (1987), "Talkers' signaling of 'new' and 'old' words in speech and listeners' perception and use of the distinction", *JML*, 26, 489-504.

SOME EFFECTS OF EXTRA- AND PARALINGUISTIC VARIATION ON THE PHONETIC QUALITY OF VOWELS

Hartmut Traunmüller
Institutionen för lingvistik, Stockholms universitet

ABSTRACT

F_1 at the /i/–/e/ boundary was investigated for synthetic phonated and whispered vowels in which the higher formants and F_0 had been varied. The vowels were presented with and without a leading phrase whose overall F_1 was also varied. The results are discussed in terms of how listeners 'tune in' to a speech signal in order to recover the linguistic information. The effect of cues to this tuning was observed to vary due to interactions as well as spectral and temporal restrictions.

INTRODUCTION

Within the frame of the modulation theory of speech [1], speech perception is seen as a process in which listeners tune in to the 'carrier' (and clock rate) of a speech signal. In order to recover the phonetic quality of vowels they evaluate the deviations of the properties of the signal from those which they expect of a neutral vowel produced by the same speaker with the same vocal effort, in the same voice register, and with the same type of phonation.

In qualitative terms, this theory can explain a large number of experimental results, including the observed dependence of perceived vowel quality on various intrinsic and extrinsic variables, such as F_0 and the formants above F_2 within a vowel or its context. The theory allows for listeners to exploit any kind of cues for tuning in. The experiments to be reported are intended to show which cues actually govern listeners' expectations concerning the frequency position of F_1 when the speaker is unknown and unseen.

For this purpose, the i/e boundary value of F_1 was investigated in a set of experiments which allow conclusions about the contributions of the following variables: F_0 in the vowel itself; F_0 in the vowel and in a leading phrase; F_0 in the leading phrase alone (when the vowel is whispered); whispering vs. phonating; F_1 in the leading phrase; formants above F_1 in the vowel itself; formants above F_1 in the vowel as well as in its leading phrase; and the perceived age and sex of the speaker.

METHOD

Subjects

There were 29 listeners, 15 male and 14 female. They were speakers of standard Swedish, all but one grown up in eastern central Sweden. Nine of them were affiliated to this institute, the rest were university students who were moderately paid.

Stimuli

A female speaker of standard Swedish produced some samples of the phrase *och nu hörs nog snarast V* [ɔˈnʉːhœʂɳʊgsnɑːrast 'V:] 'and now it's rather likely to hear V' where V stands for the names of the vowels *i* and *e*, pronounced [ij^{ə}] and [e^{ɐ}]. There was always a pause of about 300 ms before the vowel. The utterances were recorded in an anechoic chamber, digitized at 16 kHz, 16 bit/sample, and subjected to LPC analysis with 17 reflection coefficients, 20 ms Hamming window, 5 ms progression. One of the leading phrases and a smoothed interpolation between [ij^{ə}] and [e^{ɐ}] were chosen as models for resynthesis with various modifications of the frequency positions of F_0, F_1, and the formants above F_1, referred to as F_h. The modifications consisted in moving F_0 and/or the F_h into positions typical of male speakers or children. For F_1 of the leading phrase, a normal and a 60 Hz (average) higher or lower position was used. Whispered vowels were synthesized using excitation by white noise, high pass filtered 2nd order Butterworth. In order to keep the number of different stimuli sufficiently small, only 3 or 4 different values of F_1 were used in each condition. In whispering, the i/e boundary was expected at a higher F_1, which is reflected in the choice of F_1 range. The combinations of modifications used are listed in Table 1. The stimuli were recorded on tape using two randomized orders, with pauses of 2.5 s between stimuli, and an additional pause after each sixth stimulus.

Procedure

Three experiments were prepared. In exp. 1, each stimulus was preceded by a phrase whose F_0 and upper formants had been subjected to the same modifications as the test vowel, while its F_1 appeared in normal and in shifted position. Phonated stimuli occurred twice, whispered stimuli once. In exp. 2, the stimuli were presented without a leading phrase, once each. In exp. 3, the different versions of the leading phrase used in exp. 1 were presented for the purpose of classification as to age (child, adolescent, adult, aged) and sex. The three experiments were run in succession, with four to seven subjects at a time.

The stimuli were presented through headphones. The subjects had to identify the vowels by marking the preprinted orthographic symbols *i* /i/, *e* /e/, *ä* /æ/, *y* /y/, *ö* /ø/ or *x* (for any other vowel they heard) on answer sheets. The wording of the leading phrase had been chosen so that all of its vowels were of type *x*, yet with the whole range of variation in F_1 represented. At the beginning of the experimental session, one example stimulus of each of the three experiments was presented for accommodation, without feedback.

RESULTS AND DISCUSSION

Figure 1 shows the identification results obtained in exp. 1 for phonated vowels with unchanged F_0 and F_h presented with an unaltered resynthesis of the original leading phrase. Since we are mainly interested in the distinction between degrees of openness, which can be assumed to be highly correlated with F_1, we are going to neglect the distinction between rounded and spread vowels. Most of the subjects did give some rounded vowel responses, but in this matter, the between-subject agreement was very low, while it was quite high for distinctions in openness or height, i.e., between /i/ and /y/ as opposed to /e/ and /ø/.

It was intended to study the F_1-values at the i/e boundary. Some of these boundaries came, however, to be located outside the range of F_1-variation used. In order to avoid the risk of substantial extrapolation errors, the values presented in the following correspond to the points where 40% of the subjects heard /i/ or /y/ and 60% /e/ and /ø/, with the few other responses neglected. This value will be referred to as 'the boundary value'.

Table 1. Combinations of leding phrases (1st column) and test vowels used. 1st digit = 0: Whispered vowel.

111	111	112	113		012	013	014	
112	111	112	113		012	013	014	
211	211	212	213		012	013	014	
212	211	222	213		012	013	014	
121	121	122	123		022	023	024	
122	121	122	123		022	023	024	
221	221	222	223		022	023	024	
222	221	222	223	224	022	023	024	025
223		222	223	224		023	024	025
322		322	323	324		023	024	025
323		322	323	324		023	024	025
232		232	233	234		033	034	035
233		232	233	234		033	034	035
332		332	333	334		033	034	035
333		332	333	334		033	034	035

Digit value	1	2	3	4	5
1st digit, F0*	0.59	1.00	1.41		
2nd digit, Fh*	0.85	1.00	1.15		
3rd digit, F1	-60	+0	+60	+120	+180

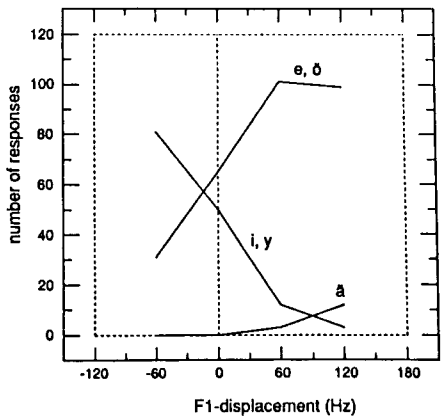


Figure 1. Identifications of phonated vowels with unchanged F_0 and F_h, presented with an unaltered resynthesis of the original leading phrase.

In Figure 2, the boundary value in the whispered stimuli is plotted against that of the otherwise identical phonated stimuli. Evidently, the boundary is situated at higher values of F_1 in the whispered stimuli. This is what could be expected, since F_1 in whispered vowels is usually higher than in the same vowels if phonated. Figure 2 also shows that this difference (in Hz) increases with F_1.

The effects of changes in F_0, F_h, and F_1 of the leading phrase are shown in Figures 3 to 5 for the cases in which the other variables remained unchanged.

The effects of F_0 and F_h were strongly nonlinear. The effect of decreasing F_0 from the original female value was, on average, considerably smaller than that of increasing it, but this can only be asserted for the phonated vowels, or for those cases in which the boundary was no higher than 2.3 units with unchanged F_0. The frequency range within which listeners expect F_1 appears to be limited so that they do not expect it to be shifted further down even if F_0 is lower. For the whispered vowels, this lower limit was not reached.

In contrast, the boundary shift effected by increasing F_h from the original female value was, in absolute terms, about 0.7 units smaller than that of decreasing it. F_h had even a negative effect for the phonated vowels. This can be understood if it is assumed that the variation in the higher formants which listeners are able to utilize for their tuning is restricted to a narrow range which, for the present subjects, did not include what can be found in children. It is fully plausible that this results in a negative effect when the limits of that range are exceeded. It may be that a different result would have been obtained with listeners who are more frequently exposed to speech of children.

It does not appear to make any difference whether the information on F_0 or F_h is contained only in the vowel itself or both in the vowel and in its leading phrase. There was no significant difference between these cases. The information contained in the vowels themselves was apparently enough to allow the subjects to adapt to the speaker, so that the addition of the leading phrase could not bring about any seizable improvement. This interpretation is supported by the observation that there was a significant and substantial effect of F_0 in the leading phrase when there

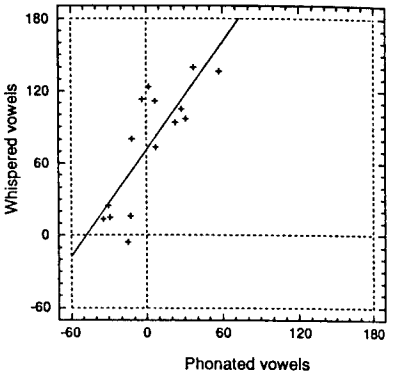
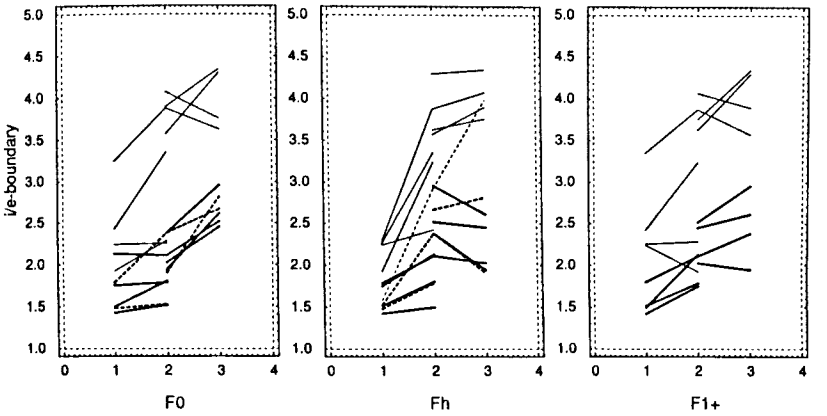


Figure 2. Boundary value (40% i, y; 60% e, ö) of F_1 in the whispered stimuli plotted against that of the otherwise identical phonated stimuli, both presented with leading phrase. Scale values indicate displacement of F_1 from the original value. Regression line also fitted ($r = 0.81$).

was no F_0 information in the vowels themselves, i.e., when they were whispered. The latter result also shows that listeners are capable of 'translating' their expectations from voiced to whispered speech, which involves an upward displacement of the i/e boundary.

The effect of F_1 in the leading phrase was, on average, not very large (slope 0.3). It was positive in 12 pairs, while it was negative in one phonated and in three whispered pairs. For voiced vowels, the effect of decreasing F_1 was, nevertheless, larger than that of decreasing F_0. In contrast with the effects of F_0 and F_h, the effects of increases in F_1 were of the same magnitude as those of decreases. The results indicate that F_1 in the leading phrase did not have a very persistent effect, otherwise the slopes in Figure 5 would have been steeper. The pause between the leading phrase and the vowel, and the long duration of the latter have apparently led the subjects to base their expectations more on the intrinsic F_0 and F_h in the vowel itself. The conclusion that listeners attach a lower weight to F_1 of the leading phrase than to F_0 and the formants above F_1 is not likely to hold in general. If the test vowel had been embedded within a phrase without pauses, we would probably have observed a considerably larger effect. This is supported by the informal observation that the phonetic quality of the vowels of the leading phrase itself was not noticeably affected by the frequency displacements in its F_1.



Figures 3 to 5. The effects on the boundary value (40% i, y; 60% e, ö) of F_1 of changes in F_0, F_h (all formants above F_1), and in F_1 of the leading phrase. Cases for which there was no other change in frequency positions connected by lines.

Line types: Thick: Phonated; Thin: Whispered vowels. Full: With a (phonated) leading phrase; Dashed: Isolated vowels.

Scale units as in Table 1.

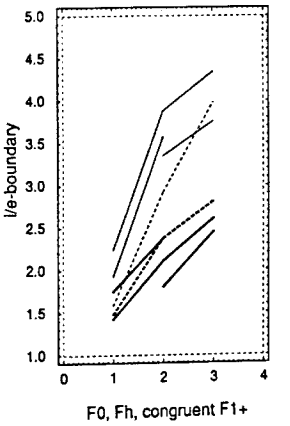


Figure 6 (to the right). Boundaries for the stimuli which simulated modal voice register in a man, a woman, and a child, with original and modified F_1 in the leading phrase. Lines drawn between pairs with congruent values of F_0, F_h, and F_1.

Figure 6 shows the shift in the boundary value for the stimuli which simulated modal voice register in a child, a woman and a man. This figure shows that, for whispered stimuli, the listeners appear to have tuned in themselves fully to the kind of speaker in question, since the slope of the lines is approximately 1.0. With the phonated stimuli, full adaptation to the speaker was apparently not reached, since the slope is smaller, about 0.6. This may have to do with the deficiency in naturalness characteristic of buzz-excited LPC speech.

It remains yet to analyze the possible effects of the listeners' perception of the speaker's age and sex.

CONCLUSIONS

The experiments have shown that all the variables analyzed so far did affect listeners' expectations concerning F_1 at the i/e-boundary, but also that the efficiency of variables is subject to various restrictions in the domains of frequency and time. They have also shown that the amount of a shift in a phoneme boundary brought about by some variable is not a valid measure of its general importance. If the listener is already tuned in to the speaker, there will be no shift at all, no matter how much additional potentially useful information is added.

REFERENCES

[1] H. Traunmüller (1994) *Phonetica 51*, 170–183.

PHYSIOLOGICAL CORRELATES OF SPEECH-HEARING INTERACTION

I.Vartanian
I.M.Sechenov Institute of Evolutionary Physiology and Biochemistry, Russian Academy of Sciences, St.Petersburg

ABSTRACT

The electrophysiological and psycho-acoustical data depicted different levels of speech-hearing interactions. The first is a "input control filter" established as independent on the higher levels of auditory processing. It supports the constancy of auditory afferent flow during vocalization as well. The next level realizes mechanisms of reinforcement of auditory flow during vocal activity of a subject. The highest levels were described as reciprocally related with subcortical levels of auditory afferent flow processing.

INTRODUCTION

It is evident that the activity of auditory and speech producing systems of humans is interconnected and interdependent. Nevertheless it can be stated that up to the present time we have only scanty information on the structure and interaction mechanisms of the sensory and motor components of the auditory-speech function. The question about the influence of speech process as a directed motor action provided by the program of the central brain conducting links remains unsolved, though the interaction between hearing and speech presents two sides of one and the same communication process. It is obvious that the interaction of speech and hearing is due to the mutual feedback activity.

The aim of present study is to investigate the manifestations and levels of speech-hearing interaction by means of auditory potentials of different latencies evoked by test stimuli against the background of vocal speech activity of a subject.

METHODS

A.Subjects.

There were 19 subjects (males and females), aged from 19 to 50, participating in the study. All of them were healthy physically and mentally and had normal tonal audiograms both for right and left ear measured before the experiment.B.

Preliminary procedure.

All the experiments were carried out in the acoustically isolated electrically screened chamber. At the beginning of the experiment each subject was offered to sing a melody of a well known song, his mouth being closed. The intensity of a sound produced by the subject was self-selected and controlled by special device for the estimation of sound pressure with microphone, put at the distance about 10 sm. from the mouth. Usually the intensity of the song corresponded to 40 dB as compared to the intensity of wideband noise and that of the same song recorded at the magnetic tape. Both were presented to the subject through the earphones binaurally in the control series of the experiment as the control background sounds.

C. Acoustical stimulation.

The stimuli were: (1) test-stimuli clicks of 1 ms duration at the level of 40 - 50 dB above threshold; (2) sustained wide band noise of 40 dB above threshold; (3) self produced melody of a well known song; (4) the same song melody produced by the same subject

recorded in the majority of subjects only at the background of noise.

2. P_2 and N_2 components (the place of generation is supported to be the thalamo-cortical auditory structures) can be recorded only at the background of the total acoustic influence. However, during the subject's song production, P_2 and N_2 components in response to test clicks increase as compared with the response to test signals without any background and decrease while listening a song from the tape-recorder or at the background of noise (as well as P_1 and N_1 components of LLAEP and MLAEP). Peak latent periods are changeable, there being a tendency to their increase at the background of all the acoustic background influence (song melody self-production, in listening the same melody as external sounds and in noise), though this tendency doesn't reach any significant values (the results for 9 subjects).

DISCUSSION

I. Thus, initial SLAEP waves evoked by test stimulus increase while at the background of listening to a song melody from the tape-recorder or to the background noise, but do not change, if the subject is vocalizing the song himself. Wave 5 of SLAEP can increase or decrease during subject's song production, but these changes are not statistically relevant. Therefore, an incoming afferent flow changes under the influence of external sounds, but remains unchanged at the background of subject's vocalization of song melody. It means that vocalization does not hinder external sound afferentation and the afferent flow is controlled by a "filter", decreasing the influence of the sounds vocalized by the subject. One can suppose that the mechanism which allows to produce vocal speech and to listen simultaneously to external sounds without their substantial distortions at the level of the auditory input is realized at the auditory periphery due to the activity of the feedback system from the vocalization centers.

II. The fact that waves of MLAEP evoked by test stimuli increase sharply at the background of song melody production and slightly increase or even decrease during listening to a tape-recorded song melody or in the background of noise, can be considered as an indication of vocal speech producing and auditory systems interaction. The interaction must be realized at the levels of midbrain and diencephalon and is directed to the reinforcement of an external signal, to its distinguishing at the background of the subject's own vocal activity.

Such a reinforcement mechanism contributes to the significance of the income information conservation process due to vocal-hearing control. The decreasing of some MLAEP waves, in the conditions of external stimulation (tape-recorded vocalizations and noise), shows a well-known phenomenon of masking one stimulus by another depending on physical parameters of the stimuli.

III. The increase of amplitude of N_1 P_1 and P_2 N_2 waves evoked by test-stimulus at the background of song vocalization and their decrease at the background of external acoustical stimulation completely repeats the phenomena related to the electrical activity of the subcortical brain structures. Accepting the viewpoint that the LLAEP waves reflect the activity of thalamo-cortical level one can believe that subcortical processes are directed to the reinforcement of afferentation evoked by external sounds at the background of the appropriate vocal-speech activity of a subject, in conditions of interference of two signals (both the test and the background ones)

recorded and reproduced by the tape. Test stimuli (1) were presented without any background sounds or against the background stimuli (2), (3) and (4) in the successive series of electrophysiological recording of auditory evoked potentials.

D. Electrophysiological experiments.

Auditory evoked potentials (AEP) of various latencies from the vertex were recorded with the help of standard method. According to Picton et al., (1974) AEP were classified as (1) short-latency - SLAEP (PLP up to 8 ms); (2) middle-latency - MLAEP (PLP 8-40 ms) and long-latency - LLAEP(PLP 40-350 ms).

RESULTS

I. Short-latency potentials - SLAEP.

1. The amplitude of waves 1-4 (the place of generation - auditory nerve, cochlear nucleus, superior olivary complex, lateral lemniscus) weekly depends on all types of sounds, representing the background of test clicks. There is a tendency ($0,1>p>0,05$) for the wave amplitude increase at the background of sounds presented binaurally from the tape-recorder (a song melody) and from the noise generator (white noise). There were no amplitude differences in response to the test-stimulus before and during the reproduction of song melody by the subject. The peak latency periods of the waves evoked by the test stimuli at the background of different sounds did not change.

2. The amplitude of wave 5 (the place of generation - lateral lemniscus, inferior colliculus) increases weekly during the subjects song production, without the difference reaching any significant level in comparison with the reaction to test stimulus. At the same time, the increase of the amplitude of wave 5 at the background of a noise and of a song melody presented from the tape-recorder is statistically significant ($p<0,01$). The increase of the amplitude at the background of noise is observed to be higher than at the background of the song produced by the subject as well as presented from the tape (the data for 10 subjects).

II. Middle-latency potentials - MLAEP.

1. The amplitude of wave evoked by test stimulus with peak latency less than 20 ms (the place of generation being midbrain, probably inferior colliculus and diencephalon) at the background of the subject's song production increases sharply, being higher than at the background of listening the same song melody from the tape recorder and white noise. Peak latent period of some waves is decreased during subject's song production ($p<0,01$).

2. The amplitude of waves with peak latency of 20 ms or more (the place of generation is probably diencephalon and thalamic nuclei) behaves similar to the amplitude of other MLAEP waves. Their peak latent periods shortened during song generation and kept of the same value as in response to the test stimulus presented without background or against the background of listening to the song melody from the tape-recorder and to noise (the results for 7 subjects).

III. Long-latency potentials - LLAEP.

1. All the components of LAEP (P_1 N_1 , P_2 N_2 , P_3 N_3) can be distinctly recorded only in response to the test clicks. During the song production by the subject the components P_3 and N_3 disappear, being recorded in response to click not in all subjects and not regularly in one and the same subject. Similar results were also demonstrated during listening to the self-produced song melody presented from the tape. Masked and diminished in amplitude these components of LLAEP can be

at the auditory input, and possible mutual depression of afferent flow at the different brain levels, that reflects in the evoked potentials in response to external sounds and subject's vocal activity. Cortical components P_3 and N_3 of LLAEP behave differently in principle as they do not appear at all in the conditions of subject's own sound production. Cortical activity to external stimuli is depressed at the background of vocal speech activity. One can think that physiological filtration of the afferent flow is accomplished at the subcortical levels, where the afferent flow is distributed to the executive structures of the brain. The cortical generators of P_3 and N_3 components may be required for activation of some directed attention and only with its participation can reflect different properties of the afferent flow.

CONCLUSIONS

1. The interaction between vocal speech and auditory system is fulfilled at all levels of the brain and is depicted in amplitude value and shape as well as in latency of the auditory evoked potentials. Most evident are the interactions reflected at the levels of middle- and long-latency auditory evoked potential generators.

2. The waves of middle-latency potentials (electrical correlates of the midbrain activity, probably of diencephalon and thalamus as well) in response to the test signal are increasing at the background of vocal speech activity and decreasing at the background of external sound action.

3. The initial components of long-latency auditory potentials (P_1 N_1 , P_2 N_2) behave in the same manner as middle-latency potentials. The wave N_2 P_3 has reciprocal relations to initial waves, especially sharp differences being found at the background of the subject's own vocal speech activity.

ACKNOWLEDGEMENTS

The research described in this publication is supported in part by Grants No NVS000 and NVS300 from the International Science Foundation and Russian Government.

DETECTING GHOST PHONEME : THE "LIAISON ENCHAINEE" IN FRENCH

Sophie WAUQUIER-GRAVELINES

Laboratoire de Psychologie Experimentale, Université René Descartes, Paris 5, C.N.R.S, 28 rue Serpente, 75006 Paris, FRANCE.

Equipe Linguistique et Informatique E.N.S Fontenay-St Cloud 31 avenue Lombard 92260 Fontenay-aux Roses FRANCE

ABSTRACT

This paper presents 2 experiments in which the detection of latent consonant of the "liaison enchaînée" in French is observed. Results suggest that the specific phonological nature of this segment impedes the subjects in locating word's boudaries.Apparently, they do not treat the words one by one in a strictly left-to-right parsing but rather use the phonological organisation of the speech stream to find word boundaries.

Key-words: Psycholinguistics, Speech Segmentation, Prosody.

1-SEGMENTATION AND "LIAISON ENCHAINEE" IN FRENCH.

"Understanding" spoken language is first of all a process of recognition of discrete words in continuous speech signal. This means that the hearer must locate the boundaries between words of each utterance that he hears. Now, contrary to a written text where words are isolated by blanks, speech signal does not comprise clear and systematic cues signalling the beginning and the end of a word. This is particularly true for French language in which the phonological phenomenon of "liaison enchaînée" (Encrevé, 1988) can remove the left boundary of word beginning with a vowel

The "liaison enchaînée" consists of the following double phenomenon : when two vowels are in contact at a word boundary, a latent consonant appears at the boundary between the two words and it is resyllabificated at the attack of the second word : ie "bon ami" (good friend): "bon" [bõ] and "ami" [ami] when the words are produced separately, but [bonami] when they appear together (figure 1).

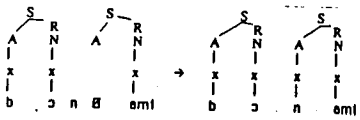


Figure 1 : The "liaison enchaînée"

The resyllabification of latent consonant occurs over the word boundaries. The consequence of such a phenomenon is the lack of correspondence between syllable boundaries and word boundaries and the creation of an erroneous beginning of word (*nami does not exist in French).

Moreover, this latent consonant has a singular status. As a matter of fact, it belongs lexically and morphologically to the first word (it is used in derivation "bon" [bõ] / feminine "bonne" [bon]) but when it is realized, it appears in the surface structure, it is located on the attack of the following word.

One can assume that this fact creates a difficulty for the segmentation process. How can a hearer locate the left boundary of a word when this word he has to access, has on the attack of the first syllable, a consonant that does not belong to its deep representation. In other words, how can the hearer treat this latent consonant on-line, how can he recover [bõ] and [ami] in the following utterance [bonami] which would be syllabificated as following : [[bo/na/mi] ?

In the following experiences we observed how the hearers resolve these questions during their perception of continuous speech.

2-EXPERIMENTS

2-1: Experiment 1

Because of the actualisation of the latent consonant on the attack of the first syllable of the word "ami" in "bon ami", the hearer must recognize on-line the specificity of this latent consonant in order to avoid a wrong segmentation (ie " bõ nami"). In order to test if the hearer showed a particular sensitivity with regard to this specific consonant, we used the Generalized Phoneme Detection (GPD) proposed by Frauenfelder & Segui (1989). This task consists in detecting target phonemes that might appear anywhere in an utterance. Thus, the subjects do not know where the target phoneme can appear. They cannot built a "expectation strategy" based on the place of the target in the word. The assumption was the following : if the subjects show a greater difficulty in detecting the phoneme (ie "n") in "un bon ami" [ɛ̃bonami] than in "un bon navire" [ɛ̃bõnavir], this means that this segment must be treated specifically and must be distinguished from the initial consonant. Consequently, this means that this kind of boundary requires a specific process which permits to remove the ambiguity. On the contrary, if there is no difference of treatment between initial consonants and liaison consonants, this might suppose that the subjects are not sensitive to the specificity of this segment.

Materials

8 pairs of sentences were built : each comprising one item with /t/ as initial phoneme and one with /t/ as liaison phoneme (i.e. " un grand éléphant" / "un grand téléphone"). They were inserted in test sentences. Test sentences were inserted in two lists of 70 sentence fillers. Half of them contained the target either at the beginning of the word (tapis), or inside the word (beauté), or in the end of the word (Colette). Each sentence contains only one target phoneme /t/. There were also 20 practice items. Half of them contained liaison /t/

Procedure

Sentences were read by a masculine speaker who did not know the aim of the experiment so that he should produce the liaison /t/ in a completely natural way. They were recorded on a REVOX 516 and digitized on Macinstosh SII. On a second channel, inaudible for the subjects, clicks were placed on the burst of the stop consonant targets.These clicks triggered a clock, which was stopped by the subject's keypress response.
The burst of the stop consonants were watched on a spectrogramm.

Subjects first listened to the 20 practice items over stereo headphones. They were instructed to respond as rapidly as possible by pressing a response key when they heard the target phoneme /t/ anywhere in each sentence. At the end of these practice trials, when the subjects did not detect any liaison /t/, they were explicitely asked to detect also the liaison phonemes. Then began the experimental trials.

Subjects

The subjects were 23 undergraduate students from the University of Paris V. They were French native speakers without any known hearing disabilities.

Results

For the analysis of the data, the reaction times below 100 milliseconds and those above 2,000 milliseconds were eliminated.

	Liais /t/	Init /t/	Diff
RT	619	573	46
% miss	7,8%	4,6%	

Table 1 : Reaction times (RT) in milliseconds and missing percentage (% miss) in detection of the target /t/ in pairs like "un grand éléphant" (liais /t/) and "un grand téléphone" (init /t/)

A Student test was performed taking the initial / liaison condition as random factor. The t values are given for subjects (t) and for items (t'). The analysis shows a significant difference of reaction times between the initial condition and the liaison condition for items ($t'(1,8) = 2,87$ $p = 0.02$) as for subjects ($t(1,19) = 2,89$ $p = 0.0009$).

Discussion

The results confirm the assumption that we have formulated above : the liaison consonants have been detected

more slowly than the initial ones. The process of this kind of segments is apparently more complex than the monitoring of initial phonemes. In "un grand éléphant" [ɛgratelefa], the resyllabification of the consonant induces some ambiguity with respect to the boundary. The subjects have a great difficulty to remove this ambiguity in order to be able to locate the left boundary and to access the word "éléphant" because of the realization of /t/ at the beginning of the word.

As we have seen, the "liaison enchaînée" implies a *phonetic* alteration of the speech signal but that is bound by *phonological* constraints. As a matter of fact, the actualization and the resyllabification of the liaison segment does not occur anywhere in the speech stream. The frequency and the distribution of the liaisons are bound by the nature of the phonological domain in which they appear (De Jong 1990). Thus, the liaison is considered obligatory in the clitic group (i.e "un ami") and quasi-obligatory in the phonological phrase ("un bon ami"). De Jong has shown that it is actualized by 99% of speakers in CG and by 75 % of speakers in PP. The more syntactically cohesive the group, the more obligatory the liaison. One can assume that this phonological production constraint also comes into play during the perception. In order to examine if this phonological constraint is a factor of the perceptual process of liaison phoneme, we replicated the first experiment but with liaisons in clitic groups.

2-2 Experiment 2

Materials and procedure

Materials was built in strictly identical way except for the nature of the domain in which the liaison appears. In this experiment, the target phoneme was /n/ as initial phoneme as liaison phoneme (i.e. " un navire" a ship [ɛnavir / "un avion" a plane [ɛnavjo]). The procedure was strictly the same. The subjects were 37 undergraduate students from the University of Paris V.

Results

Subjects have had many difficulties in performing the task. In spite of the increase of the number of subjects, the number of detected liaison target phonemes was insufficient to perform a statistic analysis. Then, the missing detections had been systematically analyzed.

	Liais /n/	Init /n/
% miss	40 %	14 %

Table 2 : missing detections (% miss) for detection of the target /n/ in pairs like "un avion (liais /n/) and "un navire"t (init /n/)

A Student test was performed taking the initial / liaison condition as random factor. The t values are given for subjects (t) and for items (t'). The analysis shows a significant difference of percentage of detection between the initial condition and the liaison condition for items (t'(1,7) = 4,22 p= 0.0039) as for subjects (t (1,36) = 5,75 p = 0.00002). Although, the subjects had been explicitely asked to detect also the liaison consonants, they missed many target phonemes. They didn't seem to perceive this kind of segment, as if they were deaf in regard to these liaison consonants.

Although the results of experiment 1 (RT) cannot be directly compared with those of experiment 2 (% of missing responses), in both cases the subjects have shown many difficulties to detect liaison consonants, but most caracteristically in clitic groups. This detection seems quite impossible as if the more obligatory the liaison, the more difficult the target detection.

3-DISCUSSION

This difficulty can be interpreted in two ways. One can first make the hypothesis that initial consonants and liaison consonants do not have the same acoustic realization. In that case, the differences between liaison condition and initial condition could be assigned to a subject's sensitivity to this kind of information. Thus, this would mean that they have used low-level acoustic information to locate the words' boundaries and process the segmentation of the speech stream.

Acoustic hypothesis

It is traditionally and generally admitted in phonological literature that the realization of initial consonant, and liaison consonant, are acoustically identical (Encrevé, 1988). However, a few results (Durand, 1953; Bradley & Dejean, 1990; Bresson & Grosjean, 1994) showed that the initial consonant in PP was significantly longer than the liaison consonant. More particularly, Dejean has shown that VOT and occlusion's duration are shorter in the case of liaison consonants.

We made the same measurements of our materials. The measurements were made on spectrograms in Unice (table 2).

	Init	Li	Dif	t 1,7	p
VOT	49	43.5	5.5	1,14	0.29
Occl	69.7	49.7	20	3.42	0.01
Dur	120	95	25	4.145	0.0043

Table 3: VOT, occlusion's duration (occl) and consonant's duration (dur) of /t/ in pairs like " un grand éléphant" / "un grand téléphone"

Our results confirm the inferences made by Dejean : the initial phonemes are longer than liaison consonants. One can think that this difference of duration explains the difference of RT. However, we have calculated a correlation between the results of the expertiment 1 (RT) and the durations that we have measured (VOT, occl, dur), and, we have found no significant correlation between the RT and the durations of the consonants. Moreover we have also compared acoustic realizations of initial /n/ and liaison /n/ and we have found no significant difference (table 4).

	Init	Li	Diff	t 1,7	p
Dur	61	57,8	3,2	1.15	0.29

Table 4 : duration (dur) of /n/ in pairs like ""un navire" / " un avion".

This would mean that this kind of acoustic information is not preferentially used by the hearers to resolve the ambiguity created by a liaison on a word's boundary.

One can also make the hypothesis that the phonological specificity of the liaison consonant is what the subjects have to consider.

Phonological hypothesis

In those experiments, the results have shown that the subjects have many difficulties in performing the phoneme-monitoring task for the liaison phonemes as if they could not isolate them as "phonemes". Moreover, the cohesion of the phonological domain in which this liaison appears seems to be a factor of the subjects incapacity to isolate the target from the group and to recognize it. This would mean that this kind of groups is processed as a single unit and is not split in two or three words (ie article, adjective, noun) that are isolately and successively accessed. Besides, this kind of free lexems could be lexicalized in French (ie un bon homme (a good man) > un bonhomme (a fellow).

Thus those results suggest that the segmentation in French, as it has been proposed by Christophe (1993) could be processed by phonological units larger than the words in which the lexical access could occur.

REFERENCES

[1] Besson, C & Grosjean, F (1995), L'effet de l'enchaînement sur la reconnaissance des mots dans la parole continue. *L'Année psychologique*, à paraître.

[2] Bradley, D.C. & Dejean de La Bâtie, B. (1990) . Resolving boundaries in spoken french. In R Seidl (ed) *Proceedings of the Third Australian International Conference of Speech Science Technology* Canberra: Australian Speech Science & Technology Association.

[3] Christophe, A, (1993), *Rôle de la prosodie dans la segmentation en mots.* Phd thesis, EHESS, unpublished.

[4] De Jong, D. (1990). La liaison à Orléans (France) et Montréal (Québec) in *Actes du XIIèmes Congrès International des Sciences Phonétiques*, Aout 1991, Aix-en Provence, France, pp198-201.

[5] Encrevé, P. (1988) . *La liaison avec et sans enchaînement.* Paris : Le Seuil.

[6] Frauenfelder, U.H & Segui, J (1989) Evidence for associative context effect in phoneme monitoring, *Memory and Cognition*, 17, pp 134-140.

THE ACOUSTIC CHARACTERISTICS OF WHISPERED PLOSIVES AND THEIR RELIABILITY FOR THE PERCEPTION OF 'VOICING'

Sandra P. Whiteside (1) & Kevin L. Baker (2)
(1) Speech Science, University of Sheffield, Sheffield S10 2TA, U.K.
(2) Department of Human Communication, De Montfort University, Leicester LE7 9SU, UK.

ABSTRACT

This study examines a range of acoustic characteristics of whispered plosives in initial and final position in minimal word pairs produced by two speakers. Perception tests are carried out to see whether listeners can make judgements about whether the whispered stimuli represent 'voiced' or 'voiceless' tokens. Whether the acoustic characteristics of the whispered stimuli are reliable for the perception of 'voicing' is discussed.

INTRODUCTION

From the few studies that have been carried out into whispered speech, it is apparent that listeners have little difficulty in perceiving vowels. Kallail and Emmanuel [1, 2] presented lengthened and isolated vowels to their subjects and reported that between 63 and 65% were correctly identified when whispered compared to 80% when spoken normally. Tartter [3] improved on this study by presenting whispered CVC syllables following observations by Strange [4] that in normal speech, formant transitions are important for vowel identification. Tartter [3] found a better than 80% identification rate for 10 vowels whispered by 6 speakers compared to over 90% for the normally voiced vowels.

Tartter [5] presented whispered consonant- [a] syllables to 6 listeners and found that the overall identification score was 64% with a 72% accuracy for identifying 'voicing'. However, given that her data included other consonants in addition to stops it is difficult to ascertain the levels of accuracy for the stop consonants alone. Dannenbring [6] investigated 12 subjects' ability to discriminate between whispered consonants in CV syllables, where the vowel was either /i/, /a/ or /u/. Dannenbring's results show that listeners were able to make discriminations with confidence but does not provide correct identification scores which makes his results difficult to compare with other studies. Munro [7] presented 32 whispered tokens of /p/ and /b/ in four vowel contexts to 8 listeners where he found an overall mean correct identification score of 63%. Although he showed that the whispered /b/ tended to have a steeper rise slope than the whispered /p/, these showed no relationship to the pattern of identifications. He concludes that it is dangerous to make 'inferences about perceptual mechanisms on the basis of production data alone' (p.180). The present study is intended as an preliminary investigation into whether 'voicing' contrasts in whispered stop consonants (or plosives) in words can be identified in the absence of both the laryngeal voice source and meaningful contextual information (e.g. a meaningful sentential framework). Vowel context was not controlled for. Instead, the focus of this study was placed upon the place of articulation of the whispered stop consonants in initial and final position in minimal word pairs. Listeners were asked to make judgements about whether whispered stop consonants in word initial and word final position were 'voiced' or 'voiceless'. Subsequently acoustic measurements were taken for the word initial and word final stimuli.

METHOD

We presented whispered stimuli to subjects who were given a forced choice for their identification. The forced choice was between the presented stop consonant and its 'voiced' or 'voiceless' counterpart. For example, if the whispered token A PAT was presented to the subject, then the word pair A PAT and A BAT was visually presented as the choice for identification.

Subjects

The authors served as speakers for the recorded speech samples. Both speakers are native speakers of British English and are in their late twenties. Five female and five male subjects with normal hearing served as subjects for the perceptual part of the study. All listeners native speakers of British English with an age range of 20 to 34 years.

Stimuli

The speech samples consisted of 55 CVC whispered words in the frame 'a 'CVC'. This frame was used to produce even stress. They were recorded once

both by an adult female speaker (F) and an adult male speaker (M). The stimuli are shown in table 1 and form 30 minimal word pairs for stop consonants in word initial position and 30 minimal word pairs for stop consonants in word final position (5 of the words are used in more than once). The minimal word pairs represented bilabial (B), alveolar (A) and velar (V) places of articulation.

Table 1. Whispered minimal word-pairs

	Word Initial	Word Final
B	a pat/ a bat	a lap/ a lab
	a peat/ a beat	a tap/ a tab
	a pack/ a back	a swap/ a swab
	a pay/ a bay	a cop/ a cob
	a pig/ a big	a cap/ a cab
A	a tip/ a dip	a pat/ a pad
	a tab/ a dab	a lit/ a lid
	a tuck/ a duck	a fat/ a fad
	a tart/ a dart	a sort/ a sword
	a toe/ a doe	a lout/ a loud
V	a cod/ a god	a lack/ a lag
	a cold/ a gold	a tack/ a tag
	a cape/ a gape	a back/ a bag
	a coat/ a goat	a tuck/ a tug
	a cap/ a gap	a rack/ a rag

The 60 words were repeated once and randomised into a list for recording.

Recording

Each whispered word was recorded while the speaker was seated in a sound proof chamber. The whispered speech was recorded digitally using an Apple Macintosh Classic II computer via a microphone connected to a Farallon MacRecorder™. The sampling rate was set at 22kHz (8 bit). The MacRecorder digitizer filtered the analogue sound with a cut off of 11 kHz.

Perception Tests

Subjects were seated in the sound proof chamber with a loudspeaker and a computer 'mouse'. Outside the chamber the Apple Macintosh was placed in view of the subject through a window in the chamber, and connected to the mouse. A Hypercard™ (Apple Computer Inc., 1990) program written by the second author, was used to play the speech samples from the stimuli list, present the appropriate word pair on the computer screen, and to record the judgements made by the subjects. The subjects used a computer mouse to play back the stimuli and make their forced choices about the stimuli they were presented with. Each subject repeated the experiment so that they made judgements of both the male and female speech stimuli.

Acoustic Analysis

Possible acoustic cues to the perception of 'voicing' were investigated for the whispered stop consonants. These acoustic cues were examined using a KAY Computerised Speech Lab (CSL) Model 4300. The whispered speech stimuli were transferred from the Apple Macintosh computer onto digital audio tape (DAT) and then transferred on to the KAY CSL using a sampling rate of 10 kHz. The methods of analysis used for each of the measurements are outlined below.

For the word-initial stimuli the following measurements were taken: i) The amplitude (dB) of the plosive burst using the graphical results of an algorithm which computes an energy envelope in dB SPL from the speech pressure waveform of the whispered speech sample; ii) The interval (ms) between the peak amplitude of the plosive burst and the peak amplitude of the following noise-excited vowel from the computed energy envelope (dB SPL) using the graphical interface provided by the CSL; iii) The amplitude difference (dB) between the peak of the burst and the following vowel. This was done using a similar method as for i) and ii); iv) The closure duration (ms) of the initial plosive measured from the end of the preceding schwa (/ə/ to its release; v) The release phase (ms) of the initial plosive, measured from the point of the plosive's release to the onset of F1 in a wide band FFT spectrogram and vi) The overall energy (SPL dB) of the CVC using the same method as measures i) to iii). The statistical results of these analyses can be found in table 4 below.

For the word-final stimuli the following measurements taken were: i) the frequency (Hz) of the first formant (F1) offset preceding the closure for the word-final plosive, using an FFT wideband spectrogram and a graphical interface which allows the measurement of formant frequency values; ii) The duration (ms) of the noise-excited vowel preceding the closure, given that for post-vocalic plosives one of the acoustic cues of voicing is the duration of the preceding vowel, where a shorter vowel duration cues voicelessness [8]. This was done using the FFT spectrograms and the graphical interface. The duration of the vowel was taken from the point immediately following the plosive burst of the preceding plosive until the acoustic closure for the final plosive. So

for example, for the stimulus A PAT (/ə 'pæt/) the duration of /æ/ would be taken immediately following the plosion of /p/ until the acoustic closure for /t/; iii) The duration (ms) of the acoustic closure following the vowel and preceding the final release of the plosive; iv) The energy (SPL dB) of the release burst of the final plosive using the method described above and v) The overall energy (SPL dB)of the CVC as described above. The statistical results of these analyses can be found in table 5 below.

RESULTS AND DISCUSSION

Perception Tests

Table 2 provides a summary of perception test results. χ^2 tests were carried out on the identification scores with the assumption that the expected identification of the consonants would be at chance level (i.e. 50%). These results are given in table 3.

From table 2 we can see that the mean correct perception scores range from 41% to 100%. This represents an overall mean of 77% and 96% for the word initial and word final stimuli respectively. If we look at the results in more detail we find a variation in the identification results for each place of articulation. For example the 'voiceless' alveolar stimuli are identified for both the male and female stimuli with most accuracy (mean of 98.5%). In addition, the 'voiced' bilabial stimuli for the male speaker are identified with the least level of accuracy (41%) followed by the 'voiced' velar stimuli of the female speaker (42%). What is evident from table 2 is that the 'voiced' word-initial stimuli are identified with lower levels of accuracy compared with their 'voiceless' counterparts, a finding also made by Tartter [5].

For the word final whispered stop consonants the number of stimuli correctly identified ranged from 60% to 100% with an overall mean identification score of 96%, much higher than the word initial scores. These findings suggest that the listeners had little trouble identifying whispered stop consonants in word final position. The correct identification of the word initial and word final whispered consonants are significantly above the chance level of 50% expected if identification of the consonants was based on 'voicing' which of course is absent in our stimuli (see table 3).

Tables 4 and 5 shows the t-scores and their significance levels for the acoustic parameters of the 'voiced' and 'voiceless' pairs of word initial and word final stimuli

Table 2. Summary Table of Perception Scores (%)

	(M%, F%) Mean Word Initial	(M%, F%) Mean Word Final
Bilabial -v /p/	(80, 89) 84.5	(99, 99) 99
+v /b/	(41, 69) 55	(89, 91) 90
Alveolar -v /t/	(97, 100) 98.5	(99, 99) 99
+v /d/	(63, 83) 73	(86, 99) 92.5
Velar -v /k/	(90, 98) 94	(98, 100) 99
+v /g/	(72, 42) 57	(94, 99) 96.5

Table 3: χ^2 values for perception scores assuming chance levels of 50%.

	Word Initial Male Female	Word Final Male Female
Bilabial	122.5 289	402 423
Alveolar	268 377.5	395.5 481
Velar	260 332	428 490.5

All values are $p \leq 0.0001$.

respectively. We can see from these scores that of the 36 t-scores for the word initial stimuli, only 10 (5 bilabial, 3 alveolar and 2 velar) are statistically significant, whereas for the word final stimuli 16 out of the 30 scores are statistically significant. The latter findings lend some support to the better perception scores for the word final stimuli. However, although the acoustic data for the word initial stimuli show less statistical significance, the χ^2 values given in table 3 are statistically significant. This suggests that there was enough information in the whispered plosives for listeners to make accurate judgements about 'voicing' in the absence of laryngeal voicing.

From table 4, we can see that significant differences between the 'voiced' and 'voiceless' tokens were variable and patchy. For example, highly significant differences were found for both speakers in the 'release phase' of the bilabial stimuli. However, for the 'burst amplitude' parameter, no significant differences were found between the 'voiced' and 'voiceless' stimuli . The findings suggest that different acoustic cues in the whispered stimuli may be operational for different tokens and different speaker characteristics in the perception of 'voicing' in plosive-initial position.

Table 5 also shows variation in the levels of statistical significance for all the stimuli, however it also shows that the vowel duration and duration closure parameters show significant differences for all the 'voiced' and 'voiceless' stimuli. This suggests that these parameters may be playing a key role in the perception of voicing. Given that the vowel and closure duration preceding the final plosive are available to the listeners for a longer period of time, it is probable that these acoustic characteristics are serving as a robust cues in the perception of voicing .

However one must also bear in mind that there may be other acoustic cues operating in the perception of voicing for the word-initial stimuli that we have not considered in this study. Further research is planned in this area.

Table 4. T scores for Word Initial Acoustic Parameters .

Ac. Param.	Bilabial	Alveolar	Velar
Burst Amp.			
M	0.227	-1.532	-0.348
F	0.708	-2.013	-0.686
Peak to Peak Dur.			
M	3.885*	0.157	0.903
F	-1.796	-3.131*	1.191
Amp. Diff.			
M	0.722	1.719	-1.113
F	3.059*	1.364	0.985
Closure Dur.			
M	1.444	3.258*	-1.135
F	-0.61	-2.011	-5.63**
Rel. Phase Dur.			
M	17.77**	0.333	9.66**
F	13.51**	5.44**	2.031
Overall Energy			
M	-1.363	-2.607	1.574
F	-3.721*	1.114	0.459

* significant at $p \leq 0.05$,
** significant at $p \leq 0.01$

Table 5. T scores for Word Final Acoustic Measures.

Ac. Param.	Bilabial	Alveolar	Velar
F1 offset			
M	-1.395	-2.619	-5.17**
F	-2.978*	-2.671	-3.13*
Vowel Dur.			
M	5.1**	4.07*	3.983*
F	4.433*	4.362*	6.066**
Energy of Rel.			
M	-0.723	0.843	-0.657
F	1.026	2.479	1.646
Dur. of Closure			
M	17.39**	4.172*	7.665**
F	7.621**	6.022**	6.491**
Overall Energy			
M	1.62	-0.627	0.065
F	3.566*	-0.75	1.156

* significant at $p \leq 0.05$,
** significant at $p \leq 0.01$

REFERENCES

[1] Kallail, K. L. and Emmanuel, F. W. (1984a). An acoustic comparison of isolated whispered and phonated vowel samples produced by adult male subjects, *Journal of Phonetics*, vol. 12, 175-186.

[2] Kallail, K. L. and Emmanuel, F. W. (1984b). Formant frequency differences between isolated whispered and phonated vowel samples produced by adult female subjects, *Journal of Speech and Hearing Research*, vol. 27, 245-251.

[3] Tartter, V. C. (1991). Identifiability of vowels and speakers from whispered syllables. *Perception and Psychophysics*, vol. 49, 365-372.

[4] Strange, W.(1989) Evolving theories of vowel perception, *Journal of the Acoustical Society of America*, vol. 85 2081-2087.

[5] Tartter, V. C. (1989). What's in a whisper? *Journal of the Acoustical Society of America*, vol. 86, 1678-1683.

[6] Dannenbring, G. L. (1980). Perceptual discrimination of whispered phoneme pairs. *Perceptual and Motor Skills*, vol. 51, 979-985.

[7] Munro, M. J. (1990). Perception of 'voicing' in whispered stops, *Phonetica*, vol. 47, 173-181.

PERCEPTION OF ORAL RELEASE RATE FOR INITIAL VOICED STOPS

David R. Williams
Sensimetrics Corporation, Cambridge, MA 02139, USA

ABSTRACT

Listener preferences for variations in oral release rate (ORR) of initial labial, alveolar and velar voiced stops were tested. Three five-member continua were synthesized using a set of articulatory parameters to control the increase in oral cross-sectional area at stop release. For each continuum, naive and experienced subjects judged the typicality of each member against all others. Faster ORRs were preferred for labial stops than for alveolar and velar stops.

INTRODUCTION

Because of differences in mass of the primary articulator and contact length, initial rates of increase in cross-sectional area of the oral constriction may differ for stops made at the various places of articulation. Based on theoretical considerations, acoustical analyses and some physiological data [1, 2], faster rates of increase in cross-sectional area are expected for stops formed at the lips than for those formed by the tongue tip or by the tongue body. These differences in oral release rate (ORR) have acoustic and aerodynamic consequences which may play an important role in the perception of stop place of articulation.

This paper examines the perception of syllable-initial voiced stops that were synthesized with differences in ORR. The synthesis was achieved by means of a new approach which employs a set of articulatory parameters to control a Klatt synthesizer [3]. The purpose of the experiments was twofold. First, there was an interest in assessing the approach itself to determine the precision with which control parameters must be specified in order to achieve acceptable synthesis of certain phonetic categories. Second, it was of interest to examine listeners' ability to discriminate within-category variations for a previously untested articulatory variable.

THE HL SYNTHESIS APPROACH

The underlying motivation for the synthesis approach described here is to combine the simplicity of control that characterizes articulatory approaches to synthesis with the accuracy and computational efficiency of traditional formant synthesis [4, 5]. This hybrid approach employs a small set of high-level (HL) parameters to construct an acoustic-articulatory utterance specification which is then transformed by means of mapping relations into a specification in terms of the larger set of lower-level (LL) acoustic parameters needed to control a KLSYN88 formant synthesizer [6]. In effect, the HL synthesis system is a formant synthesizer preprocessor.

The HL synthesis parameters and their functions can be described in terms of three broad classes:

Class 1 parameters control the first four natural frequencies of the vocal tract (***f1***, ***f2***, ***f3***, ***f4***) and the fundamental frequency (***f0***). These parameters specify the configuration and slow movements of articulators which determine the global shape of the vocal tract.

Class 2 parameters control the cross-sectional areas of local constrictions formed by the lips (***al***) and by the tongue tip or blade (***ab***). They specify the fast movements of primary articulators that rapidly decrease/increase airflow within the oral tract.

Class 3 parameters control the cross-sectional areas of the glottal orifice (***ag***) and velopharyngeal port (***an***) and the pharyngeal volume (***ue***). They are used to specify opening/closing movements of the glottis and velum as well as active expansion or contraction of the pharynx.

The first step in determining values for the LL parameters is to calculate the pressures and flows at the supraglottal and glottal orifices using an aerodynamic model [7]. In addition to the Class 2 and 3 parameter values, inputs to the model include *agx*, the glottal orifice area as modified by supraglottal forces, and *acx*, the smallest supraglottal constriction area. The output of the model is the intraoral pressure (P_m) which, along with the orifice areas and a constant subglottal pressure (P_s) value, provides a basis for computing the LL source amplitudes **AV**, **AH** and **AF**.

Other settings and modifications of the LL parameters result from values of HL parameters specified by the user. In general, the Class 1 parameters are mapped directly to their corresponding LL parameters when the glottal area is modal and the velum is closed. An increased glottal area *agx* affects the LL formant bandwidths and the values of **OQ** and **TL**. The presence of voicing in the synthesis signal is conditional on *agx* and the calculated transglottal pressure drop (**AV** = 0 when $agx > 13$ mm^2 or $P_s - P_m < 3$ cm H_2O). Place-specific filtering of the frication is determined from a look-up table when **AF** > 0 based on the values of ***f2*** and ***f3***.

Of particular interest for the current study are rules that affect the value of **F1**. Specifically, the HL first natural resonance frequency is modified (*f1c*) when a class 2 parameter specifies a local (labial or alveolar) constriction to reflect the fact that the constriction is currently controlling its value. The value of *f1c* is approximated as the lowest frequency of a Helmholtz resonator with constriction area *acx* and with a constriction length and pre-constriction volume that are determined by the place of articulation. When the natural frequencies specify a tongue dorsum (e.g., velar) constriction, *acx* directly reflects the value of ***f1***.

[Other rules that are operative when the velopharyngeal port is open or there is lateralization or retroflexion of the tongue are not discussed here.]

EXPERIMENT

The purposes of the experiment were to (1) determine the range of acceptable ORR values, and (2) to examine the range of listeners' preferences for various ORRs in voiced stops with different places of articulation. Although ORR is acoustically a very complex variable, it is simulated here by changes in a single HL parameter which controls the rate of increase in area of an oral constriction.

SYNTHESIS

HL Input Parameters

The synthetic stimuli were modelled on three /ba/, /da/, and /ga/ utterances produced by a male talker. All stimuli were 300 ms in duration and had an ***f0*** contour which fell from 115 Hz after the burst to 85 Hz at the end of the vowel. During the vowel's steady-state portion, the natural frequency values were 700, 1150, 2400, and 3500 Hz. Initial ***f1***, ***f2*** and ***f3*** transitions were derived from the spoken utterances, and thus, were place-specific. The glottal area ***ag*** was constant at a (modal) value of 4 mm^2 throughout the stimuli.

Three five-member series of CV syllables were synthesized by varying the rate of increase in the area of oral constriction at stop release. For the labial and alveolar series, this entailed setting the slope of, respectively, the **al** and **ab** parameter trajectories to correspond to ORRs of 10, 15, 25, 50 and 100 cm^2/s. For the velar stops, the value of ***f1*** was manipulated so as to achieve *acx* values that corresponded to ORRs of 10, 15, 20, 30 and 50 cm^2/s.

Output Parameters

The effect of decreasing ORR was to increase the number of frication (**AF**) frames in the burst and to decrease the initial aspiration (**AH**) level. For the slower ORRs, voicing (**AV**) onset was not simultaneous with oral release, but was delayed by as much as 15 ms due to high intraoral pressure. Initial changes in **OQ** and **TL** were much more abrupt for the faster ORRs than for the slower ones.

METHODS

For each place of articulation series, stimulus trials were constructed by pairing each ORR with every other ORR, including itself. Five randomized blocks of the 25 distinct stimulus pairs in each series were recorded on audio tape for presentation. The stimuli were presented over headphones in a quiet room.

Subjects were asked to compare the two stops in a pair and indicate whether the first or second member was the better exemplar of the particular voiced stop. An answer was requested on each trial, even if it was thought only to be a guess.

SUBJECTS

A group of fourteen "naive" subjects were recruited from the local academic community. A short questionnaire filled out prior to testing revealed that all subjects spoke only English, and none had a history of hearing impairment. The subjects were paid for their participation.

Twelve "experienced" subjects, all volunteers from the MIT Speech Comm. Group, were also tested. This subject group included senior graduate students, postdoctoral researchers and the author. All were native speakers of English who were familiar with phonetics and had experience judging synthetic speech.

RESULTS

Although all members of a series sounded like exemplars of that stop type, there were clear differences among the stimuli within a series. Most notably, at the slowest ORR, all stops sounded somewhat voiceless and even fricated. In the alveolar and velar series, stimuli with the fastest ORRs had a somewhat intrusive /y/-glide following the stop.

Table 1. Normalized scale values for 14 "naive" subjects at each ORR. Second ORRs are for the velar series.

Normal Scale Values	ORR in cm^2/s				
	10	15	25 /20	50 /30	100 /50
/ba/	-0.99	-0.38	0.12	0.54	0.71
/da/	-1.12	-0.12	0.73	0.30	0.22
/ga/	-0.44	0.18	0.00	-0.18	0.44

Naive subjects

Table 1 shows normalized scale values for the naive subjects. The scores were computed by first transforming the percentage of times that the indicated stimulus was preferred over each other member of a series (in both positions) to normalized (z-score) units and then summing over scores (see [8]). The table shows the average normalized scale values for the pooled data. (A value of 0 represents no preference for or against.)

The scale values indicate that the naive subjects preferred the two fastest ORRs for the labial stops and the intermediate (25 cm^2/s) ORR for the alveolars. There was clearly a negative preference for the slowest ORR. For the velars, it would appear that the fastest ORR was again preferred. However, the range of scale values here are small, and the raw percentage scores did not suggest any strong preferences.

Experienced subjects

Table 2 shows the results for the 12 experienced subjects. Like the naive subjects, this group preferred the intermediate ORR for the alveolars and the faster ORRs for the labials. The lower labial scale values and spread in preference to include the intermediate ORR (/ba/) are due to the fact that three

subjects preferred the slower ORRs over the two fastest ORRs (cf. /ba/-9). Nevertheless, there was overall a strong negative preference for the slowest labial and alveolar ORRs.

Table 2. Normalized scale values for 12 "experienced" subjects at each ORR. Second ORRs are for the velar series.

Normal Scale Values	ORR in cm^2/s				
	10	15	25 /20	50 /30	100 /50
/ba/-9	-1.06	-0.30	0.31	0.43	0.62
/ba/	-1.11	-0.14	0.39	0.36	0.50
/da/	-0.51	-0.29	0.82	0.26	-0.86
/ga/	-1.05	-0.26	0.17	0.75	0.39

Unlike the data in Table 1, these data reveal a strong preference for the velar stops with the 30 cm^2/s ORR and a strong preference against the slowest velar ORR. It is also notable that the experienced subjects were much less tolerant of the /y/-glide following the release of the aveolar stop with the fastest ORR.

DISCUSSION

The results show that intermediate ORRs were preferred for all places of articulation, and that faster ORRs were also preferred for the labial stops; slower ORRs were generally not preferred. It is thus possible that a single ORR (ca. 30-40 cm^2/s) could be used to synthesize all voiced stops. Although a range of ORRs were found to be acceptable within each stimulus series, the gradient of scale values differed as a function of stop place of articulation.

CONCLUSIONS

The present findings demonstrate listeners' ability to discriminate within-category phonetic variations, and thus contribute to the ongoing discussion of phonetic prototypes. As the stimuli used here were all articulatorily plausible, the study represents a refinement on previous assessment techniques. Relatedly, these results and others [9, 10] suggest that the scope of potential prototype definitions might reasonably be expanded to embrace speech production as well as perception. [Research supported in part by a grant from NIH.]

REFERENCES

[1] Fant, G. *Speech Sounds and Features*. Cambridge: MIT Press, 126.

[2] Stevens, K. N. (forthcoming) *Acoustic Phonetics*.

[3] Stevens, K. N., and C. A. Bickley (1991) "Constraints among parameters simplify control of Klatt formant synthesizer." *J. Phonetics 19*, 161-174.

[4] Stevens, K. N., C. A. Bickley, and D. R. Williams (1994) "Control of a Klatt synthesizer by articulatory parameters." *Proceedings 3rd Int'l. Conf. Spoken Language Processes*, Yokohama, Japan.

[5] Williams, D. R., K. N. Stevens, and C. A. Bickley (1992) "Inventory of phonetic contrasts generated by high-level control of a formant synthesizer." *Proceedings 2nd Int'l. Conf. Spoken Language Processes*, Banff, Alberta, Canada, 571-574.

[6] Klatt, D. H, and L. C. Klatt (1990) "Analysis, synthesis, and perception of voice quality variations among female and male talkers." *JASA. 53*, 1070-1082.

[7] Stevens, K. N. (1993) "Models for the production and acoustic of stop consonants." *Spch. Comm. 13*, 367-375.

[8] Green, B. (1974) "Paired comparison scaling procedure." In G. M. Maranell (ed.), *Scaling: a sourcebook for behavioral scientists*. Chicago: Aldine. Pp. 93-97.

[9] Williams, D. R. (1994, June) "Modelling changes in magnitude and timing of glottal and oral movements for synthesis of obstruent consonants." *JASA 95*, 2815 (A).

[10] Williams, D. R. (1994, November) "Perception of fricatives synthesized by higher-level control of a Klatt synthesizer." *JASA 96*, 3227 (A).

CLICK ARTICULATIONS IN XHOSA: NEW PERSPECTIVES THROUGH WIGNER DISTRIBUTION ANALYSIS

Justus Roux, Grzegorz Dogil** and Wolfgang Wokurek***
** Research Unit for Experimental Phonology, University of Stellenbosch, RSA*
*** Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Germany*

ABSTRACT

Three phenomena involving click articulations are investigated using a technique which reveals time-frequency detail not normally achieved with traditional methods. The results suggest an alternative description for affricated and murmured clicks and point to idiosyncratic features in nasal clicks.

INTRODUCTION

This paper focuses on click articulations and their accompaniments in Xhosa employing a technique, which to the best of our knowledge has never been used for the acoustic analysis of these sounds. This technique, referred to as Smoothed Pseudo-Wigner Distribution (SPWD)-analysis combines both the good frequency resolution of narrow band spectrograms and the good time resolution of wide band spectrograms [1]. The acoustic description of clicks in general has received a good degree of attention over the last few decades, culminating in the work of Ladefoged & Traill [2] (henceforth L&T), presenting the most detailed phonetic description of clicks to date. Despite all the attention, there are still a number of issues related to Bantu languages such as Zulu and Xhosa that remain unresolved. We are of the opinion that SPWD-analysis may shed more light on these unresolved issues, specifically due to the fact that it yields a time-frequency resolution in which the finegrained signal structures are considerably more evident than in spectrograms and waveform representations, such as those presented by L&T. Only three **unresolved issues** regarding clicks in Xhosa will be dealt with here, i.e. those relating to, respectively, so-called affricated clicks, nasal clicks, and voiced "murmured" clicks. A comprehensive discussion of all the relevant issues appear in [3].

ACOUSTIC ANALYSIS

Material

The speech of two male native speakers of Xhosa were recorded in a sound studio with high quality equipment. The test data consisted of nonsense /VCV/ utterances where the three click types and their accompaniments (cf. Table 1) occupied the C-slot. The five basic vowels of Xhosa (/a, e, i, o, u/) varied systematically in the V-slots and the tokens were read in a carrier phrase *Ndithi.... (I say....)*

Analysis

The data were edited and analyzed with the S-Tools™ system focusing on the burst of the click and extending the window to at least the first three detectable periods of the following vowel. The

Table 1. Xhosa click consonants and accompaniments (orthographic representation). The use of phonetic symbols will be minimalized in this paper in view of some controversies. (Cf. [3]).

Accompaniment	Click type		
	Dental	Alveo-palatal	Lateral
Voiceless	c	q	x
Aspirated	ch	qh	xh
Nasal	nc	nq	nx
Murmured voiced	gc	gq	gx
Murmured nasal	ngc	ngq	ngx

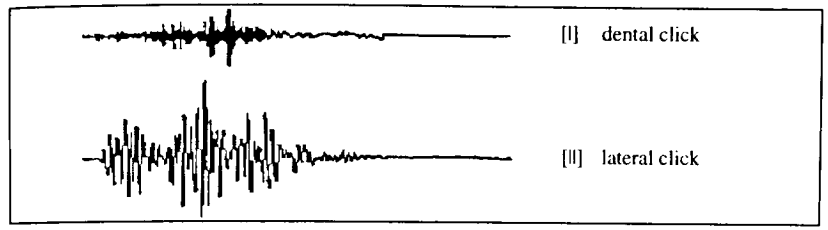


Figure 1. Waveforms of noise bursts for dental and lateral clicks indicating "considerable turbulent noise after the release" presented by Ladefoged & Traill [2:40]

theoretical foundations and implementation of SPWD is presented in [4]; suffice to say, however, that the **release transient** of a stop is shown in SPWD as a combination of a vertical line followed by horizontal lines. The vertical line represents an impulse like signal component caused by a sudden release of air behind a constriction. The horizontal lines are interpreted as damped oscillations (formant frequencies) forming the response of the vocal tract to the exciting plosion-pulse. Noise bursts are represented in SPWD as an irregular grid filling a specific area of the time-frequency plane.

RESULTS AND DISCUSSION

Affricated clicks

The earliest descriptions of click sounds mention that some clicks, notably the dental/alveolar "c" [ǀ] and the lateral "x" [ǁ] clicks are pronounced with friction or as "affricative" segments. This view is also supported by L&T. In everyday phonetic terms an affricate is considered to be a stop with affricated (fricative) release [5,6]. In an acoustic study of Xhosa clicks Sands concludes: "It seems that the lateral and the dental clicks are made with a long constriction release, which is more gradually, causing the release to be affricated." [7:3]. Referring to the respective waveforms of the dental and lateral clicks (Figure 1) L&T [2:40] state that these clicks have "... considerable turbulent noise after the release." They note a crescendo and decrescendo effect and continue: "For the noisy clicks, after the anterior closure is released, the noise increases in intensity until it reaches a maximum, after which the noise decreases in intensity." [2:41]. Their position seems to be quite clear: the silence of the closure phase of the click is broken by the anterior release after which the frication builds up to a maximum and then decreases before the transition to the following vowel is made, hence an **affricated click** is constituted.

Now consider the following SPWD-analysis of such an alleged affricated dental click in Xhosa:

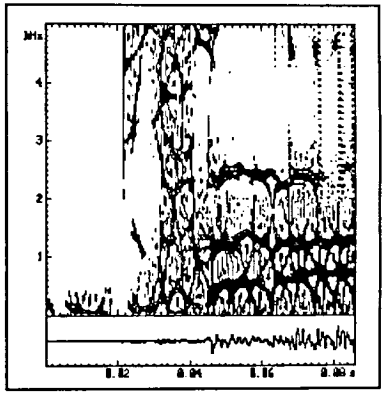


Figure 1. SPWD analysis of the noise burst of a voiceless dental click "c" in Xhosa

Following the silence of the closure phase (0-0.02s) frication excitation extending over the whole spectrum is clearly attested up to a point (0.046s) where impulse excitation is attested (a vertical line extending into a horizontal island at 1200 Hz). This impulse then extends towards the first and second formant frequencies of the following low vowel /a/ without excessive high frequency turbulence. The acoustic pattern of this articulation is unambiguously clear: frication followed by an impulse. This order of acoustic events is not reconcilable with the notion of affrication traditionally assigned to it; as a matter of fact the opposite, **pre-affrication** actually takes place. Translated into articulatory terms this is

to be expected with sounds utilizing ingresssive air stream mechanisms; a backward (and downward) movement of the anterior closure (due to lower pressure within the points of closure) creates friction up to the point where the seal is finally broken with enough energy to result in a pulse-like excitation.

L&T's interpretation of the waveforms (Fig 1), namely that the turbulent noise occurs **after the release** cannot be supported. The first deviations from the zero-line does not necessarily imply a "release"; these deviations are indeed representative of frication prior to the (full) release attested in the following pulse. An exact definition of the notion "release" would clarify L&T's position considerably. In the case of the other "noisy" click, i.e. lateral [ll], there is likewise no evidence of an impulse excitation followed by turbulence to justify affrication in the classical sense of the word. What is present is high intensity frication over the whole spectrum with some suggestion of a weak impulse-like excitation indicating the final breaking of the seal. Pre-affrication is not clearly attested in these forms, however, they are characterized by high intensity frication throughout the articulation.

The frication and, or preaffrication of the "noisy" clicks are clearly distinguishable from the transient impulse-like click "q" [!]:

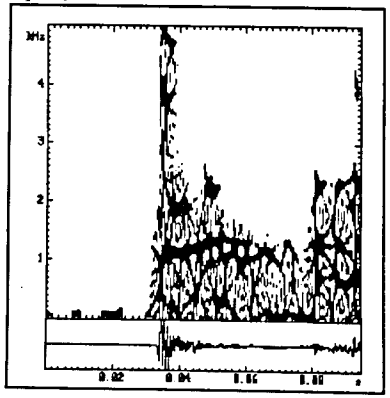


Figure 2. SPWD analysis of the noise burst of a voiceless alveo-palatal click "q" in Xhosa

Nasal Clicks: L&T [2:46] correctly refer to these clicks as "nasal clicks" and not as "nasalized clicks" (cf. [7]). These clicks contrast lexically with other click forms and are not the product of a specific phonological process. In viewing the structure of the particular waveform, L&T [2:47] conclude "The clicks (...) occur almost at the end of an accompanying nasal." A detailed SPWD-analysis of the articulation of these sounds, however, reveal rather interesting features not normally detectable in a waveform presentation alone.

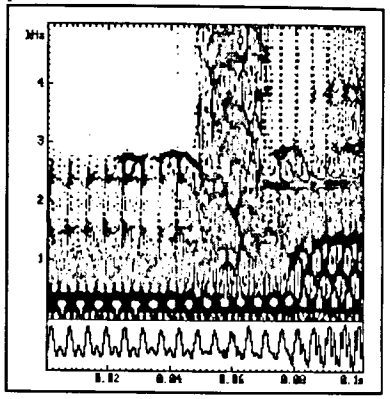


Figure 3. SPWD analysis of the noise burst of a nasal dental click "nc" in Xhosa

The nasal waveform is observable from the beginning of the window up till approximately 0.08s with the click clearly **superimposed** on this wave at a point approximately 0.06s into the signal. This superimposition of a click without significantly changing the structure of the nasal waveform is characteristic with all three click types in our data. In a certain sense it may even be argued that the click is an accompaniment of the nasal and not vice versa. This predominance of the nasal and the superimposition of the click surely shed some light on the process of click acquisition where children initially tend to nasalize all clicks [8] [9], as well as on a phonological process in which a nasal preceding a click surfaces phonetically as a single murmured nasal click [3].

Voiced "murmured" clicks

The clicks represented orthographically as gc, gq, gx are described by L&T [2:46, 47] as "murmured" clicks because they are accompanied by a "murmured velar plosive [g]". This click is transcribed with a voiced velar symbol with a dieresis under it (implicating breathy voice) preceding the click symbol: [g̈!],. However, referring to the waveform they state that "...there is no breathy voice during the closure", and furthermore, that the murmur "... is not accompanied by strong breathy voice during the release of the closure as it is in languages such as Hindi or Marathi." Although they find no breathiness in the signal as such they maintain the use of the dieresis in view of the observation that the murmured nasal belongs to a set of depressor consonants (allegedly exhibiting some degree of breathy voice, cf. [10:477]) that has a lowering effect on the tone of the following vowel. The question then remains what constitutes the perceptual murmur? Catford [11:101] indicates that Bell may have first used this term in 1867 describing it as "whisper and voice heard simultaneously." and that Ladefoged's interpretation assumes "one form of whispery voice." Consider now the following SPWD-analysis which reveals information not visible in the waveform alone:

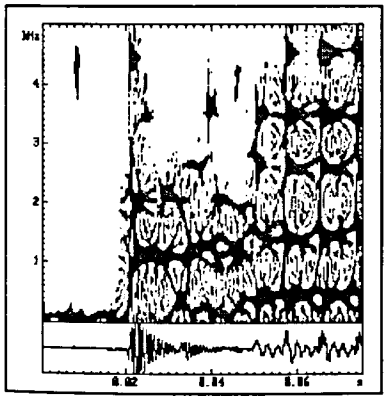


Figure 4. SPWD analysis of the noise burst of a voiced alveo-palatal click "gq" in Xhosa

A distinctive double impulse excitation is attested with the second impulse representing the release of the velar closure. Editing of the signal results in the first impulse clearly being perceived as a click and the second as a clear voiced velar stop. This impulse-like excitation following the click is attested in the other two click types as well. In fast succession, however, a voiced release may be perceived, but certainly no breathy voice. Lowering of pitch in the following vocalic segment need not have anything to do with vocal fold activity simutaneously facilitating breathy voice, but might correlate with the voicing of the short obstruent that follows the click.

REFERENCES

[1] Dogil, G. & Wokurek, W. (1991). Wigner time-frequency representation for major places of articulation in stop consonants. *Proc. 12th Int Conf Ph Sc*, Aix-en-Provence, 390-395.

[2] Ladefoged, P. & Traill, A. (1994). Clicks and their accompaniments. *Journal of Phonetics*, 22, 33-64

[3] Dogil, G. & Roux, J.C. (in prep). Unresolved issues regarding click articulations in Xhosa and Zulu.

[4] Wokurek, W. (1994). Darstellung und Untersuchung von Sprachsignalen mit Wigner-Verteilung und Spektrogramm. *Phonetik-AIMS* 1, 1-133, Stuttgart, Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung, Universität Stuttgart.

[5] Dogil, G. & Jessen, M. (1990) Phonologie in der Nähe der Phonetik. Die Affrikaten im Polnischen und Deutschen. In: Prinzhorn, M. (ed.) *Phonologie*. Westdeutscher Verlag, Opladen, 223-279.

[6] Luschützky, H-C. (1985). *Zur Phonologie der Affrikaten*. MA-thesis, Universität Wien, Institut für Allgemeine Sprachwissenschaft.

[7] Sands, B. (1991). Evidence for click features: acoustic characteristics of Xhosa clicks. *UCLA Working Papers in Phonetics*, 80, 6-37

[8] Lewis, P. W. (1994) *Phonetic and phonological aspects of click acquistion in Xhosa*. MA-thesis, University of Stellenbosch, Stellenbosch.

[9] Roux, J.C. & Lewis, P.W. (in prep.) On phonological acquisition of clicks.

[10] Laver, J. (1994) *Principles of Phonetics*. New York. Cambridge University Press.

[11] Catford, J.C. (1977). *Fundamental Problems in Phonetics*. Edinburgh, University Press.

LABIAL-VELAR STOPS IN DAGBANI

M. Pettorino and A. Giannini
Istituto Universitario Orientale, Fonetica Sperimentale, Napoli, Italia

ABSTRACT

The consonantal system of Dagbani language shows phonological contrasts among labial-velar, labial and velar stops. Labial-velars are described as having three possible articulatory and airstream mechanisms: in fact they are considered either as implosives or as bilabial clicks or as double articulations with two simultaneous closures. The experimental data allow us to exclude that Dagbani labial-velars are characterized by either a glottalic or a velaric airstream mechanism.

INTRODUCTION

Dagbani language belongs to the Gur group of Sudanic languages, exactly to the Noltaic subfamily of the Congo-Kordophanian family. Dagbani is spoken in North-eastern Ghana.

It is well known that in Dagbani, as in many other West African languages, there are voiceless and voiced labial-velar stops, that are orthographically indicated by the digraphs /kp/ and /gb/. They are generally defined as double articulations.

In phonetic literature the main difference between double and secondary articulations is that the latter are associated with a constriction ranking lower than the main articulation and having a lesser degree of stricture than the primary articulation, while the former are characterized by two simultaneous constrictions having the same degree of stricture [1] [2] [3] [4].

Ladefoged [5] considers the double articulations as secondary articulations. He says: "...unlike Pike, we will also consider sound which have two equal articulatory strictures at different places of articulation to consist of a primary articulation which is closer to the glottis and a secondary articulation which is further away from it". According to him all the double articulations involve the action of the lips that he defines "secondary articulators" because they are further from the glottis than the other stricture.

Labial-velars are described as having two simultaneous closures, one at the lips and one at the velum. According to Ladefoged [6], they can have three different airstream mechanisms: the first one, similar to the velaric mechanism used to produce clicks, is characterized by an ingressive airflow due to the decrease of intraoral pressure; the second one, similar to the glottalic mechanism used to produce implosives, is given by a lowering of the whole larynx with the closed glottis; the last possibility consists in a pulmonic egressive airflow. In some languages two or three of these mechanisms can be simultaneously employed.

However, it is useful to underline that, notwithstanding the similarities, the three mechanisms are substantially different. In fact in the velaric mechanism, there is the cooccurrence of two closures, only one of them, the lips, having an articulatory function; in the glottalic mechanism there are three simultaneous closures, two of them, the lips and the velum, having an articulatory function; in the pulmonic egressive mechanism, the two closures have an articulatory function.

The purpose of this research is to clarify, through spectrografic and electro-aerometric analysis, the articulatory mechanism of production of labial-velar stops in Dagbani.

MATERIAL

From a large corpus of about 1400 meaningful words and 6 prose passages, 500 words have been selected in order to have labial, velar and labial-velar stops in initial and intervocalic position. The whole corpus has been read in a silent room by a male speaker, aged 31, from Tamale. The dialect of Tamale has been chosen because nowadays it is considered as the standard pronounciation of Dagbani language.

Each word has been analyzed through the DSP Sonagraph 5500 KAY, using the Pitch Display program that allows to visualize the broad band spectrogram from 0 to 8000 Hz and the f_o tracing.

In order to describe the articulatory mechanism of double articulations, it is useful to compare them with labial and velar stops in intervocalic position.

DISCUSSION

Figure 1 shows the spectrograms of the words "sagbani" "sagani" and "dabari", where the tendency of the second formants of the adjacent vowels is clearly detectable because of the voiced nature of the consonants. Let us firstly consider the shifting in frequency of the second formant of the preceding vowel: it shows a slight rising transition both before [gb] and [g] and a strongly falling transition before [b]. These different trends allow us to say that the labial-velar stop starts with a velar closure. The F2 of the following vowel shows an opposite trend: it starts with a strongly or slightly rising transition after [gb] and [b] respectively and with a strongly falling transition after [g]. This means that the articulatory mechanism of [gb] ends with a labial opening.

The different rate of change of the F2 rising transition reflects a greater shifting of the tongue in [gb] than [b]. In the former, at the moment of the labial release the back of the tongue is still raised, so that it needs more time to reach the vowel configuration. On the contrary, in the latter the tongue has already reached the vowel position at the moment of the release. This is why, when [gb] is followed by a back vowel, the difference in rate of F2 transition is neutralized.

One more difference concerns the duration, being [gb] about 30% longer than [g] and [b].

All these remarks can be made also for the voiceless labial-velar stops.

The spectrographic data offer clear

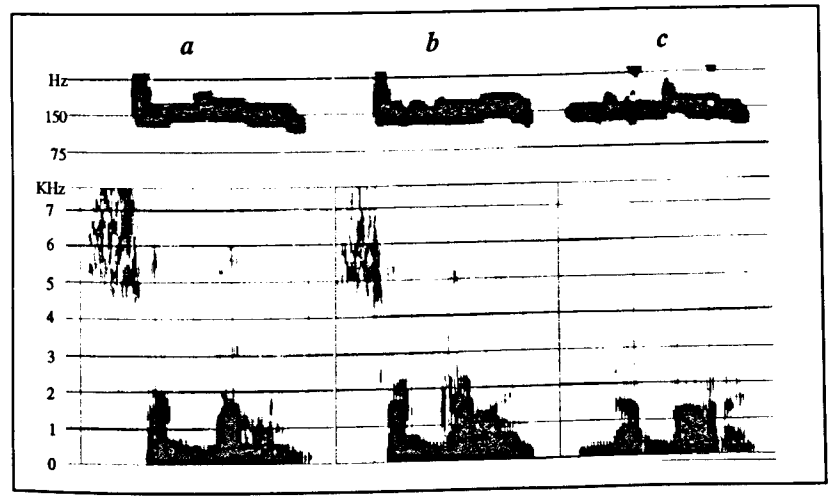


Figure 1. Broad band spectrograms and pitch display of the words "sagbani" (a), "sagani" (b), "dabari" (c).

evidence of the behaviour of the glottis, both for voiced and voiceless stops. First of all we have to say that when a voiced labial-velar stop occurs in intervocalic position, there is an uninterrupted periodical signal from vowel to vowel, through the consonant. An example of this is given by the spectrogram (*a*) of figure 1: in this case the fundamental frequency is at about 150 Hz and it is very similar to the f_o tracings of the spectrograms (*b*) and (*c*), which are produced with a pulmonic mechanism. So, we can say that also labial-velars are realized with an egressive steady flow through the glottis. The electro-areometric tracings confirm that our informant has never produced them with an ingressive airstream.

As regards the voiceless stops the spectrographic analysis points out a noticeable difference in VOT between labial-velar and labial and velar consonants. Figure 2 shows the broad band spectrograms of the words "kpali" "karili" and "pali", where [kp] [k] and [p] are in initial position followed by the stressed vowel [a].

As we can see, [k] and [p] are characterized by an aperiodic segment of about 80 ms occurring between the release and the following vowel. This turbulent airstream is generally ascribed to a delay of the glottal closing in relation to the supraglottal events: if at the moment of the release the glottis is still wide open, the stop will be aspirated, eitherwise it will be unaspirated [6] [7]. In other words, this means that in [kp] at the moment of the labial release the vocal folds are already close together and they can immediately start vibrating. Therefore, in the presence of two asyncronous releases, the glottis synchronizes itself with the velar opening rather than with the labial one. This can be explained by considering that the musculature of the glottis is more strictly related to that of the tongue rather than to that of the lips. On this subject it is enough to think about the phenomenon of Intrinsic Pitch of vowels. In fact, even though many different hypotheses have been formulated to explain this phenomenon, nevertheless in all of them the direct relationship between tongue and glottis is admitted.

CONCLUSIONS

In conclusion, the data gathered in this experimental research point out that the labial-velars in Dagbani are produced with two asyncronous closures and openings and with an egressive airstream mechanism. The diagram of figure 3 illustrates the temporal articulatory sequence of labial-velars: at the instant 1

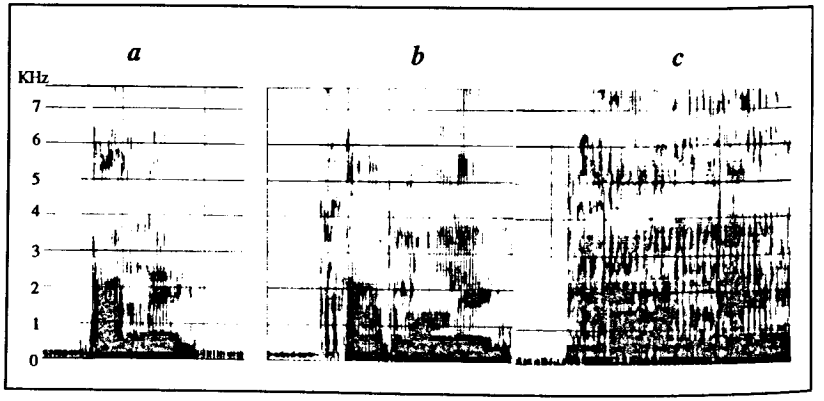


Figure 2. Broad band spectrograms of the words "kpali" (a), "karili" (b), "pali (c).

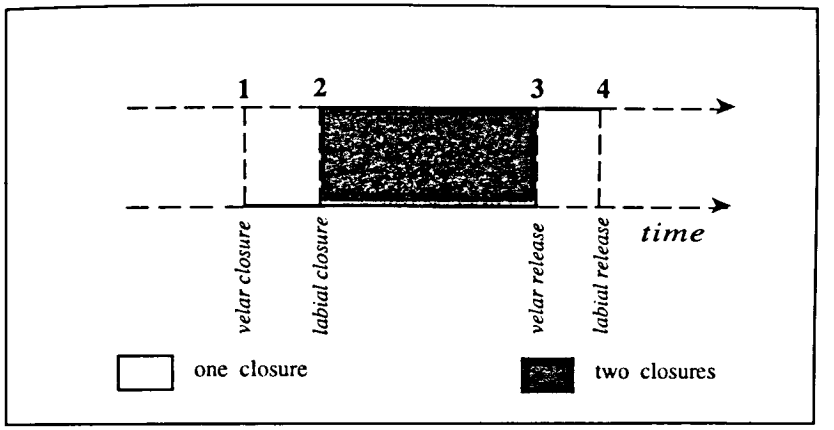


Figure 3. Temporal sequence of articulatory mechanism of labial-velars.

the velar closure is realized, at the instant 2 the labial closure is realized, from 2 to 3 there is cooccurrence of the two closures, at the instant 3 the velar closure is released, at the instant 4 the labial closure is released. In relation to this we can exclude that labial-velars are realized as bilabial clicks. In fact, even though labial-velars and clicks are produced with cooccurrent labial and velar closures, the former release the velum first, the latter the lips first.

We can also exclude that the labial-velar stops are realized as implosives, as the glottalic mechanism would cause either an interruption or a drastic drop of the vocal fold vibrations, because of the compensation between subglottic and supraglottic pressure.

REFERENCES

[1] Pike, K. L. (1947), *Phonemics*, Ann Arbor, Univ. Michigan Press.

[2] Abercrombie, D. (1967), *Elements of general phonetics*, Edinburgh Univ. Press.

[3] Westermann, D. and Ward, I. C. (1970), *Pratical phonetics for students of African languages*, Oxford Univ. Press.

[4] Catford, J. C. (1977), *Fundamental problems in phonetics*, Edinburgh Univ. Press.

[5] Ladefoged, P. (1974), *Preliminaries to linguistic phonetics*, The Univ. Chicago Press, p. 59.

[6]Kim, C. W. (1970), A theory on aspiration, *Phonetica*, vol. 21, pp. 107-116.

[7]Pettorino, M. and Giannini A. (1984), A study on aspiration, *Speech Laboratory Report*,vol. IV, I.U.O., Napoli.

AN EMMA INVESTIGATION OF LINGUAL ASSIMILATION AND COARTICULATION IN A SELECTED SET OF CATALAN CONSONANT CLUSTERS

Daniel Recasens
Universitat Autònoma de Barcelona and Institut d'Estudis Catalans, Barcelona, Spain

ABSTRACT

This paper investigates lingual activity for clusters consisting of consonants specified for adjacent places of articulation. Data on assimilatory and coarticulatory processes are interpreted in terms of the articulatory requirements involved in consonantal production.

INTRODUCTION

The investigation of sequences of consonants produced with the same articulator or with adjacent articulators is justified by the diversity of processes affecting their realization. This paper analyzes movement data for Catalan clusters with lingual consonants produced at adjacent articulatory zones, i.e., dental /t̪/ and /d̪/, alveolar /n/, /l/, /ɾ/ (a tap) and /z/, postalveolar /ʒ/ and alveolopalatal /ʎ/.

(a) Clusters /nd̪/ and /ld̪/.

In the clusters /nd̪/ and /ld̪/, alveolar C1 undergoes place assimilation thus becoming dental. Moreover C2=/d̪/ is realized as a stop (as opposed to an approximant as in the intervocalic context) in line with C1 exhibiting a lingual closure at the same (dental) location [1].

(b) Clusters /ɾd̪/, /zd̪/ and /ʒd̪/.

This paper investigates whether the tongue tip for C2=/d̪/ reaches the dental zone, remains at the alveolar zone (as for C1) or is found at an intermediate location. According to lingual movement data for Spanish [4] there is enough time for the tongue tip to reach the teeth for C2 in the cluster /ɾd̪/ because the tap C1 is produced with a fast apical movement; on the other hand, high lingual requirements for the fricative /z/ in the cluster /zd̪/ prevent the tongue tip from achieving a dental constriction for C2.

C2=/d̪/ is realized as an approximant in the three Catalan clusters /ɾd̪/, /zd̪/ and /ʒd̪/ in agreement with C1 not exhibiting a lingual closure at the same place of articulation as C2 [1].

(c) Cluster /ʎd̪/.

Catalan /ʎ/ is an alveolopalatal lateral consonant. The fact that C2=/d̪/ is realized as a stop in the cluster /ʎd̪/ implies that, analogously to /nd̪/ and /ld̪/, the two adjacent consonants share a lingual closure at the same articulatory location [1]. This paper investigates this issue as well.

METHODOLOGY

Tongue tip (TT), tongue blade (TL) and tongue dorsum (TD) movement data were collected for one Catalan speaker (the author) using an electromagnetic midsagittal articulometer (EMMA) [2]. Acoustic data were also recorded. The speaker read ten times a list of nonsense symmetrical sequences /pa'Cap/ and /paC'Cap/ embedded in the Catalan phrase 'jo guixo___ si vols' ("I chalk ___ if you want"). VCV sequences included the single consonants /t̪/, /d̪/, /n/, /l/, /ɾ/, /z/, /ʒ/ and /ʎ/. Clusters were composed of each non dental consonant followed by C2=/d̪/: /nd̪/, /ld̪/, /ɾd̪/, /zd̪/, /ʒd̪/, /ʎd̪/. The selection of stop /t̪/ and approximant /d̪/ in intervocalic position allows characterizing fine differences in closure and constriction location for C2 in the six clusters above.

A helmet with three magnetic transmitters was mounted on the head of the speaker and small transducer coils were attached to the three lingual articulators in the midsagittal plane. Coils were also placed on the bridge of the nose and on the upper incisors for head movement correction. The magnetic fields from the transmitters induce voltages in the transducers which can be converted to distance with the appropriate software. Data were digitized at sampling

rates of 625 Hz for movement and 10 kHz for speech. The movement data were rotated with respect to the occlusal plane, corrected for head movement, and extracted separately for the X (horizontal) and Y (vertical) dimensions in conjunction with derived velocities. Articulatory trajectories were labelled TTX, TTY, TLX, TLY, TDX and TDY.

A routine allowed calculating zero crossings in the velocity traces. X and Y position maxima during the consonantal period were identified at velocity minima for each articulatory trajectory; when velocity minima were not available, representative displacement points were measured at the midpoint of the acoustic waveform event associated with a single consonant or with C1 and C2 in consonant clusters. Inspection of composite X-Y temporal trajectories were used to identify one position maximum for single consonants and for the clusters /nd̪/, /ld̪/ and /ʎd̪/, and two maxima for the clusters /ɾd̪/, /zd̪/ and /ʒd̪/ (i.e., a more retracted one for C1 and a more fronted one for C2).

Statistical analyses (ANOVAs with repeated measures and Fisher post hoc tests) were performed in order to find out whether single consonants and clusters could be differentiated on the basis of X and Y position maxima. Significant effects were established at the $p<.05$ level of significance.

RESULTS

Regressive assimilation

According to Figure 1, the clusters /nd̪/ and /ld̪/ are articulated at the dental zone as shown by their TTX maximum being as front as that for /t̪/ and /d̪/. This finding confirms that C1 assimilates to C2 in place of articulation. TLX and TDX maxima also occur at a fronter position for dentals and clusters than for alveolars. TTY maxima for clusters are in agreement with C2 being a stop since they are as high as for /t̪/ (at the upper incisors) and much higher than for the approximant /d̪/ (at the lower incisors).

C1 exerts some effects on C2. There are indeed some differences in lingual configuration between /nd̪/ and /ld̪/ with the latter cluster exhibiting a significantly more posterior TTX, TLX and TDX maxima and lower TLY and TDY maxima; this finding accords with Catalan velarized /l/ involving active tongue predorsum lowering. Also, in comparison to /t̪/, /nd̪/ shows more anterior TLX and TDX maxima, a higher TLY maximum and a lower TDY maximum. All these differences are significant and cannot be assigned to differences in tongue blade position between single /n/ and a dental stop (notice that single /n/ occupies a backer and lower TL maximum than single /t̪/). Instead it appears that the two stops reinforce each other in the cluster giving rise to an increase in closure degree.

Progressive coarticulation

Clusters /ɾd̪/, /zd̪/ and /ʒd̪/ show two TT, TL and TD locations connected by lines in Figure 2, i.e., a more retracted one for C1 and a fronter one for C2. Data will be reported separately for C2 of /ɾd̪/ (a), for C2 of /zd̪/ and /ʒd̪/ (b), and for C1 of the three clusters (c).

(a) Data reveal a significantly fronter and lower TT maximum for C2=/d̪/ in the cluster /ɾd̪/ than in the clusters /zd̪/ and /ʒd̪/. The absence of significant differences in TTX maxima between C2=/d̪/ of /ɾd̪/ and single /t̪/ and /d̪/ indicates that the former realization is truly dental. A significantly lower TTY maximum for C2=/d̪/ of /ɾd̪/ than for dental /t̪/ is consistent with the former consonant being realized as an approximant; moreover, the fact that this position is significantly higher than that for single /d̪/ suggests that C1=/ɾ/ affects tongue tip location for C2=/d̪/ (notice that the TTY maximum for single /ɾ/ is also higher than that for single /d̪/).

Differences between /t̪/, /d̪/ and C2=/d̪/ of the cluster /ɾd̪/ at TL and TD are similar to those found at TT.

(b) TTX and TTY maxima for C2=/d̪/ in the clusters /zd̪/ and /ʒd̪/ are not significantly different. TTX maxima are significantly more posterior than those for single dentals (/t̪/, /d̪/) and for C2=/d̪/ of the cluster /ɾd̪/. In spite of being an approximant, the TTY maximum for C2=/d̪/ of /zd̪/ and /ʒd̪/ is significantly higher than that for single /d̪/ and for C2=/d̪/of /ɾd̪/ and as high as that for /t̪/; this accords with single /z/ and /ʒ/

exhibiting a higher TTY maximum than single /ɾ/. It can thus be concluded that C1 fricative prevents C2 from achieving a dental place of articulation and that the precise constriction place for C2=/d/ is related to the place of articulation for fricative C1.

Analogously to the TT data, C2=/d/ of /zd/ and /ʒd/ shows a highly similar laminodorsal position. TL and TD values for these clusters are significantly more posterior than those for /t/, single /d/ and C2 of /ɾd/, and higher than those for /t/ which should be attributed to /z/ and /ʒ/ occupying a higher laminodorsal position as well.

(c) There are some interesting C2-to-C1 effects in the clusters /ɾd/, /zd/ and /ʒd/.

According to Figure 2, the TT maximum for C1 in clusters does not coincide exactly with that for the same consonant in the isolated condition: in comparison with the latter, the former can be significantly more retracted (/ʒ/), more anterior (/z/) and higher and more anterior (/ɾ/). Moreover, the consonants /z/ and /ɾ/ show highly similar TL and TD maxima when produced in isolation and before /d/ (the only significant difference affects TLY maximum for /ɾ/ in the cluster /ɾd/ as opposed to single /ɾ/); however, the laminodorsal position for /ʒ/ in the cluster /ʒd/ is significantly backer and lower than that for single /ʒ/.

Cluster /ʎd/

Figure 2 indicates that the cluster /ʎd/ exhibits the same TTX and TTY maxima as single /ʎ/; in comparison to /t/, these TT maxima are fronter but neither higher or lower. The presence of a dental location for single /ʎ/ is presumably a secondary articulatory attribute occurring when the alveolopalatal closure extends over the entire alveolar zone [3]. In the cluster /ʎd/, however, it may very well be the output of a regressive assimilatory process through which C1 and C2 become homorganic. The existence of C2-to-C1 effects in laminodorsal position support this interpretation: in comparison with /ʎ/, TL and TD maxima for /ʎd/ are significantly lower and non-significantly more fronted than those for single /ʎ/.

CONCLUSIONS

Data reported in this paper confirm that C1 assimilates in place to C2 in the clusters /nd/ and /ld/. C1 preserves its place of articulation in the clusters /ɾd/, /zd/ and /ʒd/. The tongue tip for C2=/d/ reaches the dental zone after the tap /ɾ/ but not so after a lingual fricative in agreement with the articulatory requirements for the production of C1 (/ɾ/ is produced with a rapid tongue tip movement; /z/ and /ʒ/ require a highly precise tongue body positioning). While exhibiting dental contact in both contextual conditions, /ʎ/ is primarily alveolopalatal in intervocalic position and undergoes presumably regressive place assimilation in the cluster /ʎd/. Consonants which are resistant to assimilatory processes (/d/ in /nd/ and /ld/) and to coarticulatory phenomena (/z/ in /zd/, /ʒ/ in /ʒd/) are also affected by the adjacent consonant in the cluster.

REFERENCES

[1] Mascaró, J. (1991), "Iberian spirantization and continuant spreading", *Catalan Working Papers in Linguistics*, vol. 1, 167-180.

[2] Perkell, J., Cohen, M., Svirsky, M., Matthies, M., Garabieta, I. and Jackson, M. (1992), "Electro-magnetic midsagittal articulometer systems for transducting speech articulatory movements", *Journal of the Acoustical Society of America*, vol. 92, 3078-3096.

[3] Recasens, D., Farnetani, E., Fontdevila, J. and Pallarès, M.D. (1993), "An electropalatographic study of alveolar and palatal consonants in Catalan and Italian", *Language and Speech*, vol. 36, 241-262.

[4] Romero, J. (unpublished), "Articulatory blending of lingual gestures".

ACKNOWLEDGMENTS

This research was supported by NINCDS Grant A-64 to Haskins Laboratories, and by projects ESPRIT BRA 6975 and 7098 (EC) and DGICYT CE93-0020 (Spanish Government). I thank D. Whalen and P. Hoole for their comments.

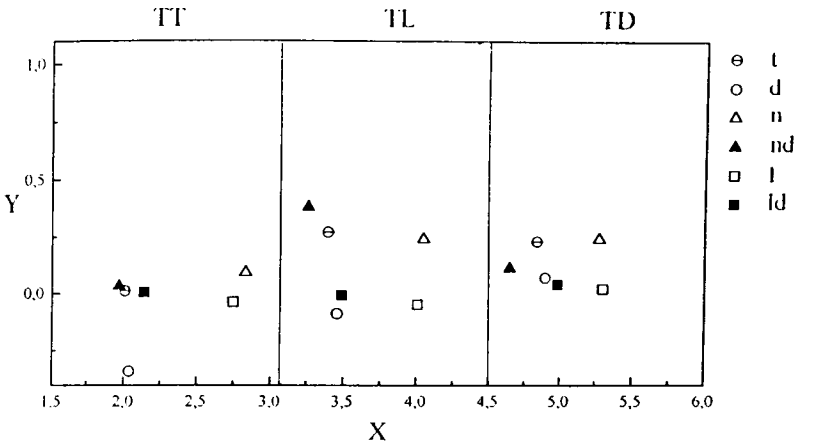


Figure 1. TT, TL and TD position maxima (in cm) along the vertical (Y) and horizontal (X) dimensions for single /t/, d/, /n/ and /l/ and for the clusters /nd/ and /ld/.

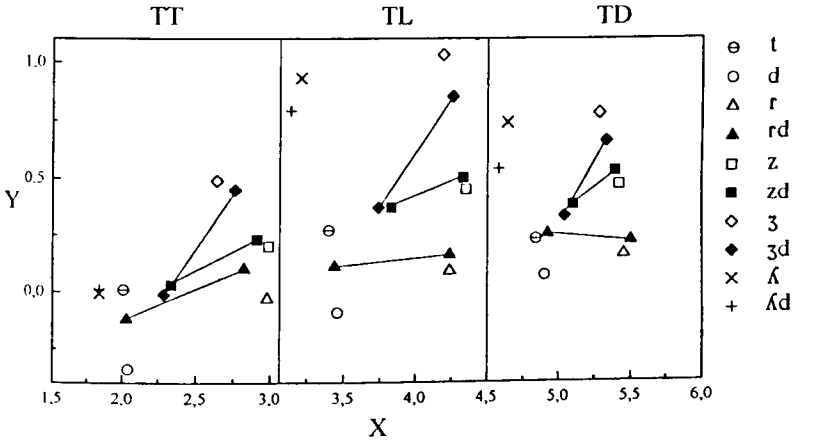


Figure 2. TT, TL and TD position maxima (in cm) along the vertical (Y) and horizontal (X) dimensions for single /t/, d/, /ɾ/, /z/, /ʒ/ and /ʎ/ and for the clusters /ɾd/, /zd/, /ʒd/ and /ʎd/. The clusters /ɾd/, /zd/ and /ʒd/ show two maxima (C1 maximum on the left, C2 maximum on the right) which have been connected by a line.

PHARYNGEAL AND UVULAR CONSONANTS ARE APPROXIMANTS: AN ACOUSTIC MODELING STUDY

*M. YEOU and S. MAEDA **
Institut de Phonétique, Sorbonne Nouvelle. CNRS, Paris.
** Ecole Nationale Supérieure de Télécommunications. CNRS, Paris.*

ABSTRACT

Idealized models based on realistic area functions are proposed for uvular /ʁ, χ/ and pharyngeal /ʕ, ħ/ of Arabic. Synthesis of speech from these idealized models was also obtained for [aCa] sequences. Findings of this study indicate that these consonants should not be considered fricatives but approximants. First, values of Ac and Ag were estimated to be higher than those for simple fricatives. Second, corresponding spectrograms usually show a vowel-like formant structure. Third, calculated and measured airflow values were outside the range for normal fricatives.

INTRODUCTION

Theoretical modeling of pharyngeal /ʕ,ħ/ was first given in [1], where formant-cavity affiliations were examined. In [2], idealized models were proposed for both pharyngeal /ʕ,ħ/ and uvular /ʁ, χ/. In this study, we propose similar models which are based on realistic area functions derived from x-ray profiles corresponding to these consonants. The x-ray profiles utilized came from [3].

1. IDEALIZED MODELS

1.1. Realistic area functions

In the process of deriving the realistic area functions, the method used for determining the sagittal distances consists of fitting circles inside the vocal tract and then defining a midline as the locus of their centers [4, 5] (see Figure 1).

Two vocal tract profiles, one for the pharyngeal /ʕ/ and the other for the uvular /χ/, were enlarged 400 % to facilitate the insertion of circles. The diameter of each circle, and the length of the segments joining these circles were measured. The area functions were derived from sagittal distances by the application of the relation $A=\alpha.d\beta$, where d is the sagittal distance, and α and β are changing coefficients in function of different regions in the vocal tract. Figure 2 gives the derived area functions corresponding to pharyngeal /ʕ/ and to uvular /ʁ/ (for details see [6]).

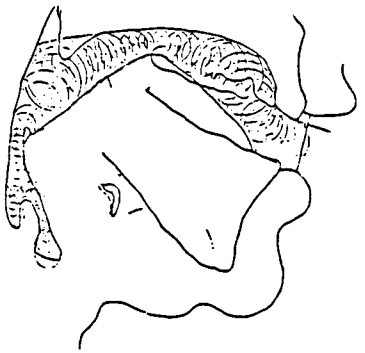


Figure 1. The 'fitting-circle' method used in deriving realistic area functions.

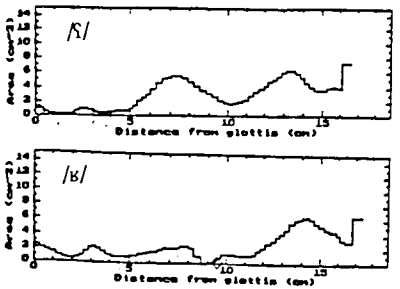


Figure 2. Area functions derived from x-ray profiles corresponding to pharyngeal /ʕ/ and to uvular /ʁ/.

The dimensions of these area functions conform to the articulatory descriptions given in [3]. For example, the vocal tract length for uvular /ʁ/ (17.2 cm) is longer than that for pharyngeal /ʕ/ (16.5 cm). This due to the elevation of larynx during the production of the latter.

1.2. Idealized area functions

Idealized models for the production of pharyngeal and uvular consonants are proposed on the basis of the dimensions provided by the realistic area functions (see Figure 3). The model consists of three uniform tubes corresponding to the back cavity, the constriction and the front cavity.

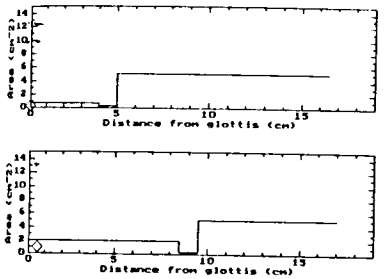


Figure 3. Idealized area functions for the production of pharyngeals (upper), and uvulars (bottom).

2. TRANSFER FUNCTIONS AND FORMANT FREQUENCIES

Many transfer functions corresponding to the idealized area functions were calculated by a simulation method of the vocal tract which includes radiation characteristics, boundary losses and the subglottic system [7]. The variable parameters changed were principally glottal area (Ag) and constriction area (Ac). The aim is to find a good correspondance between calculated and measured formant frequency values. Table 1 gives the calculated formant frequencies and airflow. It can be seen that Ag distinguish voiced consonants (Ag=0 cm^2) from their voiceless counterparts (Ag=0.20-0.25 cm^2). Moreover, the voiceless have narrower Ac than the voiced. The values for Ac and Ag given in Table 1 are found to be bigger than those appropriate for fricatives.

Table 1: Calculated formant frequencies and airflow (U) from the idealized area funtions.

	ʕ	ħ	ʁ	χ
F1 (Hz)	689	784	570	------
F2 (Hz)	1493	2040	1206	1328
F3 (Hz)	2181	2648	2304	------
F4 (Hz)	3525	3496	3157	?
Ac (cm^2)	0.35	0.30	0.35	0.20
Ag (cm^2)	0	0.25	0	0.20
U (cm^3/s)	------	596	------	439

The calculated transfer functions (Figure 4) show that only the formants associated with the front cavity are excited when the glottis is open (F2 and F4 for uvular /χ/; F1, F3 and F4 for pharyngeal /ħ/). It can also be seen that the F2 for /ħ/, though associated with the Helmoltz resonance involving the back cavity, appears to be excited. This can be explained by the fact the noise source seems to be located at the glottis for this consonant. This hypothesis is possible since Ac is larger than Ag (0.30 cm^2 vs. 0.25 cm^2), and a formant structure is well apparent in its sonogram.

3. ACOUSTIC ANALYSIS

An acoustic analysis was conducted in parallel with the acoustic modeling. Nonsense /CVV/ syllables (where /C/ is /χ,ʁ,ħ,ʕ/ and /VV/ is the long vowel /a:/) were produced in carrier phrases by 4 male Moroccan speakers. Table 2 gives the mean values for the formant frequencies taken in the middle of the consonant. Comparing these measured values (Table2) with those calculated from idealized area functions (Table1), we find a good correspondance between the two.

Table 2. Formant frequencies averaged across 4 speakers and 5 repetitions for uvular and pharyngeal consonants.

	ʕ	ħ	ʁ	χ
F1 (Hz)	710	777	616	------
F2 (Hz)	1494	1978	1252	1389
F3 (Hz)	2255	2536	2321	------
F4 (Hz)	?	3597	?	?

4. SYNTHESIS OF SPEECH

Synthesis of speech from these idealized area functions was obtained for [aCa] sequences, where C = /ʁ, χ, ʕ, ħ/. In the simulation method used [7], the noise generation is achieved by placing a pressure source along the vocal tract. Figure 6 illustrates the evolution of the parameters used in the synthesis of [aʁa]. Are shown in this figure the calculated speech waveform (signal), the glottal area (Ag), the constriction area (Ac), the fundamental frequency (F0), and the area function (AF). In an informal listening test involving 4 Moroccan listeners, the quality of the synthesized sequences was excellent in terms of both intelligibility and naturalness. Figure 7 shows the spectrograms of the synthesized sequences [aχa] and [aʕa].

5. FRICATIVES OR APPROXIMANTS

The modeling and the acoustic studies give us many indications pointing that pharyngeal and uvular consonants should not be considered as fricatives but as approximants. One definition of the difference between the former and the latter is given in [8] (cf. a phonological distinction [9] based on it). A fricative has a narrower area of constriction (Ac= 0.03-0.20 cm^2) and an airflow that is turbulent whether it is voiced or voiceless, while an approximant has a wider Ac (0.2-0.8 cm^2) and an airflow that is turbulent only when voiceless. According to this definition pharyngeal /ħ,ʕ/ and uvular /χ,ʁ/ should be considered as approximants, since their area of constriction appropriate for their modeling is in the order of 0.20-0.35 cm^2. Moreover, the spectrograms (Figure 5) corresponding to voiced /ʕ,ʁ/ indicate that the airflow is non-turbulent: absence of friction noise and presence of a vowel-like formant structure.

There is further evidence from aerodynamic data in favor of the approximant categorization. The measured airflow values for these consonants are outside the typical range for fricatives [3, 10]. Our preliminary airflow data for these consonants are in accordance with this finding. Figure 8 shows that the airflow of /χ, ħ/ is significantly higher than that of /s/. Furthermore, the airflow shape for /χ, ħ/ is single-peaked while for /s/ is double-peaked. Single-peaked airflow is characteristic of approximant consonants [11] since their supraglottal constriction area is larger than their glottal area.

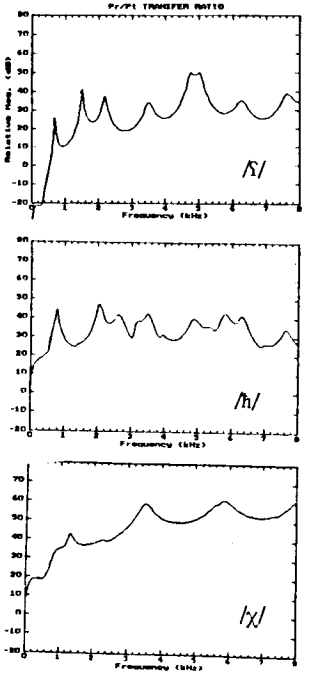
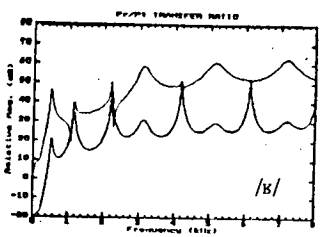


Figure 4. Transfer functions calculated from the idealized area functions. The source for /ʁ/ is both at the glottis and at 1 cm after the constriction (hence two transfer functions). For /ʕ/ and /ħ/ the source is at the glottis only, and for /χ/ it is placed at 1 cm after the constriction.



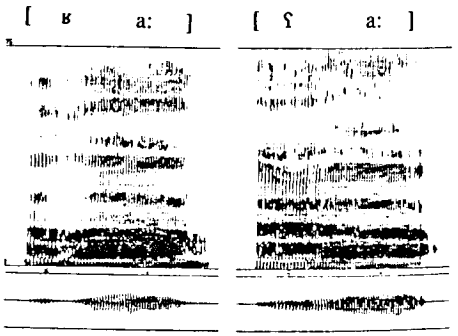


Figure 5. Spectrograms corresponding to voiced pharyngeal /ʕ/ (right) and voiced uvular /ʁ/ (left) in the context of the long vowel /a:/.

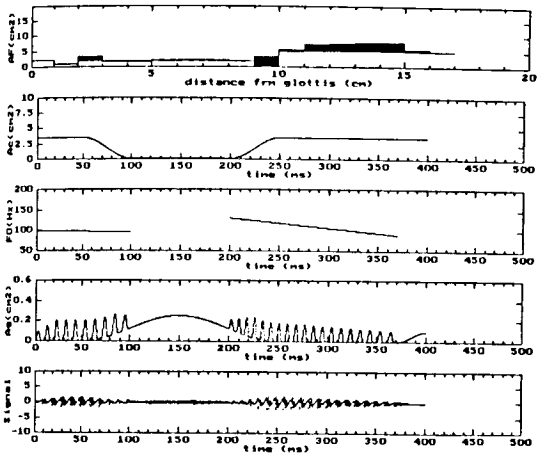


Figure 6. An illustration of some parameters used in the synthesis of [aχa]: the area function (AF), the supraglottal constriction (Ac), the fundamental (F0), the glottal area (Ag) and the speech waveform (signal).

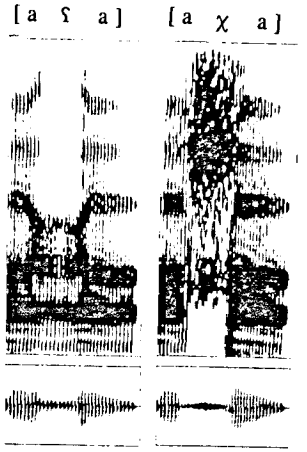


Figure 7. Spectrograms of the synthesized sequences: [aʕa] (left), and [aχa] (right).

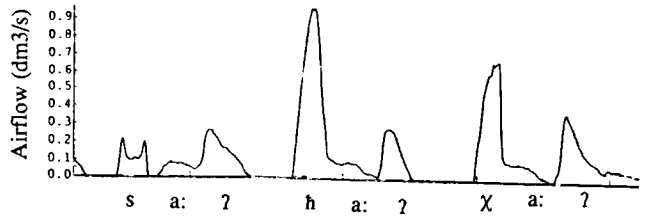


Figure 8. Airflow data (in dm^3/s) for [sa:ʔ], [ħa:ʔ] and [χa:ʔ] from one individual speaker.

BIBLIOGRAPHY

[1] Stevens, K.N. & Klatt, D.H. (1969) "Pharyngeal consonants," *Research Laboratory of Electronics MIT Quarterly Progress Report* 93: 208-216.

[2] Alwan, A. (1986) *Acoustic and Perceptual Correlates of Pharyngeal and Uvular Consonants*. SM thesis. MIT, Cambridge MA.

[3] Ghazeli, S. (1977) *Back Consonants and Backing Coarticulation in Arabic*. Ph.D. Thesis, University of Texas.

[4] Maeda, S. (1972) "On the conversion of vocal tract X-ray data into formant frequencies," Murray Hill, Bell Laboratories.

[5] Miller, I. & Fujimura, O. (1975) "From tongue model data to sound," *JASA* 57, Suppl. 1.S3.

[6] Yeou, M. & Maeda, S. (1994) "Pharyngales et uvulaires arabes sont des approximantes: Caractérisation acoustique," *XX^è Journées d'Etudes sur la Parole*. pp. 409-414. Trégastel.

[7] Maeda, S. (1982) "A digital simulation method of the vocal tract system," *Speech Communication* 1: 199-229.

[8] Catford, C. (1977) *Fundamental Problems in Phonetics*. Indian University Press, Bloomington, IN.

[9] Clements, G. N. (1990) "The role of the sonority hierarchy in core syllabification," *Papers in Laboratory Phonology 1*. C.U.P. 283-333.

[10] Butcher, A. & Ahmad, K. (1987) "Some acoustic and aerodynamic characteristics of pharyngeal consonants in Iraqi arabic," *Phonetica* 44: 156-172.

[11] Klatt, D.H., Stevens, K.N. & J. Mead (1966) "Studies of articulatory activity and airflow during speech," *Ann. N.Y. Acad. Sci.* 155: 42-55.

THE INFLUENCE OF SLOWED SPEECH RATE ON COARTICULATION: ACOUSTIC ANALYSIS OF DURATIONAL AND SPECTRAL PARAMETERS

Ingo Hertrich and Hermann Ackermann
University of Tübingen, Germany

ABSTRACT

Durational and spectral measures of coarticulation were obtained from six young female subjects. Anticipatory and retentive coarticulation differed in their acoustic patterns. Both effects showed considerable inter-subject variability. Slowing of speaking rate resulted in a decrease of retentive coarticulation only. The data corroborate the suggestion of different mechanisms underlying anticipatory and retentive coarticulation.

INTRODUCTION

During running speech neighbouring phonetic segments affect each other in various ways, which are usually referred to as coarticulation [1] [2]. Anticipatory as well as retentive coarticulation have been observed [2] [3] [4] [5]. There is some evidence that accelerated speech gives rise to increased coarticulation [6] [7] [8]. This influence of fast speaking rate on coarticulation has been explained by an increased overlap of the temporal domains or activation fields of adjacent articulatory gestures [9] [10] [11]. In analogy, reduced coarticulatory effects may be expected in slow speech.

In order to assess the influence of slowed speech rate upon anticipatory and retentive coarticulation, the acoustic signal of German sentence utterances was analyzed with respect to durational and spectral parameters. Most available acoustic studies on coarticulation considered the formants of the speech signal. However, formant analysis may be compromised by inherent shortcomings. First, it relies on the assumption that the frequency values of a limited number of distinct spectral peaks convey the relevant information. Second, formant extraction algorithms may split a single formant into two or combine two formants to a single one giving rise to incorrect assignments. In order to avoid these problems the present study relied on averaged FFT spectra without any correction or decision making algorithms such as formant tracking.

METHODS

Subjects and material

Six young females participated in the present study representing a rather homogeneous group with respect to age, sex, and education.

Three test sentences comprising the nonsense target word "getVte" (V = {a | i | u}) embedded into the German carrier phrase "Ich habe .... gelesen" ("I have read ....") were considered for analysis. The vowel /V/ of the target word "getVte" has the most prominent position. Thus coarticulation effects of /V/ on both the preceding and the succeeding /e/ and /t/ segments may be expected.

Procedure

Each of the three target words was printed on a card in bold letters. The subjects had to produce the respective test sentences ten times each at comfortable speech tempo (normal speech rate), five times with the instruction to speak somewhat slower (slow-condition) and five times at an even slower rate (extra slow-condition).

Acoustic processing

All recorded sencentes were digitized at a sampling rate of 20 kHz (after anti-aliasing filtering at 8 kHz). Analysis was performed with the CSL Speech Lab (CSL4300; KAY Elemetrics, USA).

Vowel onsets, vowel ends, and stop consonant bursts of the produced test sentences were marked (Figure 1). The following segment durations were considered for analysis: pre-accent /e/ (=SCHWA1), occlusion of pre-accent /t/ (=OCC1), voice onset time of pre-accent /t/ (=VOT1), target vowel (=V) occlusion of post-target /t/ (=OCC2), voice onset time of post-target /t/ (=VOT2), and post-target /e/ (=SCHWA2).

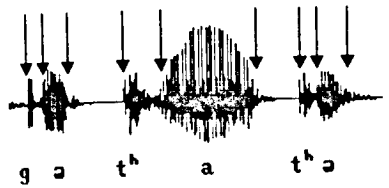


Figure 1. Target word "getate" (Vowel on- and offsets as well as bursts are marked by arrows)

Spectral analysis was performed with the following five segments (bold letters in "Ich habe **g-e-t-V-t-e** gelesen"): pre-target /e/ (=SCHWA1), aspiration of pre-target /t/ (=ASP1), target vowel (=V), aspiration of post-target /t/ (=ASP2), and post-target /e/ (=SCHWA2).

From each of the five segments an averaged FFT spectrum was made (frame length = 6.4 ms; frequency resolution = 156.25 Hz). Altogether 6 (subjects) x 5 (segments per sentence) x 3 (target vowels) x 20 (repetitions: 10 with normal rate, 5 slow, 5 extra slow) = 1800 averaged spectra were computed. The spectra, given in dB values on a linear frequency scale, were normalized for overall intensity by subtracting the mean dB value from all frequency bands.

Statistical analysis

SAS 6.03 software was used for statistical analysis (SAS Institute Inc.; Cary/NC, USA). Multivariate analyses were performed in order to test the differential contribution of the three factors SUBJ {N1-N6}, RATE {normal | slow | extra slow}, and TARG {a | i | u} to the variability of the durational and spectral parameters. The seven segment durations (SCHWA1, OCC1, VOT1, V, OCC2, VOT2, SCHWA2) were converted onto a logarithmic scale in order to compensate for the increase of variances with absolute segment durations [12].

With the spectral data a principal component analysis was performed as a first step in order to reduce the large number of variables. Post hoc analyses of spectral energy distribution were performed in order to find out the frequency bands relevant for the observed multivariate effects. To these ends spectra were averaged across repetitions and plotted together with the corresponding two-standard-error bars. This procedure allowed to identify the spectral regions of interest with respect to coarticulatory patterns.

RESULTS

Segment durations

Seven multivariate tests (the three main effects of TARG, RATE, and SUBJ and the four possible interactions) were performed with the segment durations as the numeric variables. All three factors showed significant effects. As expected, the rate condition revealed to be the dominant source of variation. Subject variability exceeded the influence of coarticulation effects. The two-factor interaction RATE x TARG was not significant when the duration of the target vowel itself was excluded from the numerical data set, indicating that the durational coarticulatory patterns, did not vary across speaking rate conditions.

The RATE factor revealed to be the dominant source of variation with respect to all segments except VOT1, the latter representing a primarily subject-specific measure. The factor TARG had its largest influence upon /V/ reflecting the intrinsic target duration: /a/ had the longest duration, /i/ was slightly shorter than /u/. Significant coarticulatory influences of TARG were observed upon OCC1, VOT1, and SCHWA2. In all subjects the relatively long intrinsic duration of the target vowel /a/ was in part compensated by reduction of one or more of the adjacent segments, predominantly the pre-accent occlusion (OCC1).

In summary, the durational analyses yielded the following results: (1) The three target vowels showed intrinsic durational variability. (2) Partially the coarticulation effects compensated the intrinsic target vowel durations. (3) The durational coarticulation patterns did not significantly interact with the speech rate condition.

Spectral analysis

For reasons of data reduction a principal component analysis was performed with all spectra up to 6 kHz (= 39 spectral bands). Components nos. 1 to 8, accounting for 88% of total variance, were used as dependent numerical variables for multivariate analyses. The categorical variables SUBJ, TARG, and RATE represented the independent factors. A separate MANOVA was performed with each of the five speech segments SCHWA1, ASP1, V, ASP2, and SCHWA2.

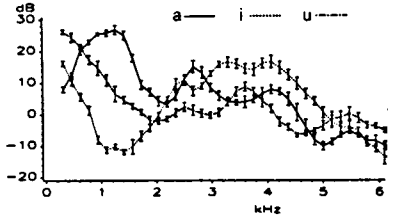


Figure 2. Target vowels /a/, /i/, /u/, (means with 2-standard-error bars across 10 repetitions)

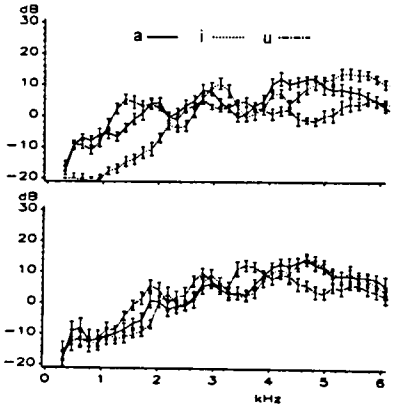


Figure 3. Coarticulation effect upon the pre-target (top) and the post-target (bottom) aspiration (means with 2-standard-error bars across 10 repetitions)

Apart from a few interactions all comparisons yielded significant results. As exemplified in Figure 2, the most significant effect was the (trivial) one of the factor TARG upon the target vowel itself.

The influence of TARG upon the pre-target aspiration had almost the same strength and exceeded the TARG-effect upon ASP2 (Figure 3). Whereas the target vowel category was the dominant source of variation of the aspiration segments, subject variability revealed to be the prevailing effect with respect to the two schwa-sounds. With respect to SCHWA1 some subjects did not show any coarticulatory behavior even under the normal speech rate condition (Figure 4).

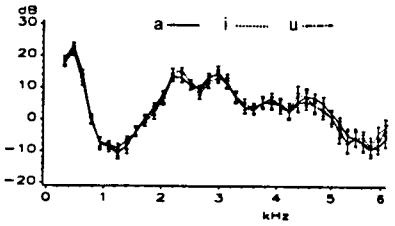


Figure 4. Absence of any coarticulation effect of TARG upon SCHWA1 (means with 2-standard-error bars across 10 repetitions of one subject under the normal rate condition)

The interactions TARG x RATE and TARG x RATE x SUBJ failed significance with respect to SCHWA1. Thus, anticipatory vowel-to-vowel coarticulation did not systematically depend on the speech rate condition. In contrast, SCHWA2 showed reduced coarticulation under the slow rate conditions in all subjects as exemplified in Figure 5.

With respect to all segments considered the interaction between target vowel category and speech rate condition (TARG x RATE) was weaker than the interaction between target vowel category and subject variability (TARG x SUBJ). This indicates that the coarticulatory patterns depend stronger on individual factors than on the speech rate condition.

To summarize the results of the spectral analyses:

(1) The pre-target consonant aspiration (ASP1) showed the strongest coarticulation effect, followed - in descending order - by post-target aspiration (ASP2), post-target schwa (SCHWA2), and pre-target schwa (SCHWA1).

(2) The target vowel category was the dominant source of the observed variability of the aspiration spectra. In contrast, the schwa-sounds primarily were characterized by individual patterns.

(3) Slowing of speech tempo had only minor influences on the aspiration segments. The schwa-segments, in contrast, showed stronger rate effects.

(4) Anticipatory vowel-to-vowel coarticulation did not systematically interact with the speech rate condition.

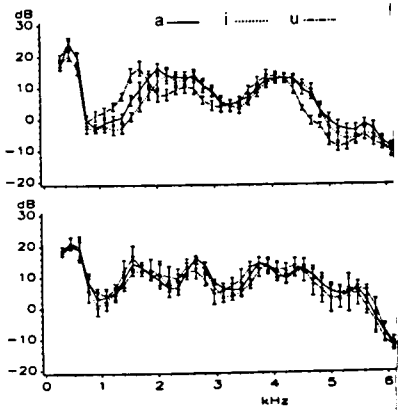


Figure 5. Coarticulation effect upon the post-target /e/ (SCHWA2) under the normal (top) and the extra slow (bottom) speech rate condition (means with 2-standard-error bars across repetitions)

CONCLUSIONS

Durational coarticulatory influences seem to reflect, at least in part, a compensation for intrinsic differences in target length. Spectral analysis showed that retentive coarticulation was weakened in slow speech, whereas anticipatory coarticulation was not. The presence of anticipatory vowel-to-vowel coarticulation was speaker-specific.

REFERENCES

[1] Daniloff, R.G., and Hammarberg, R.E. (1973), On defining coarticulation. *Journal of Phonetics*, vol. 1, pp. 239-248.

[2] Kent, R.D., and Minifie, F.D. (1977), Coarticulation in recent speech production models. *Journal of Phonetics*, vol. 5, pp. 115-133.

[3] Kent, R.D., and Moll, K.L. (1972), Tongue body articulation during vowel and diphthong gestures. *Folia Phoniatrica*, vol. 24, pp. 278-300.

[4] Parush, A., and Ostry, D.J. (1993), Lower pharyngeal wall coarticulation in VCV syllables. *Journal of the Acoustical Society of America*, vol. 94, pp. 715-722.

[5] Repp, B.H. (1986), Some observations on the development of coarticulation. *Journal of the Acoustical Society of America*, vol. 79, pp. 1616-1619.

[6] Gay, T. (1968), Effect of speaking rate on diphthong formant movements. *Journal of the Acoustical Society of America*, vol. 44, pp. 1570-1573.

[7] Gay, T. (1974), Effect of speaking rate on stop consonant-vowel articulation. *Paper presented at the Speech Communication Seminar, Stockholm, Aug. 1-3, 1974.*

[8] Gay, T., Ushijima, T., Hirose, H., and Cooper, F.S. (1974), Effect of speaking rate on labial consonant-vowel articulation. *Journal of Phonetics*, vol. 2, pp. 47-63.

[9] Bell-Berti, F., and Harris, K.S. (1981), A temporal model of speech production. *Phonetica*, vol. 38, pp. 9-20.

[10] Fowler, C.A., and Saltzman, E. (1993), Coordination and coarticulation in speech production. *Language and Speech*, vol. 36, pp. 171-195.

[11] Löfqvist, A., and Yoshioka, H. (1981), Interarticulator programming in obstruent production. *Phonetica*, vol. 38, pp. 21-34.

[12] Crystal, T.H., and House, A.S. (1988), A note on the variability of timing control. *Journal of Speech and Hearing Research*, vol. 31, pp. 497-502.

THE INFLUENCE OF THE SYLLABLE BOUNDARY ON CONSONANT-CONSONANT REALIZATIONS

Thomas Portele
Institut für Kommunikationsforschung und Phonetik, Universität Bonn

ABSTRACT

This paper describes an investigation on the influence of the syllable boundary on consonant-consonant pairs. About 2000 pairs extracted from continuos speech were analyzed. In the acoustic domain it was found that the syllable boundary has some slight influence insofar, as sounds after a syllable boundary tend to be pronounced more clearly. In the perceptive domain we found a corresponding ability to place the syllable boundary. However, both findings are not very marked.

MOTIVATION

The experience with our syllable-based speech synthesis system HADIFIX [6] indicated that the syllable boundary does not serve as a coarticulation blocker; instead we found that our system often generated hypercorrect and over-articulated speech because effects happening a the syllable boundary are not taken into account by just concatenating demisyllables. In the course of the definition of a better inventory structure [7] the question arose whether the syllable boundary has any influence on the phonetic-acoustic domain. Of course, the nature and strength of this influence was what we searched for.

METHOD

Consonant-consonant pairs were placed into 48 specifically designed short texts. All possible combinations occured in different frequencies determinated by the structure of German; the pair /kl/ appeared more often than the pair /tl/. If not prohibited by German phonotactics, the position of the syllable boundary varied between initial (preceding a consonant pair as in "betragen" /bə.tra:gən/) and medial position (between the two consonants as in "mitreißend" /mɪt.raɪsənt/), and between medial and final position (following a consonant pair).The complete corpus contained more than 2000 pairs. It was read by a woman and by a man who were unaware of the study's aim. The pairs were extracted, segmented and grouped according to the way some features were realized. For example, phonologically voiced sounds were labelled *completely devoiced*, *partially devoiced* and *voiced*; stops were tagged *not spoken*, *not released*, *weakly released*, *released*, and *released and aspirated* [2]. Fricatives, nasals and liquids were labelled as well as the degree of assimilation to the other sound of a pair. These classifications were performed only by the investigator. However; a pilot study with 200 plosive-plosive combinations and an additional labeller as well as a re-labelling by the investigator confirmed the reliability of the classification (correlation coefficient ~ 0.9 for both inter-rater and intra-rater reliability). The temporal structure of the consonant-consonant pairs was also investigated. In some cases, spectral features were taken into account. All these results were statistically analyzed regarding the influence of the syllable boundary on the distribution of these parameters.

RESULTS

The results were obtained by comparing the distribution of the labels described above in dependence on the position of the syllable boundary. The analysis yielded experimental results about reduction, assimilation and other phenomena [3], however, only the results concerning differences caused by the syllable boundary are described here.

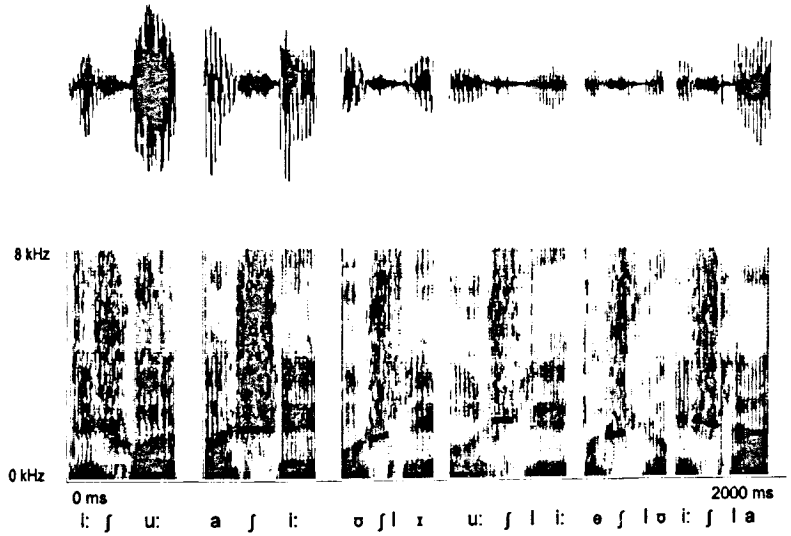


Figure 1. Context-dependent rounding of the /ʃ/ and its acoustic manifestation as a lowered formant-like structure (marked by the fat lines). A rounded /ʃ/ has a formant-like structure by 1450 Hz, an unrounded one by 1900 Hz. From left to right: /i:.ʃʊ/ (transition inside the fricative), /a.ʃi:/ (unrounded), /ʊʃ.lɪ/ (rounded), /u:.ʃli:/ (unrounded), /ɐ.ʃlʊ/ (rounded), and /ɪ.ʃla/ (unrounded). The dot denotes the phonological syllable boundary.

Plosives

The plosive-plosive combinations /pt./ and /kt./ (the dot denotes the position of the phonological syllable boundary) were compared to /p.t/ and /k.t/, respectively. It was found that the first plosive is articulated clearer in syllable final position (U-test, $p < 0.1$). A syllable final /p/ before /t/ shows longer closure duration than /p/ in /pt./.

For the plosive-fricative combinations /kv,pf,ps,tʃ/ significant effects were found only for /kv/, where a syllable initial /v/ is more often voiced or partially devoiced compared to /v/ after a syllable initial /k/ (U-test, $p < 0.025$). There is a tendency for plosives to be released more forcefully in syllable initial position; however, this phenomenon was not significant.

The combinations /.kn/ and /k.n/ differ regarding to the voice onset time of the /k/ which is longer in syllable initial position (U-test, $p < 0.05$).

In plosive-liquid combinations the liquid is less likely to be devoiced when in syllable initial position. This tendency is not significant in our data.

Fricatives

In the plosive-fricative combinations /ʃt/ and /ʃp/ the /t/ is aspirated in syllable initial position and unaspirated otherwise (U-test, $p < 0.05$). A syllable initial /ʃ/ is shorter (U-test, $p < 0.1$).

The position of the syllable boundary has no effect on the combinations /fs,ʃv,ʃm,ʃn,fr,fl/.

In /ʃ.l/ and /ʃ.r/ the /ʃ/ is longer than in /.ʃl/ and /.ʃr/, respectively (U-test, $p < 0.1$).

An interesting result is given in Figure 1. Displayed are different versions of /ʃ/ in rounded and unrounded contexts. One can easily see that the feature *rounded* is determined by the vowel that belongs to the same syllable as the /ʃ/. However, this result was not reproducable for other fricatives (but both speakers in our database produced the phenomenon for the /ʃ/).

Sonorants

In combinations with a sonorant as the first sound no significant influence of the syllable boundary could be established in our data.

DISCUSSION

The influence of the syllable boundary on consonant-consonant combinations can be found for some consonant pairs. However, it does not seem to be a strong one. No categorical differences between syllable initial and non-initial versions of suitable sounds could be observed (except for the rounded /ʃ/); sounds after a syllable boundary just tend to be pronounced more clearly. And even this might be an artifact of the test material when one assumes that the duration of a sound in a syllable or syllable part is inversely correlated to the number of sounds in the respective unit [5]. The duration of a /ʃ/ in a one-sound syllable onset is longer than in a two-sound onset. This might not be a specific influence of the position of the syllable boundary but merely a timing universal that is sometimes called *isochrony*. And, if a sound is longer, it can be pronounced more accurate [4]. It seems that the influence of the syllable boundary can be called quite negligible in German. Lip rounding, however, might be an exception, but, as long as the effect is not better reproduced (just for one sound in speech from two speakers), it might just be an oddity.

PERCEPTUAL TEST

The acoustic analysis of the data showed an influence of the syllable boundary that can be regarded largely as a timing effect. To assess whether human listeners are able to use this effect a small experiment was carried out.

Method

Ten consonant pairs were selected, namely /kt, kv, fl, kl, pf, ʃp, ʃt, ʃv, tʃ, ts/. Each pair appears in our data in two types, one with the syllable boundary between the consonants, and one with the boundary after (/kt/) or before (all other pairs) both consonants. Two versions from each type were chosen randomly. They were extracted together with the surrounding vowels. Each of these stimuli appeared twice in the test. Altogether, 80 pairs were judged by the subjects.

Each stimulus was played twice. The test was recorded on cassette and played to the subjects by earphone in a quit room. Eleven subjects participated; most of them were trained phoneticians.

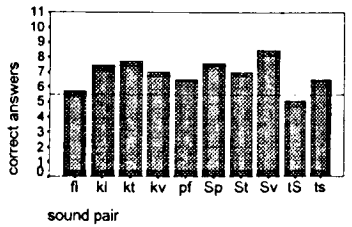


Figure 2. Results of the perceptual test displayed as correct answers for a stimulus. The horizontal line indicates the chance value.

Results

Figure 2 gives the results for each sound pair. It is easy to see that the results are better than chance value but that the correct answer rate is not overwhelmingly high. Among the subjects there are clear differences (Figure 3). Most subjects described the test as difficult; they said that they tried to guess the words according to the surrounding vowels and deduce the position of the syllable boundary from their guess applying their linguistic knowledge.

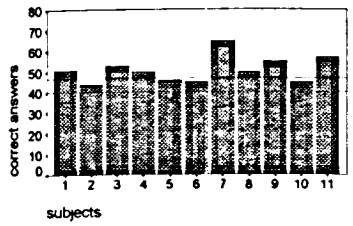


Figure 3. Results of the perceptual test displayed as numbers of correct answers from each subject. The lower horizontal line indicates the chance value, the upper horizontal line denotes the value, where the difference to the chance value becomes significant (χ^2-test, $p < 0.05$).

Discussion

Relaying on linguistic instead of phonetic hints showed to be a quite successful strategy. Therefore, the exploitable phonetic information must have been negligible, a fact that is also demonstrated by the poor recognition rates of some subjects.

GENERAL DISCUSSION

The rather obscure status of the syllable in the phonetic science [3,8] seems to be justified by the rather obscure results described above. A clear influence of the position of the syllable boundary on the properties of consonants could be established neither in the acoustic nor in the perceptual domain. The perceptual data might be consistent with some kind of low level perception of syllables as a whole that can not be applied consciously by the subjects under the artificial test conditions. But in combination with the acoustic data it is obvious that the results confirm the dubious state of the syllable in speech. However, only segmental effects were under investigation, and the syllable is an important prosodic unit [1,5]. It is difficult to distiguish between temporal and spectral domains because *target undershoot* phenomena and other reductions are a direct byproduct of a shorter sound [4].

The application of these results to concatenative speech synthesis yields to the conclusion that the syllable boundary can be neglected when choosing the units for concatenation. However, a sophisticated temporal control is necessary that does not only adjust the segmental durations according to the position in the syllable, but that supports a nonlinear timing control; for instance, aspiration is more likely to be shortened than closure duration.

ACKNOWLEDGEMENT

I thank Florian Höfer for assistance with the material, and Gerit Sonntag for carrying out the perceptual test. The research was funded by the *Deutsche Forschungsgemeinschaft.*

REFERENCES

[1] Campbell, W.N.; Isard, S.D. (1991) "Segment durations in a syllable frame." Journal of Phonetics **19**, 37-47

[2] Henderson, J.B.; Repp, B.H. (1981) "Is a stop consonant released when followed by another stop consonant?" Phonetica **39**, 71-82

[3] Kohler, K.J. (1977) *Einführung in die Phonetik des Deutschen.* Berlin: Erich Schmidt

[4] Lindblom, B. (1963) "Spectrographic study of vowel reduction." J. Acoust. Soc. Am. **35**, 1773-1781

[5] Meyer, H.; Portele, T.; Heuft, B. (1995) "Ein Silbendauermodell für die Sprachsynthese." (to appear in: *Fortschritte der Akustik - DAGA'95).*

[6] Portele, T.; Steffan, B.; Preuß, R.; Sendlmeier, W.F.; Hess, W. (1992) "HADIFIX - a speech synthesis system for German." Proc. ICSLP'92, 1227-1230

[7] Portele, T.; Heuft, B.; Höfer, F.; Meyer, H.; Hess, W. (1994) "A new high quality speech synthesis system for German." Proc. CRIM/FORWISS-Workshop on Speech Research and Technology, Munich, 284-287

[8] Tillmann, H.-G. (1964) *Das phonetische Silbenproblem.* Diss., Universität Bonn

ON THE STATUS OF THE SEGMENT IN FOUR TYPOLOGICALLY DIVERSE LANGUAGES: RESULTS FROM GLOBAL SOUND SIMILARITY JUDGMENTS

Bruce L. Derwing and Terrance M. Nearey
University of Alberta, Edmonton, Canada

ABSTRACT

On the basis of the global sound similarity judgment (SSJ) task and linear regression analysis, the phonemic segment is revealed to be a significant phonological unit in Arabic, Taiwanese (Chinese), Korean and Japanese, but with differential weightings in each case based on syllable position. A subsidiary body (CV) unit also emerged in Korean, and in Japanese the mora outranked the segment in prominence.

THE GLOBAL SOUND SIMILARITY JUDGMENT TASK

The research reported here is part of a larger cross-linguistic investigation of phonological units in languages of diverse types. One experimental technique that has proved useful in this investigation is the elicitation of global sound similarity judgments (SSJs). In this task subjects listen to systematically-varied pairs of words or pseudo-words and rate them for overall similarity in sound on some scale, usually ranging from 0 (no similarity) to 9 (identity in sound).

PREDICTING GLOBAL SOUND SIMILARITY JUDGMENTS IN ENGLISH

Following up on Vitz & Winkler's [1] attempt to predict similarity judgments on the basis of a simple phoneme-matching procedure, Derwing & Nearey [2,3] and Bendrien [4] provided strong support for the phoneme as the representational unit in terms of which English sound similarity judgments are made, while also showing a significant independent contribution of a rime (VC) unit.

PREDICTING SSJs IN OTHER LANGUAGES

Arabic

Derwing, Parkinson & Beinert [5] report on a set of SSJ results for Arabic based on a list of 47 CVCVC-CVCVC word-pairs, largely involving one-, two- or three-segment mismatches between the members of each pair (e.g., *s*akan-*m*akan, *r*asal-*r*ikal, d*ar*ab-d*uk*ub). One example each was also provided with four and with five mismatches, and one identity pair with no mismatches (katab-katab). These word-pairs were arranged into a single, randomized list for tape recording and oral presentation to subjects, except that the first six items were purposely selected to illustrate the range of variation, so as to allow subjects to mentally calibrate the zero (least similar) to nine (most similar) scale on which they were asked to rate their judgments. Subjects were 12 residents of Edmonton who were native speakers of Egyptian, Lebanese or Palestinian Arabic. The forms presented were from a "levelled" spoken dialect familiar to all the subjects.

When a series of linear regression analyses were run on the SSJ data, the segment analysis, based on matches (coded 1) or mismatches (coded 0) among the five segments in each word-pair, yielded an r^2 of 92.0%. This surpassed the results based on shared onsets and rimes (88.5%), bodies and codas (71.0%), tiers (where C-tier consists of all the consonants in each word and V-tier both vowels; 70.5%) and whole syllables (44.7%). By adding C-Tier to the variable list along with each of the five segments, the coverage for these word-pairs was increased to 93.4%, showing the significant, additive effect of the C-Tier factor. Adding the V-Tier variable contributed virtually nothing, however. This last analysis is presented in detail in Table 1 below, which shows the coefficients for each variable, as well as the *t*-ratios (all significant at $p < .01$).

We draw the following tentative conclusions from these results for Arabic: (1) Consonants contribute more to SSJ means than vowels do, as indicated by the coefficients. (Note also in this connection that only consonants are ordinarily represented in the standard orthography for this language.)

Table 1. Linear Regression Analysis for Mean SSJ Similarity Ratings in Arabic (47 CVCVC-CVCVC word-pairs)

Variables	Coefficients	*t*-ratios
C1	1.523	6.21
C2	2.038	6.84
C3	1.493	6.17
V1	1.046	5.33
V2	0.651	3.35
C-Tier	1.000	2.91

(2) Of the three consonants represented in CVCVC structures, the middle one (C2) contributes the most, perhaps reflecting the fact that it is only in the middle, intervocalic position where contrasts with consonant geminates and clusters are possible. Furthermore, (3) of the two vowels represented, the first vowel (V1) counts more than the second. (This may be an indication that a rime unit is involved, since the first vowel and the first rime were coextensive for these words, but more research will be required to clarify this point.) Finally, (4) C-Tier (scored as a match only if all three consonants were the same for a given word-pair) makes a significant, independent contribution, in addition to the individual contributions of each separate consonant. While this might be interpreted as evidence in support of the Tier model [6], the lack of relevance of the V-Tier detracts from this. Since C1...C2...C3 corresponds to the root in all of the words in this study, we suspect that this is an ancillary morphological or orthographic effect.

Taiwanese

In the Taiwanese (Chinese) phase of this study, 32 systematically varied CVC-CVC word-pairs were used as stimuli by Wang & Derwing [7], representing contrasts in the first (C1) position (e.g., *b*in33-*c*in33), the vowel (V) (e.g., t*a*n13-t*i*n13) and in final (C2) position (e.g., ci*t*21-ci*n*21), plus multiple combinations of these. (Numerals represent tone contours which were held constant within each word pair.) These items were randomized before recording on audiotape and were presented in one of two fixed orders to subjects; six duplicate items were also inserted at the beginning of each tape to illustrate the range of variation to be encountered on the test. Subjects were 105 native speakers who were recruited from four Freshman English classes at National Tsing Hua University, Hsinchu, Taiwan.

When a linear regression analysis was run on the SSJ mean ratings for these items, using the three segments in each word as variables, over 85% of the variance was accounted for (adjusted r^2=.856). The results of this analysis are shown in Table 2.

Table 2. Linear Regression Analysis for Mean SSJ Similarity Ratings in Taiwanese (32 CVC-CVC word-pairs)

Variables	Coefficients	*t*-ratios
V	4.167	11.88
C1	2.385	6.29
C2	0.780	2.06

Of the three individual segment factors, it can be seen that the vowel made the greatest contribution, followed by the initial consonant, as indicated by the relative sizes of the coefficients, and both were highly significant ($p < .001$). On the other hand, the coefficient on the final consonant factor was very small, and the contribution of that factor was only marginally significant (p =.049). In fact, if we remove the C2 factor from the analysis entirely, the amount of variance accounted for is still 83.5%. This confirms the view that Taiwanese speakers weigh the initial consonants of words more heavily than they do the final consonants, and it is perhaps less than coincidental that the initial consonant position in Taiwanese CVC words permits more than twice the range of consonantal contrasts than does the final consonant position.

Korean

Yoon & Derwing [8] report on an application of the SSJ paradigm to 48 Korean CVC-CVC word-pairs, systematically varied in composition from no phonemes in common (e.g., pin-mut) to full phonemic identity (e.g., pin-pin). These items were randomized before recording as a single list that was presented to all subjects, with four pairs inserted at the beginning of the list to preview the scale. Subjects were 15 native speakers who were students at the University of Alberta.

When a linear regression analysis was run on these data that combined the three segment variables as factors, almost 90% of the variance could be accounted for, suggesting that for Korean, as for English, Arabic and Taiwanese, segments are the most important units for predicting sound similarity. Interestingly, however, adding the CV or "body" variable to the analysis increased the coverage to 93.5%, while adding the rime variable had no significant effect. As shown in Table 3, all four of these factors were highly significant ($p < .001$), making roughly equivalent contributions to the similarity scores.

Combined with the other findings summarized in [8], this result confirms the body as an important sub-component of the Korean syllable, roughly comparable in status to the rime in English.

Table 3. Linear Regression Analysis for Mean SSJ Similarity Ratings in Korean (48 CVC-CVC word-pairs)

Variables	Coefficients	*t*-ratios
C1	0.900	7.85
V1	0.751	7.57
C2	0.710	6.20
Bo	0.756	4.66

Japanese

Finally, following up on preliminary work by Harrison [9] and Derwing & Wiebe [10], a new SSJ study was undertaken in Japanese. Stimuli were 30 real CVCV-CVCV word-pairs that were constructed around a limited set of multi-feature mismatches in one (e.g., *kami*-nami), two (*ka*ko-*ne*ko, *kamo*-*nami*), three (*kage*-*nasa*) or all four segments (*toge*-*misa*), where the segment and moraic analyses could sometimes differ. (Note that the first two examples both illustrate single mora differences, while the last three show a two mora difference.[1]) Subjects were 79 native speakers who were visiting ESL students at the University of Alberta.

When exploratory linear regression analyses were run on the results, the analysis based on the two morae accounted for 92.9% of the variance for the CVCV-CVCV items. Morever, the mora and segment analyses were clearly distinguished, with the latter accounting for only 70.0% of the variance. This large difference is reflected in the fact that, for pairs differing in one C and one V, the effect is much larger if the differences occur in two separate morae (as in pairs like *sake*-*gaka*, *kika*-*kona* or *kasi*-*nagi*), for which the overall mean rating was 0.52, than if these two differences occur within the same mora, (as in *ka*ko-*ne*ko and ta*ko*-t*ani*), for which the overall rating was 3.31. So the mora analysis is the clear winner here, since the latter accounts for more than 20% more variance with a smaller number of variables (two morae vs. four segments). Other analyses, involving units such as the rime or body, did not perform nearly as well as either of these two. However, the best analysis overall (adjusted r^2=.947) is the six-factor model analyzed in Table 4,[2] which takes both morae and all four individual segments into account, and for which five of the six variables show significant independent effects ($p(t_2) < .0001$ for both morae and $p(t_2) < .05$ for all segments except V2, for which $p(t_2)$=.0688; all variables are significant under a one-tailed analysis, which can be motivated *a priori*). In the subject analysis shown in column t_1 in Table 4, all variables are highly significant ($p(t_1) < .0001$).

Table 4. Linear Regression Analysis for Mean SSJ Similarity Ratings in Japanese (30 CVCV-CVCV word-pairs)

Variables	Coefficients	t_1	t_2
M1	3.071	23.75	8.84
M2	3.467	23.37	9.87
C1	0.555	8.19	2.17
V1	0.701	10.36	2.91
C2	0.525	10.38	2.21
V2	0.478	7.25	1.89

SUMMARY AND CONCLUSIONS

The most important finding from this study is the reaffirmation of the segment as a viable phonological unit in each of the four languages investigated. In three of the four cases, in fact (i.e., Arabic, Taiwanese and Korean), the segment clearly predominated, as the regression analyses that involved only the individual segments as factors already accounted for a substantial amount of the variance (greater than 85% throughout) in predicting the mean sound similarity ratings for the word-pairs tested. Even in the exceptional case, Japanese, where the mora unit was predominant, each of the four segments in the CVCV-CVCV comparisons were found to exhibit significant independent effects.

It is also of interest that the Japanese mora was not the only higher-order unit to emerge in this study. Specifically, the SSJ technique has revealed the influence of other sub-syllabic phonological units, in some languages, such as the rime (or VC unit) in English and the body (CV) in Korean.

NOTES

[1]While the mora and syllable units are coextensive for CVCV words in Japanese, results from other word-pairs involving constrasts between four- and five-segments have confirmed that the mora and not the syllable is the operative unit in this language, as the orthography indicates (see also [11]).

[2]Note that two *t*-ratios are provided in Table 4: t_1 reflects the reliability of the coefficients across subjects, while t_2 does the same across items, which is the only analysis done in Tables 1-3 (see Lorch & Myers [12] for discussion). Although individual data could not be processed by press time for the other languages represented here, investigations of other similar data sets have shown that all factors shown to be significant across items have remained significant across subjects.

REFERENCES

[1] Vitz, P.C. & B.S. Winkler (1973), "Predicting the judged similarity of sound of English words," *JVLVB*, vol. 12, pp. 373-388.

[2] Derwing, B.L. & T.M. Nearey (1986), "Experimental phonology at the University of Alberta," in J.J. Ohala & J.J. Jaeger (eds.), *Experimental phonology*, Orlando, FL: Academic Press, pp. 187-209.

[3] Derwing, B.L., T.M. Nearey, R.A. Beinert & T.A. Bendrien (1992), "On the role of the segment in speech processing by human listeners: Evidence from speech perception and from global sound similarity judgments," *Proc. of ICSLP 92*, vol. 1, pp. 289-292.

[4] Bendrien, B.A (1992), *Sound similarity judgements in English CVC's*, B.A. Honors Thesis, University of Alberta.

[5] Derwing, B.L., D.B. Parkinson & R.A. Beinert (in press), "Experimental investigations of Arabic syllable structure," in J. McCarthy & M. Eid (eds.), *Perspectives on Arabic linguistics* VII, Amsterdam: John Benjamins.

[6] McCarthy, J. (1981), "A prosodic theory of nonconcatenative morphology," *Linguisic Inquiry*, vol. 12, pp. 373-418.

[7] Wang, H.S. & B.L. Derwing (1993), "Is Taiwanese a 'body' language?", *Toronto Working Papers in Linguistics*, pp. 679-694.

[8] Yoon, Y.B. & B.L. Derwing (in press), "The sound similarity judgement of Korean CVC's by Korean and English speakers," *Proc. of 1994 Canadian Linguistic Association*, Toronto University Press.

[9] Harrison, K. (1992), *Syllable internal structure in Japanese phonology: Syllable, mora or segment?*, B.A. Honors Thesis, University of Alberta.

[10] Derwing, B.L. & G.E. Wiebe (in press), "Syllable, mora or segment? Evidence from global sound similarity judgements in Japanese," *Proc. of 1994 Canadian Linguistic Association*, Toronto University Press.

[11] Otake, T., G. Hatano, A. Cutler & J. Mehler (1993), "Mora or syllable? Speech segmentation in Japanese," *J. of Memory and Language*, vol. 32, pp. 258-278.

[12] Lorch, R.F., & J.L. Myers (1990), "Regression analyses of repeated measures data in cognitive research," *J. Exper. Psychology: Learning, Memory and Cognition*, vol. 16, pp. 149-157.

ACKNOWLEDGMENTS

Thanks to Kaori Kabata for her assistance with the Japanese stimuli, and to Grace Wiebe for her help with the data collection, tabulation and analysis throughout. This work was supported in part by a research grant from SSHRC awarded to the first author.

SYLLABLE SALIENCY IN THE PERCEPTION OF KOREAN WORDS

Yeo Bom Yoon and Bruce L. Derwing
University of Alberta, Edmonton, Canada

ABSTRACT

The phoneme has been assumed as the most basic phonological unit, and its universality has been proposed in the literature. A series of sound similarity judgment experiments was carried out to compare the status of the phoneme and the syllable in Korean, where both units are orthographically represented. The results showed that the syllable is the more accurate predictor of judged similarity, challenging the supposed universal primacy of the phoneme.

INTRODUCTION

The sound similarity judgment (SSJ) task employed in the present study has been found to be a useful tool for comparing the viability of phonological units across languages. The results of applying the SSJ task to English, e.g., by Vitz & Winkler [1] and other languages by Derwing & Nearey [2] have shown that the phoneme was the most basic phonological unit in most cases, and that language-specific units (e.g., the mora in Japanese) also played an important role in predicting the actual similarity scores.

In the SSJ experiments, subjects hear a series of word pairs systematically varied from sharing no common phonemes (e.g., *sit-pan*) to pairs with full phonemic identity (e.g., *sit-sit*); subjects then rate on some scale how similar each word pair is in sound. One of the strongest pieces of evidence for the phoneme came from Vitz & Winkler [1], who showed that a substantial portion of the variance in similarity scores could be explained by taking into account nothing but the number of phonemes matched between two words.

Other results showing the effects of orthography on SSJs are not surprising, since the task requires that subjects make conscious judgments, increasing the likelihood of orthographic influence. In this respect, Korean provides an interesting basis to compare the phoneme and the syllable; since both the phoneme-sized letters and the syllable-sized units are used in the orthography, the orthographic bias that favored the phoneme in English [1] and the mora in Japanese [3,5] can be neutralized in Korean.

Furthermore, there is a fair amount of evidence in Korean that suggests that the syllable is a basic level of phonological representation. First, in its orthography, individual phonemes are packaged to form syllable-sized orthographic units. Thus, for all written Korean words, syllable boundaries are straightforward. An interesting aspect of the above orthographic practice is that all syllables are written in an equi-size square, regardless of the number of phonemes in a syllable.

Another telling piece of evidence is found in the traditional poetic form *sico*, in which the syllable count is the most important metric device, e.g., the first phrase in each line always has three syllables. There is also a popular language game that shows that the syllable is a readily identifiable unit to Korean speakers. In this game, players take turns producing a new word on the basis of the last syllable of the previous word (e.g., *kakkyo* → *kyosil* → *silsu* → *su*..., etc.).

The present experiment was designed to test whether the saliency of the Korean syllable could be reflected in subjects' judgments of sound similarity. All test pairs were CV/CVC structures (where / indicates the syllable boundary) and were systematically varied from pairs that had all but one phoneme in common (e.g., CV/CVC-xV/CVC, where x indicates a mismatched phoneme) to pairs that had no common phonemes (e.g., CV/CVC-xx/xxx).

Two hypotheses were tested. First, the Syllable Hypothesis predicted that, controlling for the number of mismatched phonemes, pairs with mismatches *across* the syllable boundary (e.g., Cx/xVC) should be judged less similar than pairs with mismatches *within* a syllable (e.g., CV/Cxx), since mismatches in the former involve both syllables, while those in the latter involve only one syllable. On the other hand, the Phoneme Hypothesis predicted that there should be no significant difference between the two types of pairs, since both share the same number of mismatched phonemes. The method and predictions were based on the correlation between mean similarity scores and (i) Predicted Syllabic Similarity vs. (ii) Predicted Phonemic Similarity, as illustrated in Table 1.

Table 1. Predicted similarity on the basis of counting matched syllables and phonemes

Types of Mismatches	Predicted Syllabic Similarity S_1	S_2			Predicted Phonemic Similarity P_1	P_2	P_3	P_4	P_5		
xV/CVC	0 +	1	/2	= **0.5**	0 +	1 +	1 +	1 +	1	/5	= **0.8**
xx/CVC	0 +	1	/2	= **0.5**	0 +	0 +	1 +	1 +	1	/5	= **0.6**
Cx/xVC	0 +	0	/2	= **0**	1 +	0 +	0 +	1 +	1	/5	= **0.6**
xx/xVC	0 +	0	/2	= **0**	0 +	0 +	0 +	1 +	1	/5	= **0.4**
CV/xxx	1 +	0	/2	= **0.5**	1 +	1 +	0 +	0 +	0	/5	= **0.4**

METHOD

Subjects

A total of 117 subjects participated in the experiment on a voluntary basis. All subjects were native speakers of Korean with normal hearing. There were three groups: (i) 43 middle school students with aural stimuli only (MA); (ii) 44 middle school students with both aural and visual stimuli (MB); and (iii) 30 university students with both aural and visual stimuli (UB). These groupings were designed to test the effects of presentation mode (MA vs. MB) and age (MB vs. UB) on SSJs.

Stimuli

Twelve types of CV/CVC pairs with four tokens each were selected. The focus of our attention was on the 2- and 3-phoneme mismatched pairs, as they yield different predictions under the Phoneme Hypothesis and the Syllable Hypothesis.

The following controls were built into the stimulus pairs. First, all stimuli were real words. Second, syllable boundaries always occurred before the second C. Third, all mismatched phonemes between words were one distinctive feature away from each other. Finally, four of the identity pairs were included as control items to see if the subjects understood and were following the instructions.

Procedure

Subjects heard a series of word pairs and judged how similar each word pair sounded to them. The response measure was similarity scores on a 10-point scale ranging from zero (totally different) to nine (exactly the same). To help subjects mentally calibrate the scale, four practice pairs (one identical pair, one pair with no phonemes in common, and two pairs with some phonemes in common) were presented before the test, and the experimenter explained the approximate similarity scores for each practice pair.

The 48 pairs were recorded in a randomized order and were played back to subjects. After listening to a repetition of each word pair, subjects rated the similarity by circling the corresponding integers on the answer sheet.

RESULTS

Of the total of 117 subjects, nine subjects did not meet the inclusion criterion (four in the MA and five in the MB group). Results reported below were thus based on the remaining 108 subjects (39 MA, 39 MB, and 30 UB subjects).

To compare the Syllable and the Phoneme hypotheses, two statistics were used. First, a series of ANOVAs was run on mean similarity scores of four 2-phoneme mismatched pairs and three 3-phoneme mismatched pairs, treating both subjects and items as random factors. In these ANOVAs, the number of mismatched syllables and the number of mismatched phonemes were within-subject variables.

Analyses of Variance

Overall, the Syllable effect was a more important variable than the Phoneme effect, and the interaction between the two effects was not significant by either subjects or items in all three groups. First, in the MA group,

Table 2. Summarized results of the three groups in terms of LSD groupings

MA	MB	UB
CV/Cxx (4.90)	CV/Cxx (4.95)	CV/Cxx (5.06)
CV/xxC (3.93)	CV/xxC (4.55)	xx/CVC (4.75)
CV/xxx (3.70)	xx/CVC (4.18)	CV/xxC (4.54)
xx/CVC (3.53)	CV/xxx (3.10)	CV/xxx (3.81)
Cx/xVC (2.18)	Cx/xVC (2.76)	Cx/xVC (2.87)
Cx/xxC (1.93)	Cx/xxC (2.00)	xx/xVC (2.51)
xx/xVC (1.90)	xx/xVC (1.85)	Cx/xxC (2.48)
Mean 3.15 (sd=1.18)	3.34 (sd=1.30)	3.72 (sd=1.20)

the Syllable effect was highly significant both by subjects ($F_1[1,38] = 55.41$, $p < .001$) and by items ($F_2[1,3] = 51.48$, $p < .001$). However, the Phoneme effect did not reach significance at the .05 level either by subjects and items. The interaction was also not significant, as displayed in Figure 1.

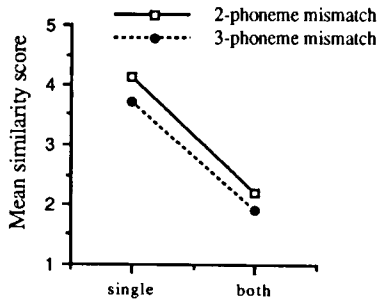


Figure 1. Mean similarity scores as a function of the number of syllables and phonemes mismatched (Middle school students with aural stimuli only, N=39)

Second, the results of the MB group, where subjects were provided with orthographic representations of stimulus pairs, were slightly different from those of the MA group, where only aural stimuli were provided. Not only did the Syllable effect again emerge as a highly significant variable ($F_1[1,38] = 42.74$; $F_2[1,3] = 30.15$, $p < .001$ for both), but the Phoneme effect was also significant ($F_1[1,38] = 25.90$; $F_2[1,3] = 16.99$, $p < .001$ for both).

Third, the results of the UB group were similar to those of the MB group, suggesting that age effect might not be important. The Syllable effect was found highly significant ($F_1[1,29] = 51.90$; $F_2[1,3] = 33.72$, $p < .001$ for both). The Phoneme effect was also significant, but with a lower level of significance ($F_1[1,38] = 9.03$, $p < .01$; $F_2[1,3] = 6.04$, $p < .05$).

Finally, to compare the seven types of pairs in detail, *Fisher's least significant differences* (*L S D*) were calculated for the three groups: MA = .81, MB = .96, UB = 1.01. Table 2 presents mean similarity scores for each type and *LSD* grouping. (Solid vertical lines include the means that are not significantly different.) Again, the Syllable effect is obvious in the above groupings. In all three groups, four 1-syllable mismatched pairs were rated higher than three 2-syllable mismatched pairs, regardless of the number of mismatched phonemes. This effect was most obvious for the MA group, where there were no overlapping types between 1- and 2-syllable mismatched pairs in terms of the *LSD* groupings.

Correlations

Using mean similarity scores for all 44 test pairs, the correlations between these scores and the Predicted Syllabic Similarity (PSS) vs. the Predicted Phonemic Similarity (PPS) were calculated for all three groups. Except for the PPS in the MA group, all correlations were higher than .80, suggesting that both the syllable and the phoneme countings could account for about 70% of the total variance in similarity scores. Again, the Syllable effect stood out in the MA group, where the coverage achieved by the PPS (53%) was much less than by the PSS (70%). The coverage of 53% was much lower than the approximately 80% that Vitz & Winkler [1] found for a variety of types of word pairs in English, suggesting that the phoneme in Korean is a less basic representational unit than it is in English.

DISCUSSION

The results reported above suggest that the syllable is the most basic and psychologically the most salient unit in Korean. The results of ANOVAs clearly confirmed the Syllable Hypothesis. This hypothesis predicted that, controlling for the number of mismatched phonemes, pairs that had mismatches *within* a syllable (e.g., xx/CVC) should be judged more similar than pairs that had mismatches *across* both syllables (e.g., Cx/xVC). It was found that the syllable effect was highly significant ($p < .001$, throughout). Furthermore, the syllable was a clear winner against the phoneme in all three groups; the phoneme effect was not significant at all in the MA group, where only aural stimuli were provided, and only marginally significant in the UB group, where both aural and written stimuli were given. The difference between the MA group and the other two may suggest that orthographic representations might have led subjects to *count* the number of letters unmatched between two written words. However, this counting must have been more difficult or impossible for subjects in the MA group, who were not provided with the written word forms.

The correlations used to compare the predictions of the Predicted Syllabic Similarity (PSS) and the Predicted Phonemic Similarity (PPS) produced a similar set of results. In the MA group, the PSS achieved much greater coverage than the PPS. In the other two groups, however, both the PSS and the PPS covered about the same percentage of the total variance. Considering that only two predicted values were used in the PSS as opposed to four predicted values in the PPS, the substantial coverage achieved by the PSS (about 70% in all three groups) suggest that the syllable is an important predictor in the SSJs of Korean words.

The difference between the MA group and the other two suggests that the presentation mode could be a variable in the SSJ experiments. Further research is required on the presentation mode effect in SSJs. However, the effect of age was evidentially not significant in the age ranges tested, judging from quite similar sets of results from both ANOVAs and correlations between the MB and the UB groups.

The primacy of the Korean syllable, the main outcome of the present SSJ experiment, is comparable to that of the mora in Japanese (see [3,4,5]). Both units are of a higher order than the phonemic segment and both find orthographic support, but only in the Korean case can the orthographic factor be winnowed out.

REFERENCES

[1] Vitz, P. C., and B. S. Winkler. (1973), "Predicting the judged similarity of sound of English words", *Journal of Verbal Learning and Verbal Behavior*, vol. 12, pp. 373-388.

[2] Derwing, B. L., and T. M. Nearey. (1994), "Sound similarity and the segment prominence: A cross-linguistic study", *Proceedings of the ICSLP*, vol. 1, pp. 351-354.

[3] Derwing, B. L., and G. E. Wiebe. (In press), "Syllable, mora or segment? Evidence from global sound similarity judgements in Japanese", *Toronto Working Papers in Linguistics*.

[4] Jaeger, J. J. (1980), *Categorization in Phonology: An Experimental Approach*. Ph.D. Thesis, University of California, Berkeley.

[5] Otake, T., H. Hatano, A. Cutler, and J. Mehler. (1993), "Mora or syllable? Speech segmentation in Japanese", *Journal of Memory and Language*, vol. 32, pp. 258-278.

ACKNOWLEDGMENTS

Thanks are expressed to Dr. E-d Cook and Dr. Rebecca Treiman for their comments on the earlier versions of the present paper. The authors also owe thanks to Dr. Sook Whan Cho, Ms. Ji Bun Kim, and students in their classes who served as willing subjects.

SPELLING ERRORS AND PHONOLOGICAL LEVELS OF SPECIFICATION

Richard LILLY
Laboratoire de Recherches Phonétiques,
Lille, France.

ABSTRACT

As shown in an earlier study of a handwritten corpus of American English [1], a large number of spelling errors cannot be explained without reference to phonological representations or operations. In this respect, reduced vowels, which appear to cause spelling errors twice as often as statistically predictable, may provide an indication that neither surface nor underlying representations are fully specified. An experiment designed to elicit spelling errors showed improved performance with words having phonologically informative derivatives. It is hypothesized that the higher level of underlying specification was responsible for lower error rates, an argument in favour of the cognitive value of certain principles of the Underspecification Theory.

INTRODUCTION

In spite of increased attention from psychologists and linguists since the publication of Frith's *Cognitive Processes in Spelling* (1980) [2], the analysis and typology of spelling errors remains a difficult and often confusing subject. One of the reasons for this situation is the fact that no classificatory tool is, so to speak, "theory-neutral". But even within a (traditional) framework, designed for the analysis of dyslexic slips or buffer memory failures, a number of statistical oddities can be discovered, all pointing to a predominantly phonological origin of spelling errors. After a brief presentation of the corpus and a discussion of the traditional typology, I proceed to show that surface representations as well as deeper constructs are involved in spelling errors. Unstressed vowels, which provide the majority of letter substitutions in the corpus, serve the argument that underspecification at the underlying level is responsible for the observed graphic indeterminacy. An experiment confirms this point a contrario by showing that an increase in underlying specification also results in improved spelling performance.

1. ERROR TYPOLOGY

1.1. The corpus

In a previous study [1], I analyzed a corpus of 204 essays, 3 to 4 pages long, written by American students of English as part of their university requirements. The precautions customary with handwritten material were taken to ensure that all collected errors were genuine, even if this meant "losing" a certain number of corrected or ill-written items. This, together with the student population under study, may explain why the corpus yielded proportionally less errors than other comparable bodies of handwritten text [3]. The *American Heritage Dictionary* [4] was used as reference, and all variants therein were considered correct (e.g. *fulfil, fulfill,* etc.). 120 essays turned out to contain one or more misspelled words, giving a total of 324. Even among misspellers, the number of faulty words per student varied a great deal (from 1 to 13, with a mean of 2.7), 23% of the subjects being responsible for 52% of the misspelled words.

1.2 Classification problems

The purpose of the study required that spelling *errors* (rather than misspelled *words*) be identified and quantified. While 3 separate mistakes can easily be isolated in **deffenitly* < *definitely* or **dissalusions* < *disillusions*, the result is less certain in **beurocracy* < *bureaucracy*, and quite impossible to assess in **oprutunities* < *opportunities*. A typology based on hypothesized causes [5] seemed open to criticism because of overlapping categories, and methodological circularity. A widely accepted structural classification (cf. [6], [7], [3]) seemed preferable, at least as a quantifying tool. An error was consequently identified each time a letter had been **Added** (**bothe*), **Deleted** (**athority*), **Changed** (**atrosious*) or **Swapped** (**marraige*). In this system, the word **beurocracy* < *bureaucracy* could be analyzed as containing 1 Addition, 1 Change and 2 Deletions. Unfortunately, this typology can produce diverging results (with obvious quantitative consequences), a fact not reported in previous research. A misspelling like **imganitive* < *imaginative* can thus receive as many as 4 different interpretations :

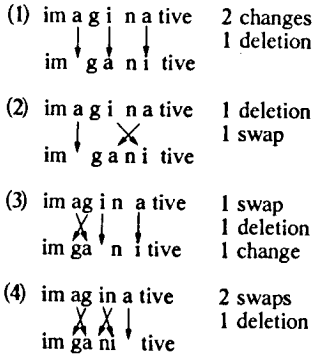


Figure 1. Four structural analyses of **imganitive*.

This difficulty was solved by imposing a certain order on the operations (i.e. Swapping, Change, Deletion, Addition), according to which more complex structural changes took place before simpler ones. Under this protocol, the four categories could function in a mutually exclusive way, yielding one and only one solution per item (solution 4 in the above example). 415 individual errors were thus unambiguously identified in the corpus, falling as follows : Added : 117, Deleted : 126, Changed : 137, Swapped : 36.

2. THE PHONOLOGICAL BASIS OF SPELLING ERRORS

2.1. The problem of units

Such classifications are not free from presuppositions, however. Let us consider errors like **acheive* < *achieve*, **thier* < *their*, etc.. The letter inversion causes them to be identified as Swaps. While factually correct, this solution misses an important point. In effect, out of 650 possible combinations of any two letters, and with a cumulated statistical probability of occurrence of .0055, *e* and *i* were involved in 72% of contiguous Swaps in the corpus... A fact for which no explanation can be offered, unless one ceases to consider letters but *graphemes*. In this respect, *ei* and *ie* happen to be the only English digraphs which are "reversible" without change of phonological value. If letters *e* and *i* have indeed been "swapped" on a superficial level, what the subjects actually did was choose the closest graphic solution to represent a given sound.

Many other statistical oddities argue in favour of considering phonological, rather than graphic representations as the operative units in spelling, leaving a small minority of errors (e.g. **convience* < *convenience*, **opionon* < *opinion*, **previuos* < *previous*, **pyscological* < *psychological*, **tevelvision* < *television*, etc.) to illustrate dyslexic or short-term memory mechanisms ("slips of the pen" proper). The level of such phonological representations remains to be discussed.

2.2. Phonetic spelling

Attempts by subjects to represent their actual pronunciation with some degree of phonetic realism are not infrequent in the corpus (e.g. **close* < *clothes*, **idenity* < *identity*, **goverment* < *government* ; **helpt* < *helped* ; **imformative* < *informative* ; **enviromment* < *environment*, etc.). The suprasegmental tendency of [r] is reflected (**oprutunities* < *opportunities*, **structrured* < *structured*, etc.), as well as the schwa deletion in the C__rV environment (**diffrently* < *differently*, **seprated* < *separated*, etc.). More generally, the frequent choice of plausible, though visually incorrect, strategies (e.g. **extreamly* < *extremely*, **lude* < *lewd*, etc.) argues in favour of surface-level driven operations.

2.3. Underlying forms and rules

The reverse tendency, namely the attempt to represent underlying, prederivational forms is also observed (e.g. **enphisis* < *emphasis* ; *respondsible* < *responsible*, **truely* < *truly*, etc.),

providing evidence that levels of representation other than the obvious surface and graphic levels have some form of psychological reality. It is equally clear that subjects are capable of modifying rule environments to regularize exceptions (e.g. : **bothe* < *both* ; **comming* < *coming*, etc.) or reinforce rule application contexts because they feel the necessity of a strong cluster (e.g. **deffenitly* < *definitely* ; **immitate* < *imitate* ; **pollitics* < *politics*, etc.) or want to avoid intervocalic voicing (**dissallusions* < *disillusions*). The interplay between this more abstract level of operations and the surface phonetic/ graphic levels is well illustrated by the numerous errors found in unstressed position.

3. SPELLING REDUCED VOWELS

3.1. Statistical evidence

The errors grouped under the Change heading present a peculiar behaviour. Out of a total of 137 errors of that type, 98 are vowels (71.5%) and 39 consonants (28.5%). Table 1 below contrasts these figures with the relative frequency of vowels and consonants in the English language (cf. Dewey [8]) as well as with their relative frequency in the corpus. The data shows that vowels are vulnerable to changes to a degree almost double what is statistically predictable.

Table 1 : relative frequency of Vowels and Consonants in Changed corpus.

	Vowels	Consonants	total
number	98	39	137
% total	71.5	28.5	100
% Dewey	38.3	61.7	100
% corpus	40.5	59.5	100

Closer examination of the Changed corpus shows that a high proportion of the Changed vowels belong to *unstressed* syllables (80/98, i.e. 81.6%).

Table 2: relative frequency of Vowels and Reduced Vowels in Changed corpus.

tot. number	98
tot. reduced	80
% Vowels	81.6

On the contrary, the remaining 18 (primary or secondary stressed) vowels were found to be misspelled because a variety of heterogeneous reasons : choice of a plausible digraph (**geered* < *geared*; *weened* < *weaned*; *teenagers* < *teanagers*), greek-style etymological spelling (**styma* < *stigma*) ; non-phononological changes (**relaxition* < *relaxation*, **intellictual* < *intellectual*; **prohibation* < *prohibition*). Clearly, the only class that presents any kind of unity is the one hosting the unstressed, reduced vowels.

3.2. Phonological underspecification

One obvious explanation for this particular vulnerability of unstressed vowels would be that the phonetic cues as to their identity are erased in such an environment. With unspecified articulatory parameters (except for the fact that it is a vowel), schwa would be characterized by *zero articulation* [9]. Now, this situation can only arise in two cases : a) if information has been deleted between the Underlying Representation and the surface (with features such as [high], [low], [back] and [tense] losing their specification) ; or b) if the UR never contained such information. Since spellers have been seen to rely on deep forms when they contain phonological information (cf. § 2.4.), their higher than normal error rate in unstressed position may be an argument in favour of case b. This (cognitive) hypothesis is in accordance with the Underspecification Theory [10] [11] [12], which argues on other grounds in favour of (variable degrees of) underlying underspecification. If we are right in postulating some form of psychological reality to this concept, any increase in feature specification should result in improved spelling performance.

4. LEVELS OF PHONOLOGICAL SPECIFICATION

4.1. The role of alternations

The fact that phonologically informative derivations (e.g. *informal, informality* ; *negative, negate*, etc.) exist in a subject's lexicon should result in such an increase in specification level. This hypothesis was tested with the following experiment. In what was presented to them as a lexical recall task, American university students were asked to supply the missing word in each of 70 unrelated sentences. All 70 target-words were words whose unstressed vowels had been misspelled at least once in the corpus. The first 40 words (part A of the test) were chosen so that derivationally related words, if any, would shed no light as to the underlying form of their reduced vowel(s). The remaining 30 items (part B of the test) were selected for the opposite reason. In this part of the test, the subjects were asked to fill in the blank and write down any "word of the same family" that they could think of.

In part A of the test, the 30 subjects found 68.6 % of the target-words (a total of 823) and made 56 spelling errors in unstressed position. In part B, they found 74.9 % of the target-words (a total of 652), and made 27 errors of that type. The error ratio (weighted by the number of target-words found) was 6.80 % and 4.14 % for parts A and B, respectively, a difference which was found to be significant ($p < .025$).

4.2. Discussion

Though clear, the improvement in performance should not be exaggerated : the subjects involved in the experiment supplied only 60% of the expected related words ; they committed a few spelling errors in spite of their knowledge (and correct spelling) of alternations ; on a few occasions, they spelled the reduced vowel of the initial word correctly, and misspelled the corresponding stressed vowel of the derived word. All in all, however, they improved their performance by 39%, which means that derivational information does help specify underlying representations.

5. CONCLUSION

The degree of indeterminacy which remains does not truly reflect the spellers' actual performance, however. A computer program, designed to simulate the above-described situation [13], still came up with improbable (and unattested) errors, until additional factors were taken into account. Among them, the familiarity of certain affix forms and the strangeness of others was found to raise or lower the probablity of occurrence of a given vowel. Phonological rules themselves bar (or impose) certain underlying vowels, e.g. after velar stops (or their softened transforms), eliminating implausible spelling errors like **eligable* < *eligible* or **nicetine* < *nicotine*. Whether such factors increase the specification of the underlying forms or on the contrary contribute to "streamline" underlying representations remains a matter for theoretical discussion and, possibly, empirical study. In the first case, subjects would use already specified representations as the basis for spelling ; in the second, they would reconstruct underlying forms, spelling so to speak "by rule".

REFERENCES

[1] Lilly, R. (1993), "Spelling errors as evidence as phonological operations". - Paper presented at the *28h Colloquium of Linguistics*, Graz, Austria.

[2] Frith, U. (1980), *Cognitive Processes in Spelling*, London: Academic Press.

[3] Wing, A.M, and Baddeley, A.D. (1980), "Spelling errors in handwriting: a corpus and distributional analysis", in: Frith, U. (1980), pp. 251-277.

[4] The *American Heritage Dictionary*, 2d ed. (1985), Boston: Houghton Miflin.

[5] Hotopf, N. (1980), "Slips of the tongue", in: Frith (1980), pp. 287-307.

[6] Conrad, R. (1964), "Acoustic confusion in immediate memory", *British Journal of Psychology*, 63, pp. 74-84.

[7] Chedru, F. and Geschwind, N. (1972), "Writing disturbances in acute confusional states", *Neuropsychologia*, 10, pp. 343-353.

[8] Dewey, G. (1970), *Relative Frequency of English Spellings*, New York: Teachers' College Press, Columbia University.

[9] Durand, J. (1990), *Generative and Non-Linear Phonology*, London: Longman.

[10] Archangeli, D. (1988), "Aspects of Underspecification Theory", *Phonology*, 5.2, pp. 183-207.

[11] Keating, P. (1988), "Underspecification in phonetics", *Phonology*, 5.2., pp. 275-292.

[12] Goldsmith, J.A. (1990), Autosegmental and Metrical Phonology, Oxford: Blackwell.

[13] Lilly, R. (1994), "Towards a computer simulation of spelling (in)competence", paper presented at the *29th Colloquium of Linguistics*, Aarhus.

LEARNING OF A PHONOLOGICAL COMPONENT FROM BREF CORPUS

A. Mailland, M. de Calmès, G. Pérennou
Institut de Recherche en Informatique, Toulouse, France

ABSTRACT

We introduce a phonological component model based on contextual phonological groups (cpg's) and multi-pronunciation groups (mpg's). The first ones are word substrings having several pronunciations depending on the context. The second ones are HMM like model of pronunciation in a given context.

The purpose of this paper is both to describe the current version of this phonological component and to present the results obtained from BREF Corpus.

INTRODUCTION

Recent developments of speech recognition have proved that taking in account phonological information improve speech recognizers - see for exemple [1], [2].

This is a point that the authors of BDLEX project had in mind [3]. Then, we have developed a phonological model compatible with HMM modeling [4].

We present here the method used in the model for learning, the parameters of this model from corpora and we give the results obtained from BREF corpus.

THE MHAT PHONOLOGICAL MODEL

The general model – the MHAT model (Markovian Harmonic Adaptation and Transduction) is described in [4]. We use here a particular model where the lexicon contains the phonological representation of inflected words.

The figure 1 shows the structure of the model.

Syntactic level S

The representations are the surface forms which are generated by the grammar. They consist of word-class strings. (S,S) transformations insert word boundaries : # for required liaison, ≠ for optional liaison and ! for prohibited liaison. These boundaries depend on the syntactic stucture of the sentence.

These boudaries play an important role in French phonology. In [5], we have proved that a markovia biclass model is a good approximation of the (S,S) transformation (a ideal model must include prosodical features).

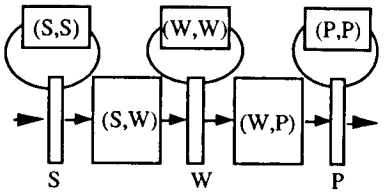


Figure 1 - MHAT model.

W-level

The representations are strings of both subword units and word boundaries #, ≠ or ! . In our model, the subword units are contextual phonological groups (the cpg's) defined as strings of interdependent phonemes having context-dependent realisations. For example the word «grandes» (big) has, in our model, the W-representation gRã‹~də› instead of the standard representation /gRãdə/. The first three units are the same in the two representations (they are called trivial units in our model) but the last one in the W-representation covers two phonemes which have interdependent and context-dependent realisations.

An harmonic transformation (in the sense given in Golsmith [6] , see also [4]) adapts W-representations by rewriting all cpg's into context-independent phonological units and deleting all boundaries. The result is a new W-representation, called here phonotypical representation, which consists of a string of multi-pronunciation groups (the mpg's units). For example, the W-representation of «grande fenêtre» is gRã‹~də› # ‹fə›nɛ‹trə› which becomes the W'-representation gRã(~də) fŒnɛ(trə›) where three units have multiple pronunciations depending on the speaker,

the style ... but not of the context. The first one is (~də) which can be pronounced [dŒ] or [n]. The second one is (trə›) which can be pronounced [tRŒ] or [tR] or [tR]. The third one is Œ which have three possible realisations : [ø], [a], [œ]. The other units have only one pronunciation and are considered as trivial mpg's. In order to represent cpg's and mpg's, we adopte the following conventions :

- if the phonetic substring x has one and only one pronunciation, their phonological and phonotypical representations are x (x is a trivial cpg and a trivial mpg).
- if x is constituted by interdependent phonemes (that is the phonetic realization of one among them which is dependent on the phonetic realization of the others) then :

• ‹x› is a cpg
• (x) is a mpg

If ‹x› depends only on its context within the word, it's a trivial cpg which will not be affected when it will be inserted in a sentence. Then, the notation (x) is used instead of ‹x›.

More details are given in previous papers [7], [8], [9]. Probabilities can be assigned to phonetical rules. Thus, the phonetical rule for (~də) which associates two realisations : dŒ and n is represented by :

(~də) —> dŒ (0.4) ! n (0.6)

In this way, mpg's can be seen as hidden Markov model subword units.

Phonetic level

At this level the P-representations consist of strings of phonetic units in a given alphabet (here the IPA of the standard transcription of French). Thus, the last example can have several P-representations : [gRãdœ fŒnɛtR] [gRãn fŒnɛtRœ] , [gRãn fŒnɛt] ...

Internal adaptations occur also at this level for taking into account coarticulation effects. This will not be discussed here.

Here, the model used is simplified. It takes into account only end word mpg's describing phonological phenomena such as liaison. Few not trivial internal mpg's are conserved for foreign origin words.

GENPHON system

The purpose of GENPHON is to transform an orthographic form into a phonotypical one. This single transcription takes all the various possible pronunciations into account. It consists of three units :

• a lexical access module which permits to generate the phonological form with syntactic categories from orthographic form. A phonological form is composed of cpg's and mpg's.
• a module for positioning the phonological boundaries.
• a phonological module which yields the phonotypical transcription by using phonological rules.

For each word of a sentence, the phonological representation and syntactic class are retrieved from a BDLEX-derived lexicon [10].

Boundaries are positioned according to the bigram model introduced in [5] based on the work of P. Delattre [11].

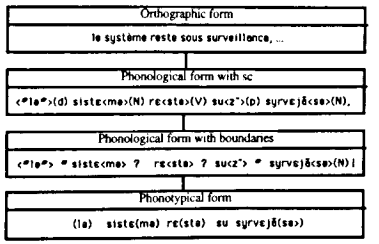


Figure 2 : Example of phonotypical generation.

In order to generate the phonotypical form of the utterance from the phonological form with cpg's, we use a set of phonological rules [5]. For every cpg's class, this rulebase provides a mpg for a given context.

Figure 2 shows an example of phonotypical form generation.

THE LEARNING SYSTEM

The proposed learning system (cf. fig.3) uses a variant of the alignment tool VERIPHON [12], developed at IRIT. It affords both advantages of being specially well adapted to align mpg like groups and of supplying statistics about the pronunciation variants observed within the corpus. As input, it takes a phonotypical transcription stemmed from the phonological component and its corresponding phonetic transcription proceeded from speech corpora. The system yields the aligned utterance

represented by a string of mpg/phonemes couples.

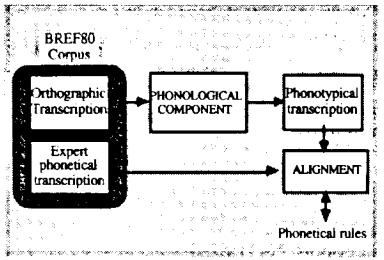


Figure 3 : Learning system.

RESULTS

GENPHON and the learning system have been experienced on the BREF80 Corpus [13]. The results have some phonetic implication and allow us to specify the impact of the phonological alterations on the automatic speech recognition.

BREF80 has been transcribed in two steps :
1) a phonetic transcription is yielded by GRAPHON, the grapheme-to-phoneme conversion system of LIMSI,
2) this transcription has been rectified by experts.

It is composed of 5323 sentences pronounced by 80 speakers in a dictation style.

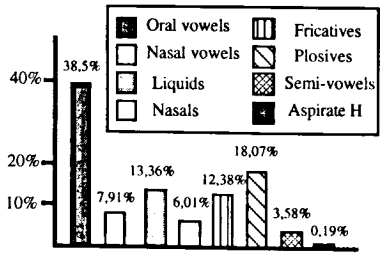


Figure 4 : Trivial mpg distribution.

BREF80 includes 345 000 occurences of mpg. This means 387 distinct mpg's including 40 trivial mpg's. These trivial mpg's account for over 90 % of the occurences.

Most of trivial mpg's is represented by standard phonemes. Their distribution is given figure 4. Nearly 50% are vowels.

For our purpose, the most important phenomena occur in the non-trivial ending mpg's. Figure 5 shows their distribution by phonetic class.

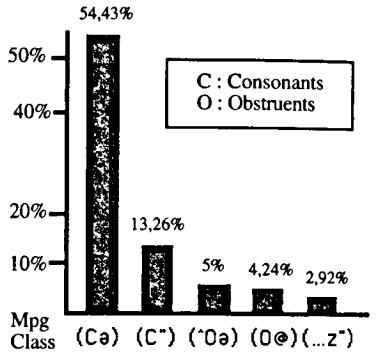


Figure 5 : Non-trivial mpg distribution.

The maximum for the (Cə) class is certainly due to numerous monosyllabic pronouns, articles, ...

(C") represents a latent consonant, for example in the word *encombrant* /ãkɔ̃brɑ̃(t")/.

(^Oə) and (O@) are respectively in the words *poursuite* /pursɥi(^tə)/ and *tarif* /tari(f@)/.

(...z") represents the end of a plural word, for example in *nouvelles* /nuvɛ(ləz")/.

Here, we can only present two results about important phonological phenomena of french : the liaison and the schwa elision. More exhaustive results are given in [9].

Frequency of liaison realization depends on numerous factors [11]. Using this model, the study of the liaison processing has been made in [4].

We give, figure 6, liaison frequencies for some final mpg's in a plural word.

Mpg Class	Liaison	Non-liaison
(Wəz")	23,91%	76,09%
(Bəz")	12%	88%
(^Qəz")	12,04%	87,96%
(^QLəz")	24,32%	75,68%

Figure 6 : Liaison after a plural word where W : set of liquids and nasals, B : set of voiced obstruents, Q : set of unvoiced obstruents and L : set of liquids.

Non-liaison is always more frequent. The more frequent liaisons produce in the

(^QLəz") class (in this case schwa is pronounced). The average of liaison frequency is 18,06% for these classes.

We give, figure 7, schwa elision frequencies on Consonant-Schwa mpg's. These results shows that the kind of consonants plays a part in the elision phenomenon.

Mpg	Realizations			
(Lə)	LŒ	42,69	L	57,31
(Nə)	NŒ	24,36	N	75,64
(Fə)	FŒ	38,28	F	61,72
(Pə)	PŒ	76,23	P	23,77

Figure 7 : Frequencies per cent of schwa elision where L : set of liquids, N : set of nasals, F : set of fricatives and P : set of plosives.

These results illustrate the fact that frequencies of phonological phenomena are very variable according to mpg's in which they appear. This implies that a such component must be train from large corpora.

CONCLUSION

Using the phonological model based on contextual phonological groups (cpg's) and multi-pronunciation groups (mpg's), we have showed that it was possible to learn automatically pronunciation likelihoods associed to mpg's. This learning has been made thanks to a phonological lexicon where words are represented in cpg's, and two bases of rules. The first one defines cpg's. The second one describes mpg's and is learned from a transcribed corpus.

Achieved results concern a speaker population in a given communication situation : text reading.. Such results show a learning ability for a task of oral man-machine communication. They allows to show the impact of phonological variations in oral production and better to place the role of the phonological component in speech recognition systems.

Such a phonological component is compatible with speech recognition systems based on HMM.

REFERENCES

[1] J.-L. Gauvain, L.F. Lamel, G. Adda, M. Adda-Decker, "Speaker-Independent Continuous Speech Dictation", Eurospeech93, pp. 125-128.

[2] E.P. Giachin, A.E. Rosenberg & C.-H. Lee, "Word Juncture Modeling Using Phonological Rules for HMM-Based Continuous Speech Recognition", Computer Speech and Language,pp.155-168 (1991).

[3] G. Pérennou, "BDLEX : A Data And Cognition Base Of Spoken French", Proceedings of ICASSP, Tokyo, pp. 325-328 (1986).

[4] G. Pérennou, "Phonological Component in Automatic Speech Recognition. The case of Liaison Processing", Levels in Speech Communication - Relations & Interactions, pp. 211-223, 1995.

[5] G. Pérennou, "Introduction aux groupes à prononciations multiples suivi d'un sous-ensemble phonologique du français", IRIT report 94-07-R, 1994.

[6] J. Goldsmith, "Autosegmental & Metrical Phonology", Basil Blackwell, 1990.

[7] A. de Ginestel-Mailland, G. Pérennou, M. de Calmès, "Une approche de la phonologie en reconnaissance de la parole", Interface des mondes réels et virtuels, Montpellier, 22-26 Mars 1993, pp. 345-354.

[8] A. de Ginestel-Mailland, M. de Calmès, G. Pérennou, "Multi-Level Transcription of Speech Corpora from Orthographic Forms", Eurospeech93, pp. 1441-1444.

[9] A. Mailland, M. de Calmès, G. Pérennou, "Transcription multi-niveau d'un corpus de parole", JEP94, pp. 309-313.

[10] G. Pérennou, D. Cotto, M. de Calmès, I. Ferrané, J.-M. Pecatte, "Le projet BDLEX de base de données lexicales du français ecrit et parlé", Séminaire Lexique, Toulouse, 21-22 Janvier 1992, pp. 153-171.

[11] P. Delattre, "Studies in French and comparative phonetics", Mouton & Co. The Hague, 1966.

[12] G. Pérennou, H. Kabré, M. de Calmès, J.M. Pécatte, N.Vigouroux, "Une approche de l'étiquetage automatique indépendant du locuteur", Séminaire variabilité du locuteur, Avignon, 20-21 Juin 1989, pp. 61-67.

[13] L Lamel, JL Gauvain, M Eskénazi, "BREF, a Large Vocabulary Spoken Corpus for French", Eurospeech91, Genova, 24-26 Sep 1991. pp. 505-508.

VOWEL INVENTORIES AND VOWEL PROCESSES WITHIN OPTIMALITY THEORY

Abigail R. Kaun
University of California, Los Angeles

ABSTRACT

Optimality theory allows for a unified analysis of the role of feature combination markedness in both segmental inventories and contextual phonological processes. The data and analyses presented in support of this proposition focus on the distribution of rounded vowels in vowel inventories and on the patterns observed in rounding harmony systems.

INTRODUCTION

The principles which underlie the cross-linguistic patterns observed in the structure of segment inventories play a role in syntagmatic phonological patterns as well. In *The Sound Pattern of English* [1], this relationship was characterized by means of marking conventions characterizing marked and unmarked feature combinations. Marking conventions constituted a theory markedness in segment inventories. Furthermore, given their status as elements of Universal Grammar, any derivation yielding marked feature combinations would necessarily entail language-specific rules, thus adding complexity to the grammar. Under the assumption that simplicity is favored over complexity in grammars, it follows that, all else equal, rules required in order to generate marked feature combinations will be typologically dispreferred.

OPTIMALITY THEORY

In optimality theory [2], [3] the phonological *rule* has no formal status and is instead replaced by constraints on surface representations. While optimality theoretic constraints refer to surface well-formedness, they are not necessarily surface-true. Crucially, constraints may conflict with one another in the sense that a given representation may satisfy one constraint while violating some other constraint.

Such conflicts are resolved by means of a hierarchical constraint ranking. Constraint hierarchies are characterized by what has been termed *lexicographic* ordering: a given constraint overrides all lower-ranked constraints. Thus, a single violation of any given constraint is worse than multiple violations of one or more lower-ranking constraints. The constraints themselves constitute part of Universal Grammar, however their relative ranking is determined on a language-specific basis.

For a given input (or lexical) representation, output (or surface) representations are determined by means of an algorithm which compares an array of candidate output structures and evaluates the extent to which they satisfy the constraint hierarchy. The optimal candidate is that which best satisfies the constraint hierarchy and it is that candidate output structure which surfaces. Clearly, the value of a given output candidate will be determined in part by the degree to which it constitutes a faithful rendition of the input.

The constraints which insure input-output fidelity are referred to as *faithfulness* constraints. Faithfulness constraints require that specifications present in the input remain in the output (PARSE) and that specifications absent in the input are not added to the output (FILL).

SONORITY AND VOICING

Within this model it is thus possible to characterize both surface inventory patterns and contextual phonological patterns. For instance, consider the well-known observation that sonorants are typically voiced. In Maddieson's survey of the segment inventories of 317 languages [4], only 3.4% of nasal consonants, 3.3% of approximant laterals, 2.2% of trills and 1.9% of taps/flaps were voiceless. From this, two conclusions are relevant: (i) voiceless sonorants are rare, and (ii) voiceless sonorants are attested.

Optimality theory is well-suited to account for both of these facts by means of the interaction between constraints on markedness and constraints on faithfulness. Voiceless sonorants will be allowed to surface only if both PARSE[voice] and PARSE[sonorant] outrank the constraint dictating that [-voice] and [+sonorant] are incompatible, which I will label *[-vce, +son]. Of the six logically possible constraint hierarchies in Figure 1, only the first two listed will allow voiceless sonorants to surface:

PARSE[vce] >> PARSE[son] >> *[-vce, +son]
PARSE[son] >> PARSE[vce] >> *[-vce, +son]
PARSE[vce] >> *[-vce, +son] >> PARSE[son]
PARSE[son] >> *[-vce, +son] >> PARSE[vce]
*[-vce, +son] >> PARSE[vce] >> PARSE[son]
*[-vce, +son] >> PARSE[son] >> PARSE[vce]

Figure 1. Constraint Hierarchies.

In addition to the rarity of voiceless sonorants in segment inventories, the resistance of sonorants to processes of devoicing can also be attributed to the markedness constraint *[-vce, +son]. In Russian, for instance, the fact that obstruents are targeted by word-final devoicing while sonorants are not can be analyzed as an instance in which *[-vce, +son] outranks the constraint (or set of constraints) which gives rise to final devoicing. In Angas, a Chadic language of Nigeria [5], sonorants are subject to final devoicing. Thus, in this language *[-vce, +son] ranks lower than those constraints which conspire to yield final devoicing.

VOWEL CONSTRAINTS

A number of very solid generalizations can be made regarding the occurrence of rounded vowels in vowel inventories. With respect to the height dimension, it has long been observed that low vowels are typically unrounded. Of the 523 low vowels recorded in Maddieson's survey, only 37 (or 7%) were rounded. Therefore, it is very clearly safe to say that low rounded vowels are marked. In *The Sound Pattern of English* [1], this correlation is captured in Marking Convention XI, which states that the unmarked value of [±round] is [-round] in the context [__, +low]. In optimality theory, this correlation must be expressed as a constraint on feature incompatibility. I will label the relevant constraint *ROLO. *ROLO dictates that low rounded vowels are dispreferred.

Similarly, front rounded vowels are attested, but typologically dispreferred. Of the 1019 front vowels recorded in Maddieson's survey, only 61 (or 6%) were rounded. This percentage should be compared with the percentage of back vowels in the survey which were rounded. Of the 964 back vowels recorded, 901 (or 93.5%) were rounded. It is therefore reasonable (and not novel) to claim that front rounded vowels are dispreferred. Marking Convention XI in *The Sound Pattern of English* captured the relationship between backness and rounding as well. In optimality theory, we establish a constraint which I will label *FRORO, expressing the dispreference for front rounded vowels. By the same token, back unrounded vowels are clearly in the minority suggesting the need for a third constraint which we may label *BAK-RD.

Now just as the markedness constraint referring to sonority and voicing was shown to play a role both in the shaping of segment inventories and in the typology of final devoicing, the constraints on rounding and the height and backness dimensions can also be shown to participate both in determining the content of vowel inventories and in the typology of vowel harmony.

BACKNESS HARMONY

In the native vocabulary of Hungarian, front and back harmonic vowels do not co-occur within a word [6]. The so-called "neutral" vowels (*i(:)* and *e:*) freely co-occur with vowels of both harmonic classes.

Front	Back	Neutral
y, y:	*u, u:*	*i, i:*
œ, œ:	*o, o:*	*e:*
ɛ	*a, a:*	

Figure 2. Hungarian Harmony Classes.

It is immediately apparent from the harmony classes shown in Figure 2 that the neutral vowels are the front vowels which lack a back counterpart. Thus, the failure of *i(:)* and *e:* to participate in harmonic alternations is apparently linked to the absence of *ɯ(:)* and *ʌ:* in the surface inventory. Those constraints giving rise to backness harmony, discussed in Kaun [7], rank lower than *BAK-RD. And similarly, *BAK-RD outranks the PARSE constraints which would force *ɯ(:)* and *ʌ:* to surface. Stated differently, the grammar places a higher priority on the avoidance of the marked feature combination [-round, +back] than it does on the need to maintain backness harmony throughout the word.

ROUNDING HARMONY

More interesting cases from the perspective of the point I make in this paper are to be found in rounding harmony systems. The interest of these cases lies in the fact that the markedness constraints *ROLO and *FROLO are surface-violated, that is low (or lower mid) rounded vowels and front rounded vowels do occur in the relevant languages. Nonetheless, the effects of these constraints are clearly visible in the observed rounding harmony patterns.

*FROLO

In the Mongolian dialect Shuluun Höh [8], both front and back rounded vowels are present in the surface inventory. Furthermore, both front and back rounded vowels occur as the output of rounding harmony. Rounding harmony is triggered by non-high vowels and targets non-high vowels. Some examples are shown in Figure 3, where the targets of harmony are underlined. In addition to rounding harmony, Shuluun Höh exhibits a system of ATR harmony, described in detail in Svantesson [8]:

jorɔxø̲:lɔxtʃ-	'president'
nɔxœ̲:-	'dog'
dɔro̲:-	'stirrup'

Figure 3. Shuluun Höh root harmony, data taken from Svantesson [8].

Front and back vowels do not exhibit entirely parallel distributional patterns, however. While both front and back rounded vowels occur as the output of rounding harmony within roots, only back rounded vowels are found as the output of rounding harmony in suffixes. Examples of suffixal vowels are underlined in Figure 4:

nɔxœ:-gɔ̲:r	'dog-instrumental'
dɔro:-gɔ̲:r	'stirrup-instrumental'
mœrj-tæ̲:	'horse-comitative'
(*mœrj-tɶ̲)	
obst-te̲:	'grass-comitative'
(*obst-tø̲)	

Figure 4. Shuluun Höh suffix harmony, data taken from Svantesson [8].

In optimality theory this asymmetry is elegantly captured by means of constraint ranking. Specifically, the constraint which gives rise to rounding harmony within roots, which I will label $\text{EXTEND}^{\text{R,Root}}$ outranks *FROLO (see Kaun [7] for a full analysis). The analogous constraint which operates at the level of the word, $\text{EXTEND}^{\text{R,Word}}$, is ranked below *FROLO. Stated in prose, it is more important to extend the domain of rounding throughout the root than it is to avoid generating the typologically marked front rounded vowels. The avoidance of such vowels in Shuluun Höh takes precedence over the importance of extending the domain of rounding throughout the entire word. The relevant partial constraint hierarchy is therefore:

$$\text{EXT}^{\text{R,Root}} \gg \text{*FROLO} \gg \text{EXT}^{\text{R,Word}}$$

Shuluun Höh thus constitutes a case in which a constraint on feature combination markedness plays an active role in the phonology while not imposing a limitation on the surface segment inventory. The constraint *ROLO can be shown to play a role in some rounding harmony languages which feature low (or lower mid) rounded vowels in the surface inventory.

*ROLO

In Turkish, rounding harmony is triggered by high and non-high rounded vowels, but targets only high vowels. Thus, a high vowel suffix such as /-Im/ 'first person singular possessive' surfaces with a rounded vowel whenever the preceding vowel is rounded. Vowels in Turkish are subject to backness harmony as well, as indicated: *buz-um* 'my ice', *kɔl-um* 'my arm', *gül-üm* 'my rose', *gœz-üm* 'my eye'. A non-high suffix vowel surfaces as unrounded regardless of the quality of the preceding vowel: *buz-da* 'on the ice', *kɔl-da* 'on the arm', *gül-de* 'on the rose', *gœz-de* 'on (in) the eye'.

The failure of non-high vowels to undergo rounding harmony is attributable to the constraint *ROLO . In Turkish, the relevant PARSE constraint must rank above *ROLO, giving rise to non-high rounded vowels on the surface. Non-high rounded vowels are not found as the output of harmony however, and this distributional fact can be accounted for if we assume that *ROLO ranks above the relevant EXTEND constraints. The distribution of the marked vowels ɔ, œ in Turkish is therefore quite restricted. These vowels occur where required to satisfy PARSE, i.e. in positions lexically specified as [+round, -high]. They fail to occur in positions where EXTEND would dictate in their favor due to the influence of the higher ranking constraint *ROLO. We thus have the partial constraint hierarchy shown here for Turkish:

$$\text{PARSE}^{\text{R}} \gg \text{*ROLO} \gg \text{EXTEND}^{\text{R}}$$

In Kachin Khakass, a Turkic language documented in Korn [9], high and non-high rounded vowels are found in the surface inventory. Rounding harmony applies only between a high trigger and a high target, however:

kuʃ-tuŋ	'of the bird'
kɔzuk-ta	'in the nut'
(*kɔzuk-tɔ)	
ɔk-tuŋ	'of the arrow'
(*ɔk-tuŋ)	
pɔl-za	'if he is'
(*pɔl-zɔ)	

Figure 4. Rounding Harmony in Kachin Khakass, data from Korn [8].

An additional constraint is relevant to the Kachin Khakass pattern which, in Kaun [7], is labeled $\text{UNIFORM}^{\text{R}}$. This constraint is operative in a variety of other rounding harmony languages including the Mongolian and Tungusic languages, and dictates that a single [+round] autosegment in the phonology should correspond to a uniform articulatory setting in the phonetics. $\text{UNIFORM}^{\text{R}}$ rules out harmony when the trigger and target are of distinct heights, since the lip activity involved in the articulation of non-high rounded vowels is distinct from that involved in the articulation of high rounded vowels [10].

The constraint hierarchy for Kachin Khakass is therefore the following:

$$\text{PARSE}^{\text{R}} \gg \text{UNIFORM}^{\text{R}}, \text{*ROLO} \gg \text{EXT}^{\text{R}}$$

This hierarchy gives rise to a decidedly limited range of harmony configurations in the language. EXTEND^{R} may prevail only when its satisfaction entails violations of neither $\text{UNIFORM}^{\text{R}}$ nor *ROLO .

CONCLUSION

I have shown that optimality theory provides a unified means of characterizing the role of markedness constraints on feature co-occurrence in both inventory structure and in contextual phonological phenomena. In this model constraints are universal, but also violable. We therefore expect to find cases in which a given markedness constraint is ranked quite high, imposing limitations on both the segmental inventory and syntagmatic phenomena (e.g. Hungarian). In addition, we expect to find cases in which the markedness constraints rank somewhat lower, and fail to impose restrictions on the shape of the surface segmental inventory. In such cases, the markedness constraints may still be expected to play a role in certain contextual phonological manifestations. It is this situation which is encountered in Shuluun Höh, Turkish, and Kachin Khakass.

REFERENCES

[1] Chomsky, N. & M. Halle. (1968), *The Sound Pattern of English*, Cambridge: MIT Press.

[2] Prince, A. & P. Smolensky. (1993), *Optimality theory: constraint interaction in generative grammar*, ms., Rutgers University and University of Colorado at Boulder.

[3] McCarthy J. & A. Prince. (1993), *Prosodic morphology I: constraint interaction and satisfaction*, ms., University of Massachusetts & Rutgers University.

[4] Maddieson, I. (1984), *Patterns of Sounds*, Cambridge University Press.

[5] Burquest, D. A. (1971), *A preliminary study of Angas phonology*, Zaria: Institute of Linguistics.

[6] Ringen, C. O. (1988),*Vowel harmony: theoretical implications*, Garland Publishing, Inc.

[7] Kaun, A. R. (1995), *The typology of rounding harmony: an optimality theoretic approach*, doctoral dissertation: University of California, Los Angeles.

[8] Svantesson, J.-O. (1985), "Vowel harmony shift in Mongolian", *Lingua*, vol. 67, pp. 283-327.

[9] Korn, D. (1969), "Types of labial vowel harmony in the Turkic languages", *Anthropological Linguistics*, vol. 11, pp. 98-106.

[10] Linker, W. (1982), *Articulatory and acoustic correlates of labial activity in vowels: a cross-linguistic study*, doctoral dissertation: University of California, Los Angeles.

A MINIMALIST APPROACH TO PHONOLOGY

David Michaels
University of Connecticut

ABSTRACT

This paper explores a minimalist approach to consonant cluster phenomena in phonology. The approach avoids language particular devices such as context sensitive rules and rule ordering. Instead a single universal operation *Attract F*(eature) is used to relate surface to lexical representations.

INTRODUCTION

The minimalist program [1] assumes that there is a language faculty. The initial state of this faculty is Universal Grammar (UG) which maps data into a Grammar, its final state. Early generative approaches to both syntax and phonology carried a heavy descriptive burden in the form of a grammar with complex rules. In the minimalist program the burden is shifted to UG with a single rule that can relate any two items at any stage of a derivation. Work along these lines is proceeding in syntax with some success. For phonology, however, it has been argued that the minimalist approach is inappropriate [2], that in phonology explicit rule ordering and the intermediate structures they specify are necessary.

In this paper, I investigate how the minimalist program can work in phonology. In the phonological literature there are many examples of deep ordering relations that hold among phonological rules. For example, in the analysis of Southern Paiute (SP) [3,4], there is rule ordering that is seven rules deep: C-deletion, Gemination, Spirantization, Stress, Degemination, V-devoicing, Sonorant devoicing. I show that in a system with X-bar projections of syllable structure, the setting of coda and stress parameters, a theory of markedness and the principle *Attract F(eature)*, the various combinations of Gemination, Spirantization and Devoicing can be accounted for without context sensitive rules and rule ordering. For example, a morpheme final abstract consonant is posited in SP. C-deletion deletes this final abstract consonant in word final position, while in nonfinal position Gemination causes that consonant to take on the features of a following consonant. Where the abstract consonant is deleted, the preceding vowel devoices (V-devoicing). The problem is how to account for such phenomena without context sensitive rules.

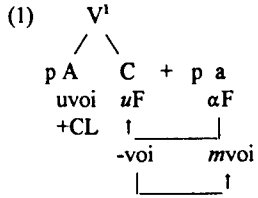
THE THEORY

I assume a lexicon that contains array of morphemes with phonological representations in which segments that alternate for some feature *F* are represented as unmarked for that *F* (*uF*). Otherwise segments are +/-*F* as appropriate. I also assume X-bar projection of syllable structures: where every vowel (V^0) projects a Rhyme (V^1) which optionally licenses a coda, and V^1 projects a Syllable (V^2) which obligatorily licenses an onset. The single operation is *Attract F* which can relate any two nondistinct *F*s in a phonological representation subject to certain universal constraints. First, the two *F*s cannot be related across a syllable head (a locality constraint). Second, unmarked or marked (*u/m*) *F*s are nondistinct from +/- *F*s. Thus, only the *u/m F*s of a consonant (or syllable head) can be related to the +/- *F*s of an adjacent consonant (or head of an adjacent syllable). The attraction between *u/m F*s and +/- *F*s is motivated by the requirement that all *F*s be phonetically interpretable, a requirement of the phonetic interface with phonology. Thus, *u/m* values for *F*s are motivated by learnability considerations since these *F*s vary in the data, and *Attract F* is motivated by phonetic interpretation which requires that *u/m* values have +/- interpretations.

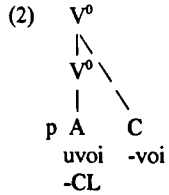
SOUTHERN PAIUTE

To return to gemination and deletion in SP, let us say that *Attract F* creates a chain of *F*s between two adjacent consonants (...C_1+C_2...), where C_1 is the morpheme final abstract consonant, abstract since it alternates as a copy of whatever C_2 follows it. Thus, C_1 is *u/m* for all its place *F*s and is interpreted by *Attract F* for those *F*s specified in C_2 (Gemination). C_1 is always [-voice]. C_2, however, alternates for voice. In particular, it is [+voice] between vowels. I assume it is lexically [*m*voice]. In word final position, there is no local consonant to chain with the final abstract consonant, however, its [-voice] can interpret the [*u*voice] of the preceding vowel. These derivations are illustrated in (1) and (2) respectively.

(1)



(2)



SP places stress on alternate Vs, but never on a final V. (1) illustrates a sequence of a stressed syllable followed by an unstressed syllable. Here, the stressed syllable licenses a coda position which is filled by the abstract C. The unmarked place *F*s of C attract the specified place *F*s (αF = +ant, +cor, -high, -back) of *p*. Since no syllable head intervenes, interpretation by attraction of features is possible. In (2), a final unstressed syllable, there is no licensed coda position and no segment to the right of C, and the consonant to the left is too far away (across a syllable head) to interpret C. Thus, C adjoins to the preceding V. C's [-voice] percolates to the adjunction node where it merges with A's [*u*voice] to yield a voiceless vowel. In this way we get both C-deletion and V-devoicing to follow automatically from the syllabification algorithm. Thus, in this account there no separate gemination, devoicing and deletion rules and hence no ordering between them.

In order to get this account to work under minimalist assumptions, UG is assumed to provide a syllabification algorithm, a theory of markedness, feature percolation and the operation *Attract F*, subject to locality constraints. In addition, I assume that stressed syllables license coda positions while unstressed syllable do not. Let me formulate this last condition on codas as the result of a coda licensing feature *CL* already illustrated in (1, 2). Thus, instead of assigning stress by rule, vowels that alternate for stress are [*u*CL], stressed vowels are [+CL], unstressed vowels are [-CL]. Languages that are stressed from the right have a suffixed [+CL] affix. Those that are stressed from the left have a prefixed [+CL] morpheme. In either case, [*u*CL] vowels attract [+CL] subject to the usual locality constraint. However, [+CL] requires an available segment to be licensed in coda to be realized on a particular syllable. Thus in (1), [+CL] falls on the first syllable to license the coda position. Once C is licensed as coda, it cannot adjoin to V^0, hence the contrast with (2).

JAPANESE

Let us turn now to a similar phenomena in Japanese. Here the initial consonant of the past tense suffix assimilates for voice to the final consonant of the verb stem, and

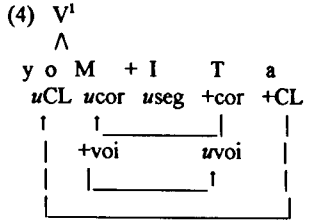
the final consonant of the verb stem assimilates for place to the initial consonant of the past tense suffix. Where the stem is vowel final, a voiceless *t* emerges. When the stem ends in *s*, an *i* emerges between the stem final *s* and the suffix *t* and palatalizes the *s* to *š*. I assume that the lexical representation of the past suffix is *-ITa*, where *I*, which alternates with Ø, is unmarked for the feature [segment] and *T* which alternates for voice is unmarked for that feature. Thus, we get examples as in (3).

(3)

	Past	*UR*	*Gloss*
a.	mita	mi+ ITa	look at
b.	yonda	yoM+ITa	read
c.	kašita	kaS+ITa	lend

In the case of (3a) *mita*, the suffix *I* adjoins to the stem *i* and *T* gets its unmarked interpretation via universal marking conventions [3]. In the case of (3b) *yonda*, the stem final nasal alternates with *m* and is unmarked for [coronal] (compare *yomu* (nonpast)). The derivation of *yonda* from lexical *yoM+ITa* is illustrated in (4).

(4)



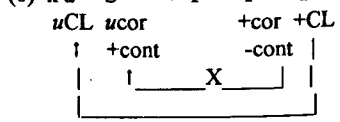
Thus, the coronal value of *T* interprets *M* giving *n* and the voice value of *M* interprets *T* giving *d*.

However, in both instances *Attract F*, in the preceding derivation, applies across lexical *I*, apparently violating the locality constraint prohibiting attraction of *F*s across a syllable head. Also, we must explain why *I* does not take *M* as its onset and form a syllable. In the latter case, I assume that the past tense suffix *ITa* is specified [+CL] in Japanese. Thus the preceding vowel with [*u*CL] can attract [+CL] if it can license a coda. The stem vowel is followed by a consonant *M* which can satisfy the coda condition. Therefore, *M* is licensed in coda position. *I*, unmarked for [segment], must be invisible for the purposes of *Attract F*. Since *Attract F* has applied, *I* cannot be a syllable head. Under the assumptions of the X-bar account of syllable structure, a well-formed syllable must have an onset. Since *M* is licensed in coda position by the preceding syllable, it cannot be onset to *I*. *I*, therefore, being [*u*seg] and without an onset cannot project a syllable. Assuming that only segments in syllable structure constituents can be phonetically interpreted, the noninterpretation of *I* here is accounted for. This analysis is consistent with the derivation of (3a) *mita*, since the *I* of the suffix surfaces there only through adjunction to the adjacent stem vowel which has on onset and can therefore project a well formed syllable, as illustrated in (5), to yield *i*.

(5) $m[i,I]_i$ ta

But what then of (3c) *kašita*. Here, under the analysis in (4), we would expect *s* to be licensed in the coda of the first syllable which is [+CL] by attraction from *ITa*, and *I* to receive no phonetic interpretation yielding **kasta*, instead of *kašta*. Let us assume that *Attract F* in Japanese must also satisfy an identity condition on codas. That is, the *Attract F* chain must link to the feature [continuant] as well as the place features. In that case the derivation would look as follows for *kas+ITa*.

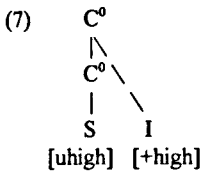
(6)



Here the mismatch in identity for the

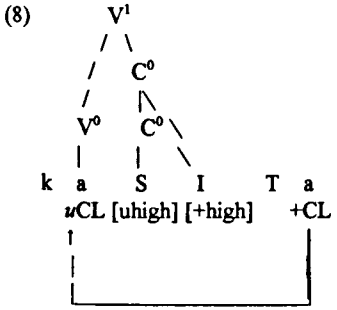
feature [continuant] blocks *Affect F* from interpreting the coda position. Instead *S* and *I* adjoin and the [*u*high] of *S* and the [+high] of *I* merge at the dominating C^0 adjunction node as illustrated in (7).

(7)



Notice, so far, *I* does not block the movement of consonantal features across it and can move features from itself to an adjacent consonant. Since it behaves in all respects like a consonant, let us assume that is what it is. Thus, the resulting cluster *ši* is interpreted as a voiceless consonant plus a voiceless vowel (or consonant) and can be interpreted in the licensed coda position of the preceding syllable, as illustrated in the derivation of (8).

(8)



CONCLUSION

An interesting result of this approach to consonant cluster phenomena is that Southern Paiute and Japanese look the same with respect to gemination. In each case, it is the [+CL] feature on the vowel of one syllable that licenses a coda position. This position is filled by a variable segment, one with unmarked features, that results in the attraction of features from the following onset. There are differences, of course, between Southern Paiute and Japanese. In the former the [+CL] property is affixed to every word, and because of the lengthy strings of affixes, is propagated across alternate syllables. In Japanese, [+CL] is a property of a small class of affixes, all of which have the property of licensing a coda in a preceding stem. Also, Japanese has an intervening abstract vowel *I* which emerges only if *Affect F* cannot related the consonants which surround it.

REFERENCES

[1] Chomsky, N. (1993), "A Minimalist Program." In K. Hale & S. Keyser (eds.), *The View from Building 20*. Cambridge: MIT Press.

[2] Bromberger, S. & M. Halle (1991), "Why Phonology is Different." In A. Kasher (ed.), *The Chomskyan Turn*. Oxford: Blackwell.

[3] Chomsky, N. & M. Halle (1968), *The Sound Pattern of English*. New York: Harper & Row.

[4] Sapir, E. (1933), "The Psychological Reality of Phonemes." In D. Mandelbaum (ed., 1949) *Selected Writings of Edward Sapir*. Berkeley: U of California Press.

THE VOICE SOURCE IN PROSODY

Gunnar Fant and Anita Kruckenberg
Dept of Speech communication and Music Acoustics, KTH, Stockholm, Sweden

ABSTRACT

The role of voice source parameters as acoustic correlates to prosodic categories have been studied experimentally and within the frame of production theory. Source parameters contribute more to shaping phrase group intensity profiles and to emphatic stress than to non-focal stressed/unstressed contrasts. The covariation of F_0 and source amplitude and source-tract interaction effects are important components of a generative theory.

INTRODUCTION

This is a progress report from an ongoing project on Swedish prosody [1]. One purpose is to discuss the role of the intensity parameter which in early Swedish phonology [2] was supposed to be crucial for the so called "Expiratory accent", while experience from the last decades of speech analysis and synthesis points at segmental durations and F_0-patterns as the primary acoustic correlates of stress in Swedish and English. A recent attempt to reconsider the relative salience of the intensity domain [3] has focused on spectral tilt as a more important parameter than overall intensity.

We shall also expand on the relation between F_0 and source parameters in the realisation of accentuation and attempt physiological and acoustic explanations. Source tract interaction phenomena [4] promote a production oriented view of articulatory effort as a component of prominence.

The intensity of a voiced sound has a complex relation to source and filter functions. The source may be extracted by inverse filtering, i.e. a process of removing the filter function from the sound. The glottal flow, thus recovered, is specified by an amplitude factor E_e and a few shape parameters. E_e is defined as the slope of the glottal flow at the closing discontinuity which usually coincides with the amplitude of the peak of the glottal flow derivative [5]. The most important shape parameter within the LF model of glottal flow [5-6] is the frequency F_a at which the source spectrum attains an additional -6dB/oct slope. An increase of emphasis is usually accompanied by an increase in both E_e and F_a. Formant amplitudes are proportional to E_e while an increase in F_a provides a relative gain at higher frequencies.

GLOBAL CONTOURS

The voice source has an important role in establishing groups and boundaries, e.g. shaping the onset, rises and declination within a phrase. From studies of prose reading we have found a typical phrase intensity contour with an initial rise lasting about 100 ms usually followed by a declination of about 4 dB per second and a more rapid decay in the last 400 ms ending with a final abduction gesture in prepause voiced segments or with a creaky voice termination at a voiced juncture. The main part of the contour may be level or show a rise-fall indicating a prominent focal domain in partial conformity with the F_0-contour. The declination of intensity is in part the automatic consequence of the F_0 dependency of E_e, in part a decline of about 3 dB per octave fall in F_0 at constant E_e and to a part the consequence of a decreasing lung pressure within the phrase. A final abduction gesture in the last 50-100 ms is associated with the main attributes of breathy voicing, e.g. increasing decline of the spectral slope (decreased F_a) and increased F1 bandwidth. Phrase junctures with continued voicing are often produced with creaky voicing accompanying the local F0 minimum.

Another aspect of the global contour is an alternation of rises and falls of overall voice intensity and tempo within a paragraph which adds an element of engagement.

STRESS AND ACCENTUATION

Our studies of prose reading have confirmed the relative subordninate role of intensity as a stress correlate [1]. Within a corpus of about 200 syllables

the average difference in intensity, defined as sound pressure level in a lowpass (LP) 1000 Hz band comparing stressed and unstressed vowels was 2.5 dB and about 1 dB higher values with highpass (HP) 1000 Hz measures. E_e measures were closely proportional to LP measures. No significant differences between phonemically short and long vowels were found.. Maximally open stressed vowels showed 1-2 dB higher LP and E_e values than maximally close vowels and also a greater stressed/unstressed contrast than more close vowels. The maximally constricted phase of long stressed [u:] [ʉ:] [i:] [y:] are not included in these data. They showed an additional weakening of the order of 3-6 dB.

Source amplitude and intensity play a greater role in emphatically stressed words. A separate study of specially constructed "lab sentences" contrasting in word accent types and in a variation of the place of focal accentuation showed a more substantial range of variation.

These are illustrated in Figures 1-2. As discussed in [1] the source amplitude E_e follows F_0 up to a critical frequency above which E_e tends to decrease. For the male subject, Figure 1, it was found at F_0=130 Hz. A closer analysis reveals a 1.7 power proportionality of E_e with respect to F_0 in the low frequency ascending branch. The break is especially apparent in the female voice, Figure 2 where the focal intonation peak overshoots the critical F_0=215 Hz causing a local intensity minimum. The asymmetry of the surrounding maxima suggests the presence of a larger subglottal pressure in the rising than in the falling part of the F_0 contour. The minimum is not always present. It was found in two out of 4 subjects and was occasionally missing for the subject AK of Figure 2. It can be less apparent in the intensity than in the E_e or in the speech waveform display, see Figure 1, which to some extent might be explained by the general relation of intensity being proportional to F_0 at constant F_0.

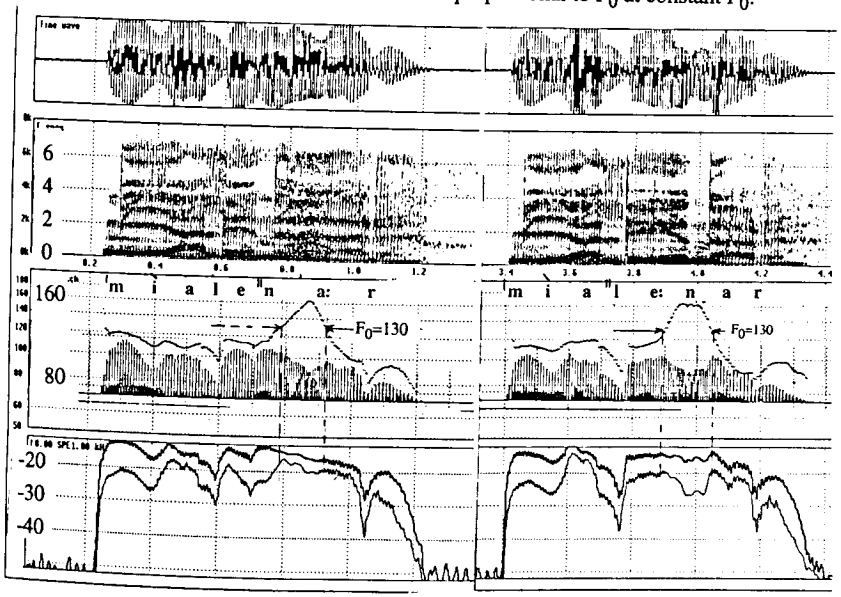


Figure 1. Male subject, contrasting stress location. E_e is superimposed on the F_0 display. The lower intensity curve is produced with high frequency preemphasis.

A study of the first and second syllables of the contrasting words "Lenár" [lenɑ:r] and "Lénar" [le:nar] both in focus showed large intensity effects for subject AK. The overall unweighted step in SPL from the first to the second vowel was -6 dB for "Lénar" and + 3dB for "Lenár".

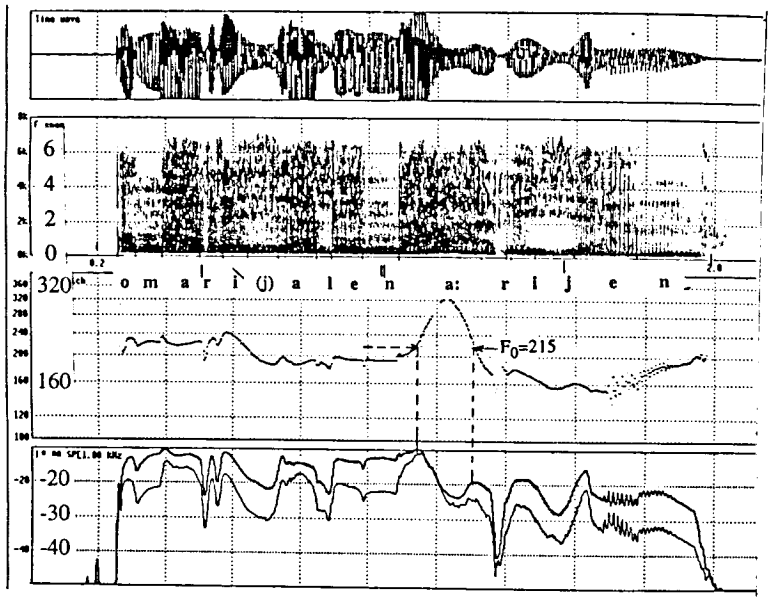


Figure 2. Female subject. Lexical stress on second syllable of "Lenár". Processing and display as in Figure 1.

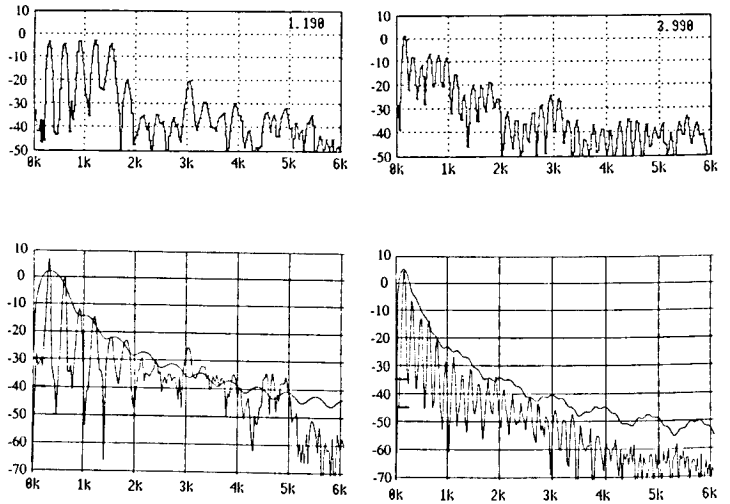


Figure 3. Spectra of stressed long [ɑ:] upper left from "Lenár" in Figure 2 and of unstressed [a] upper right from the contrasting "Lénar" together with their corresponding source spectra below. The envelopes in the source spectra pertain to a LF model match.

The corresponding high frequency preemphasized measures (+22 dB increase from 200 Hz to 5000 Hz) were -6 dB and +9dB respectively which reveals a significant high frequency gain in the stressed [ɑ:].

This is further illustrated in Figure 3 by the corresponding amplitude-frequency spectra of the two vowels and their source spectra derived from inverse filtering. They differ by a factor 2 in the F_a parameter which was 400 Hz for the stressed vowel and 200 Hz for the unstressed vowel. Subjects GF, on the other hand, produced much less intensity contrast, and even a reverse trend with the two contrasting words out of focus.

The intensity minimum in the long vowel [i:] of "Maria" in Figure 2 and "Mia". in Figure 1 is a typical instance of the articulatory gesture towards a [j] element in the middle of the vowel imposing a source filter interaction. The local dip in source amplitude E_e is substantially increased under influence of focal stress. Similarly, the E_e and the intensity minimum of the [*l*] in "Lénar" is enhanced and prolonged versus the [*l*] in "Lenár", see Figure 2, or versus the same word out of focus which enhances the consonant-vowel contrast at the onset of the stress gesture, i.e. at the P-center [7]. This is an additional stress correlate to consider [1].

DISCUSSION

Stressed syllables do not differ significantly from unstressed syllables in terms of source amplitude and high frequency contents unless produced with a marked emphasis which implies an increased subglottal pressure. This finding is coherent with the general conclusions in [8]. Continuously scaled measures of perceived prominence of syllables in prose reading have been correlated to duration and local F_0 measures [1] which both provided correlation coefficients of the order of r=0.9 to be compared to r=0.5 for recent studies of intensity and E_e.

The essential element of a focal versus nonfocal accentuation is the F_0 contour with or without significant increases in duration or in source properties. The individual variability is large and need to be related to the possible presence of a local subglottal pressure pulse and its synchrony with the glottal F_0 gesture and the stressed vowel onset.

Increase of stress also affects the consonant vowel articulatory and acoustic contrast not only in terms of formant patterns but also in terms of glottal source efficiency, i.e. a reduction of source amplitude and high frequency contents with increasing articulatory narrowing. This interaction may cause a narrow vowel to lose intensity at increased stress levels. Another instance of reversed stress intensity relation is when F_0 in focal accentuations overshoots a high critical value above which source amplitude and intensity is reduced unless backed up by an increasing subglottal pressure.

ACKNOWLEDGEMENT

This research has been supported by a grant from The Bank of Sweden Tercentenary Foundation.

REFERENCES

[1] Fant, G. and Kruckenberg, A. (1994), "Notes on stress and word accent in Swedish", *Proc. Int. Symp. on Prosody*, Sept. 18 1994, Yokohama. Also published in *STL-QPSR* 2-3 1994, pp.125-144.

[2] Elert, C.C (1968), *"Allmän och svensk fonetik"*, Almquist & Wiksell.

[3] Sluijter, A.M.C. and van Heuver, V.J. (1994), "Spectral balance as an acoustic correlate of linguistic stress". Manuscript submitted to *J.A.S.A.*

[4] Bickley, C.C. and Stevens, K.N. (1986), "Effects of a vocal tract constriction on the glottal source: Experimental and modelling studies", *Journal of Phonetics* 14, pp. 373-382.

[5] Fant, G., Liljencrants, J. & Lin, Q. (1985), "A four-parameter model of glottal flow", STL-QPSR 4/1985, pp. 1-13.

[6] Fant, G., Kruckenberg, A., Liljencrants J. & Båvegård, M. (1994), "Voice source parameters in continuous speech. Transformation of LF-parameters", *ICSLP-94*, Yokohama.

[7] Marcus, S.M. (1981), "Acoustic determinants of perceptual center (P-center) location", *Perception and Psychophysics* 30, pp. 247-256.

[8] Stevens, K.N (1994), "Prosodic influences on glottal waveform: Preliminary data", *Int. Symp. on Prosody*, Sept. 18 1994, Yokohama, pp. 53-63.

GLOTTAL LEAKAGE STUDIED BY MEANS OF SIMULTANEOUS VIDEO-STROBOSCOPY AND FLOW MEASUREMENT

B. Cranen and F. de Jong
Nijmegen University, Dept. of Language and Speech, Phonetics section
P.O. Box 9103, 6500 HD Nijmegen, The Netherlands

ABSTRACT

Laryngoscopic studies often report that a glottal leak is quite common in females and children but not in males. This seems in contradiction with the general finding that a dc-offset component in the glottal flow is normal for both females and males. In this study this phenomenon is examined in more detail. Stroboscopic images were recorded on videotape simultaneously with oral flow. The latter signal was used to estimate glottal flow waveforms by means of inverse filtering. We come to the conclusion that the glottal flow may have a dc-offset component even though the corresponding images do not reveal a leak.

INTRODUCTION

During the last few years the insight has grown that most people, both males and females, show a dc-offset flow when producing vowels at moderate loudness levels. From data presented in the literature it can be inferred that this dc-offset flow may attain values of 10–20% of the peak-to-peak value which, in our view, can only be explained by assuming a leak. In a recent paper the acoustic consequences of glottal leakage were discussed from a theoretical point of view [1]. In that paper the concept of a parallel leak (a leak completely separated from the time-varying membranous glottis, e.g. in the form of an opening in the cartilaginous portion of the glottis) was introduced to explain certain phenomena from an acoustic point of view. At the same time, though, it was recognized that the plausibility of such a parallel leak still needs to be proven. This paper presents a first exploration in that direction.

In various laryngoscopic studies it has been reported that a leak in the posterior commissure is quite common for females and children. However, for males, these studies invariably seem to suggest complete glottal closure. The latter is in contradiction with the finding that (for moderate loudness levels) not only the flow of females but also the flow of males generally contains a dc-offset component. This contradiction can in our view only be solved by assuming that (1) the laryngoscopic examination itself causes the subject to adopt a different phonation behavior with and without viewing equipment, (2) there exists a flow generating mechanism which does not require a leak (e.g. vertical motion of the folds), or that (3) the laryngoscopic view is not complete in the sense that not every possible leak is always visible.

EXPERIMENTAL SET-UP

In order to decide whether one of the aforementioned hypotheses is more likely than the other, we carried out a pilot experiment on two male subjects of which we collected video-stroboscopic image data of the vocal folds (stored on a video tape) *simultaneously* with measurements of the following signals

- oral flow measured by means of a circumferentially vented mask
- electroglottogram (EGG)
- microphone signal of the sound outside the mask (speech produced by the subject and the ENT clinician who handled the fiberscope commenting the images)

The image data, collected by means of a Kay LS9100 Rhino Laryngeal Stroboscope of which the stroboscopic flashlight was triggered by the EGG signal, were stored on a Super VHS video recorder. The microphone signal was fed to the audio channel of this VCR. All non-image signals (oral flow, EGG, and microphone signal) were stored on an FM recorder with a frequency response that was flat upto 5 kHz. Since the microphone signal comprising the subject's speech and comments of the ENT clinician was stored on both recording devices, this signal allowed (a rather crude) synchronization of the video images on the one hand, and the signals on the FM recorder on the other. However, for the stationary type of signals we analyzed in this pilot experiment (the vocal fold behavior generally did not significantly change over intervals of one or more seconds), we considered this sufficiently accurate.

In order to be able to collect image data simultaneously with oral flow data, we made a hole in the mask approximately at the level of the nostrils through which the endoscope could be inserted under an approximately horizontal angle. To ensure an airtight seal between mask and endoscope, we clamped a soft, flexible piece of silicone rubber with a hole slightly smaller in diameter than the endoscope, between the mask and a small metal plate. Both the mask and the metal plate contained a hole slightly larger in diameter than the endoscope. In order not to render the mask useless when the endoscope is not used, the hole can be closed by a metal plug of the right diameter.

Before the experiment started the flexible endoscope was fed some 15 cm through the hole in the mask (approximately the point where it needs to be to yield an adequate image when the mask is in place on the subjects face). Next the fiber (with the mask some 15 cm from the tip) was inserted via the nostril and gently pushed forward until the mask reached the subject's face so that he could push the mask firmly against his face. After the clinician adjusted the fiber's insertion depth to obtain a good quality image, the subject started to phonate steady vowels (/ae/) in as natural a fashion as possible.

The signals recorded on the FM-recorder were digitized at a sampling rate of 10 kHz (12 bit amplitude resolution) and stored on a digital computer. Next, the flow signals were calibrated, using the data of two calibration measurements: one set of calibration data was recorded at the beginning of the recording session and one at the end. Subsequently, each period of the flow signal was inverse filtered using formant estimates extracted from the closed glottis interval of that very period (using a pitch-synchrounous covariance LPC analysis). The analysis window (with a fixed duration of 3.6 ms) was chosen to begin at the the location of the peak in the EGG-derivative (after shifting this signal the appropriate amount of samples to account for the acoustic delay of the flow signal from glottis to sensor). The dc-offset flow value per glottal period was obtained by taking the minimum in a low-pass filtered version of the inverse filtered flow (cut-off at 1 kHz).

RESULTS

Since inverse filtering is less error prone for sounds with a high first formant, only recordings of the vowel /ae/

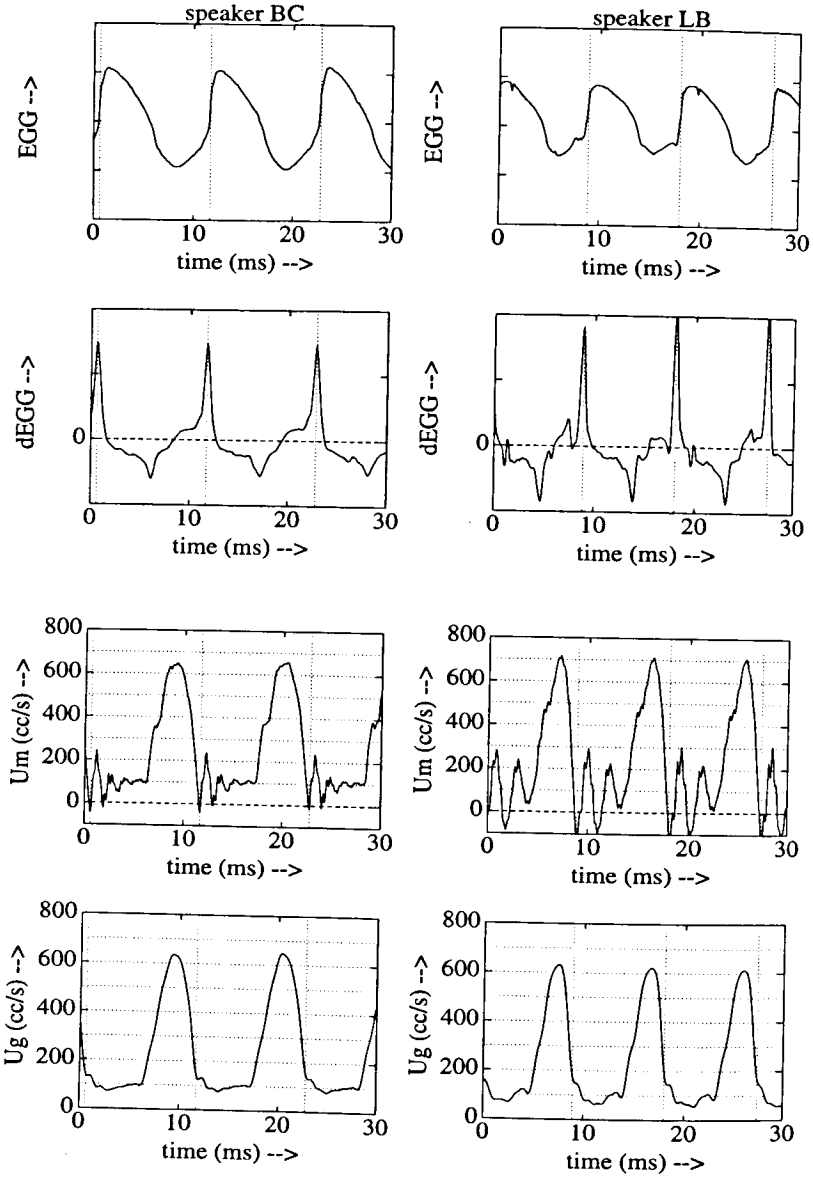


Figure 1: From top to bottom are shown the EGG, the derivative of the EGG, the oral flow signal and the inverse filtered flow signal [left: speaker BC (from segm #1); right: speaker LB (from segm #3)]. Vertical dotted lines denote the automatically detected moments of glottal closure. The videostroboscopic images that were collected simultaneously with these signals did *not* show any leak.

are considered here. From the recorded signals 3 segments per speaker of 1.5–3 seconds were selected during which, according to the information on the audio channel, glottal closure was complete. For each speaker one representative example is shown in figure 1. The corresponding image data on the video tape were inspected to verify that closure was indeed complete. For each segment, the average dc-offset (μ) and the standard deviation (σ^2) were calculated. The values are shown in table I (number of glottal periods in each segment is denoted by N).

Table I: dc-offset flow (cm^3/s)

	speaker BC			speaker LB		
segm	N	μ	σ^2	N	μ	σ^2
#1	167	77	51	140	135	19
#2	291	65	21	320	64	11
#3	183	56	20	334	57	20
total	641	66	32	794	73	33

Note that for speaker LB segment #1 comprises the /ae/ portions of a /paepae.../ utterance while the other two segments are from /ae/ vowels spoken in isolation. For speaker BC all three segments are from isolated /ae/ vowels. Unfortunately, with this speaker the fiber caused to much discomfort during /paepae.../ productions to warrant natural phonation.

DISCUSSION

Although the videostroboscopic images did not reveal any leak, both speakers show a non-negligible dc-offset value in the inverse filtered flow. For the isolated vowels this dc-offset amounts approximately 60 cm^3/s which is close to 10% of the peak value ($\approx$ 600 cm^3/s). For the vowel parts of the /paepae.../ utterance (LB: segment #1) we find almost double that value.

Regarding the hypotheses formulated in the introduction we come to the following conclusions. The fact that we observe a dc-offset flow even with an optic fiber in place suggests that it is not likely that phonation behavior is very different (at least not in a qualitative sense) whether viewing equipment is present or not. Thus, we have to explain our findings either by assuming that the laryngoscopic view is not complete in the sense that not every possible leak is always visible, or by assuming that a combination of squish flow and vertical motion of the folds is responsible for the extra flow.

The latter possibility is rather unlikely as well. In order to get a dc-offset flow of 60 cm^3/s during the entire closed glottis interval ($\approx$ 5 ms) requires that 0.3 cm^3 of air is displaced. If we estimate the area of the membranous part of the vocal folds that can take part in the vertical motion to be 0.6x1.5=0.9 cm^2 (which we think is an over-estimation), this would imply that the average vertical displacement of the folds during this interval would still have to be approximately 0.3 cm which seems irrealistically large.

As a consequence, we believe that the major part of the dc-offset flow is due to a leak. In order to account for a dc-offset value of 10% of the peak flow value, a leak opening is required which is approximately 10% of the maximum glottal opening. A reasonable estimate of the latter is 15–20 mm^2. Since it is very unlikely that an opening of 2 mm^2 would easily be overlooked, we are forced to postulate that such an opening does exist, but that the view on that opening is blocked by the arytenoids. In order to draw decisive conclusions, ways will have to be found to get reliable 3D images of the glottal region during phonation.

REFERENCES

[1] Cranen, B. and Schroeter, J. (1995) "Modeling a leaky glottis", *Journal of Phonetics*, vol. 23, pp. *unknown yet.*

SUPRALARYNGEAL RESONANCE AND GLOTTAL PULSE SHAPE AS CORRELATES OF STRESS AND ACCENT IN ENGLISH

Agaath M.C. Sluijter[1], Stefanie Shattuck-Hufnagel[2], Kenneth N. Stevens[2], Vincent J. van Heuven[1]

[1]Dept. Linguistics/Phonetics Lab., Leiden University/ Holland Institute of Generative Linguistics, The Netherlands
[2]M.I.T., Research Lab. of Electronics, Cambridge MA, USA

ABSTRACT

In a production experiment, it is shown that unstressed syllables have a smoother and slower vocal fold closing movement than stressed syllables. As a result the spectrum of stressed syllables is characterized by an increase in high-frequency emphasis. Accent, but not stress, is additionally characterized by a slightly increased open quotient and an increased amplitude of voicing.

INTRODUCTION

Sluijter and Van Heuven [1,2] showed that high-frequency emphasis is a powerful acoustical and perceptual correlate of linguistic stress in Dutch. They assumed that high-frequency emphasis arises because of the way the vocal folds and the glottis are configured during phonation when producing stressed syllables ("stressed" here refers to the main-stressed syllable of a word; "unstressed" means secondary stress or lower). Increased vocal effort, as is needed to produce stressed syllables, generates a more strongly asymmetrical glottal pulse, vocal fold abduction is faster, the maximum amplitude of movement following the opening time is greater and the closing phase gets shorter, so that the trailing flank of the glottal pulse is steeper. Certain of these differences are manifested in the spectrum at low frequencies, in the vicinity of the lowest three harmonics, whereas other differences modify the spectrum at mid and high frequencies. The low-frequency part of the spectrum is determined by the gross shape of the waveform. The relation between higher harmonics and the lowest few mainly depends on the speed of glottal closure. The faster the glottis closes, the more pulse-like the excitation signal will be, resulting in a harmonic spectrum with an increased tilt at high frequencies [3]. A more gradual pattern of glottal closure, as Sluijter and Van Heuven assumed to be the case for unstressed syllables, however, yields a steeper negative spectral slope (i.e. low-frequency emphasis).

In the present study we tried to replicate these results for American English, examining the possible cause of the high-frequency emphasis in more detail. Is it brought about by a change in the glottal pulse (e.g. longer glottal closure, faster vocal cord adduction) or by a change in the supralaryngeal tract (wider mouth opening, leading to an upshift of F1)? We investigated differences in the glottal vibration pattern for stressed and unstressed syllables, inferring glottal parameters from selected characteristics of the audio signal [3,4].

METHOD

Subjects, material and procedure

Six speakers of American English (three male, three female) produced four noun-verb minimal stress pairs, *'export~ex'port, 'uplift~up'lift, 'compact~com'pact, 'digest~di'gest* as well as their reiterant mimicries */bi:bi:/, /bɛbɛ/ and /ba:ba:/* with and without focal accent in fixed carrier phrases. In (1) an example is given of the condition with a pitch accent on the target (+F), in (2) without a pitch accent (-F), (target words in italics, accent position in bold face).

1. Please produce *'compact* for him again.
   Please produce *'baba* for him again.
   Please produce *'bibi* for him again.
   Please produce *'bɛbɛ* for him again.
2. Please produce *'compact* for him again.
   Please produce *'baba* for him again.
   Please produce *'bibi* for him again.
   Please produce *'bɛbɛ* for him again.

Each response type was recorded twice, the second time with the items in reversed order. This procedure yielded 640 utterances. Only the initial syllables of the lexical items, *export* and *uplift* and the final syllables of *digest* and *compact* were used for further analysis (where underline indicates the syllable that was analysed, in both its stressed and unstressed forms). Of the reiterant items, we will present only the data of the vowel */a:/* and */ɛ/*.

Measurements

a) Effects of the filter function of the vocal tract: To control for differences in the shape of the vocal tract due to stress and/or accent, we measured formant frequencies (F1..F3) of stressed and unstressed vowels. Only F1 was used in the present study; the difference, ΔF1, was calculated between F1 for each individual utterance and mean F1 across speakers and conditions, for each vowel (mean F1: */a/* 760 Hz, */ɛ/*: 605 Hz, lexical material: 637 Hz).

b) *The open quotient*, expressed as a percentage of the total period, determines the time during which the glottis is open. The primary acoustic manifestation of a narrow glottal pulse, i.e. a decrease in open time, is a reduction of the amplitude of the fundamental in the source spectrum relative to adjacent harmonics [5]. The amplitude difference between the first two harmonics (H1-H2), therefore, is an estimate of the open quotient (OQ). The stronger H1, the larger OQ [3]. H1 and H2 were corrected for the influence of F1, yielding the measure $H1^*$-$H2^*$ (see [3]).

c) Completeness of closure (bandwidth of F1): The amount of minimal flow (i.e. glottal leakage) varies over loudness conditions [6]. Louder voices tend to have a smaller minimal flow than soft voices. We therefore suggest that minimal flow also varies with stress. When the glottis is not closed during phonation, glottal resistance can contribute to energy losses in the F1 range, adding significantly to the F1 bandwidth (B1). B1 is estimated from the amplitude decay rate during the first two cycles of the F1 oscillation. To reduce interference by higher formants, the waveforms were filtered in a 600 Hz frequency band centered around F1.

d) Degree of opening over the entire glottal cycle: The amplitude of F1 (A1) depends on the degree of opening over the entire glottal cycle, i.e., A1 is influenced by OQ and the glottal aperture during the open phase, whereas B1 is not. We measured the difference between $H1^*$ and A1.

e) Skewness of glottal pulse, rate of closure: The spectrum of the glottal waveform at mid and high frequencies is influenced primarily by the abruptness of the glottal closure. There are two ways in which glottal closure can differ. If the closure is nonsimultaneous along the length of the vocal folds, there is a more gradual cutoff of airflow. A more abrupt closure introduces less negative spectral tilt in the higher frequency region. Another way in which glottal closure can differ is related to the duration of the closing portion, i.e. rate of closure (RC), which directly influences the skewness (SK) of the glottal pulse. As the slope of the closing phase gets steeper (keeping OQ constant) the amplitudes at mid and high frequencies increase relative to amplitudes at low frequencies. We derive information on the skewness of the glottal pulse and the rate of glottal closure by taking the difference between the amplitude values of the first harmonic ($H1^*$) and of F2 and F3 (A2 and A3). Both A2 and A3 are corrected for their dependence on F1, and F1 and F2, respectively, yielding the measures $H1^*$-$A2^*$, and $H1^*$-$A3^*$ respectively (see [3]).

f) Amplitude of voicing ($H1^*$): When intensity increases, subglottal airpressure will also increase, which directly influences amplitude of voicing. One of the main acoustic effects of an increase in subglottal airpressure is an increase in H1. We hypothesize that stressed and unstressed syllables differ in the amplitude of the glottal pulse.

g) Overall intensity: In addition we measured the overall intensity value of the stressed and unstressed syllables. We expect no differences between stressed and unstressed syllables in condition -F, whereas we do expect differences in +F.

An overview of the physiological

	Physiology	Acoustics
Filter	a) shape of vocal tract	F1, F2, F3
Source	b) open quotient (OQ)	$H1^*$-$H2^*$
	c) completeness of closure, glottal leakage	B1
	d) degree of glottal opening	$H1^*$-A1
	e) skewness of glottal pulse (SK), duration of closing portion (RC)	$H1^*$-$A2^*$ and $H1^*$-$A3^*$
	f) amplitude of voicing (AV)	$H1^*$

dimensions in which glottal pulses of stressed and unstressed syllables can differ, and the acoustic parameters from which these differences can be derived, is presented at the top of this page.

All measurements were made at the F1 maximum in each target syllable, i.e., when the mouth is maximally open. The resulting measures were averaged over speakers and over vowels, both reiterant and lexical. We compared the averaged values of stressed and unstressed vowels *paradigmatically* for each focus condition separately.

RESULTS AND CONCLUSIONS

Figure 1 shows the differences between stressed and unstressed syllables (solid and hatched bars, respectively) in condition +F (in focus, i.e. with a pitch accent) and condition -F (out of focus, i.e. without a pitch accent) of selected parameters. $H1^*$-A1, F2 and F3 were not significantly influenced by accent and/or stress, and are therefore not presented in the figure.

Results indicate that glottal pulses are more sinusoidal in unstressed syllables: high-frequency emphasis (SK and RC) is weaker, indicating smoother and slower vocal fold closing movement. Counter-intuitively, B1 was found to be wider for stressed than for unstressed vowels. This effect is even stronger for accented, stressed syllables. We assume that this effect is caused by the increased subglottal pressure with which stressed syllables are produced. Due to this pressure, the arytenoid cartilages remain abducted throughout the cycle, whereas the glottis is entirely closed over a part of the cycle of vibration when producing unstressed syllables. Accented stressed vowels are additionally characterized by an increased AV ($H1^*$) and a slightly increased OQ ($H1^*$-$H2^*$). The transfer function of the vocal tract differs only in the mouth opening dimension (F1) showing an overall tendency towards greater opening for stressed vowels, irrespective of accentuation.

We investigated to what extent the glottal shape parameters, intensity and ΔF1 by themselves could be used to discriminate stressed from unstressed, as well as accented from unaccented vowels in linear discriminant analyses across speakers, conditions and vowels. Table 1 gives an overview of the percentages correct classification.

Table 1 Correct classification (%) of stressed and unstressed syllables (stress), and of accented and unaccented syllables (accent) in condition +F and -F, and across conditions (all), using a supralaryngeal parameter (ΔF1), glottal parameters and overall intensity separately and in combination.

	accent	stress		
	all	+F	-F	all
1. ΔF1	68	69	64	65
2. glottal	88	90	73	75
3. 1+2	88	91	72	75
4. intensity	69	64	53	61
5. 1+2+4	88	90	74	76

Almost 90% correct classification of (pitch) accented syllables was reached using only glottal parameters as predictors. Intensity and ΔF1 are less powerful predictors by themselves; moreover adding them to the glottal parameters did not significantly improve the results.

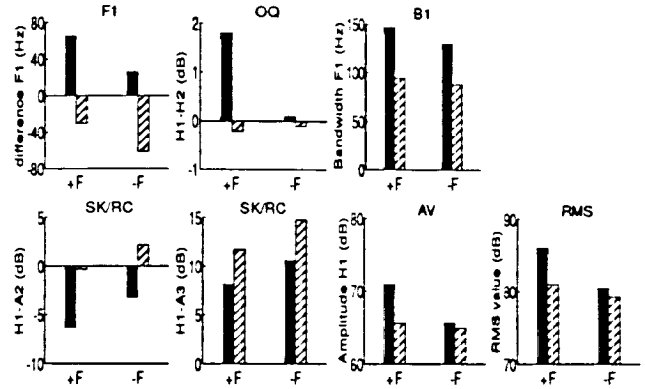


Figure 1. Effects of stressed (solid bars) and unstressed (hatched bars) syllables in condition +F (with pitch accent) and condition -F (no accent) on selected acoustic parameters (see text).

In condition [+F], percent correct classification of accented syllables is somewhat higher than in the above mentioned analyses. The relative strength of glottal versus other parameters, however, is virtually identical.

When separating stressed (both accented and unaccented) from unstressed syllables across conditions, almost 75% correct classification was reached. In this case, just as in condition [-F] (separating stressed, unaccented from unstressed syllables) especially the predictive strength of the glottal parameters decreases. Nevertheless, even in condition [-F], the contribution of glottal parameters always outweighs that of the other parameters, which means that glottal differences are the most reliable correlates of stress.

In this paper we studied the correlates of stress and accent other than F0 contour and duration, i.e. concentrating on the distribution of spectral energy. In this domain of correlates we conclude that accent and stress are mainly characterized by differences in the shape of the glottal pulse, rather than differences in the supralaryngeal tract.

Future research is needed to determine if stressed and unstressed syllables will sound more natural in synthesized speech if we also take aspects of vocal fold vibration and their effect on the spectral balance into account. Also, our results potentially contribute towards more accurate identification of accented and/or stressed syllables in automatic speech recognition systems.

REFERENCES

[1] Sluijter, A.M.C. and Heuven, V.J. van (1995). "Spectral balance as an acoustic correlate of linguistic stress," *J. Acoust. Soc. Am.* (submitted).

[2] Sluijter,A.M.C, Heuven, V.J. van and Pacilly, J.J.A. (1995)."Spectral balance as a cue in the perception of linguistic stress," *J. Acoust. Soc. Am.* (submitted)

[3] Stevens, K.N. and Hanson H.M. (1994) Classification of glottal vibration from acoustic measurements, in O. Fujimura, M. Hirano (eds) *Vocal fold physiology: Voice quality control*, San Diego: Singular, pp. 147-170.

[4] Fant, G., Liljencrants, J. and Lin, Q. (1985), "A four-parameter model of glottal flow", *Speech Trans. Lab. Q. Prog. Stat. Rep. 4*, Royal Institute of Technology, Stockholm, pp. 1-13.

[5] Klatt, D.H. and Klatt, L.C. (1990), "Analysis, synthesis, and perception of voice quality variations among female and male talkers", *J. Acoust. Soc. Am. 87*, pp. 820-857

[6] Holmberg, E.B., Hillmann, R.E., and Perkell, J.S. (1988), "Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal and loud voice", *J. Acoust. Soc. Am. 84*, pp. 511-529.

INTERACTIVE VOICE SOURCE MODELLING

Mats Båvegård and Gunnar Fant
Dept. of Speech Communication and Music Acoustics,
KTH, Stockholm, Sweden

ABSTRACT

Interaction ripple is caused by the non-linear relation between glottal flow and trans-glottal pressure, the latter containing components of vocal tract oscillatory modes evoked during the past history of the phonatory process.

This study deals with the non-linear transformation from glottal area function to glottal flow, with and without constant leakage invoked by a glottal chink. Earlier observations on a certain high frequency boost associated with the glottal chink [1], [2], have been verified. A separate study of the audibility of interaction ripple has been undertaken.

INTRODUCTION

Interaction ripple is the superposition of quasi-random variations within a voice cycle of the glottal flow which is commonly observed in inverse filtering of the speech signal. A number of aerodynamic-acoustic model simulations have revealed the general mechanism causing interaction ripple, i.e. [2] [3].

The origin is a non-linear perturbation of glottal flow caused by vocal tract oscillatory modes superimposed on the trans-glottal pressure.

The term interaction in a broader sense involves all aspect of a complete supra- and sub-glottal coupling and associated departures from normal phonation. In breathy phonation the coupling causes shifts of formant frequencies and bandwidths and the appearance of sub-glottal formants and noise. These can be studied in a complete articulatory analog.

FLOW SIMULATIONS

The simulations have been performed with our articulatory speech synthesiser, TRACTTALK [2]. The synthesiser consists of three parts, a sub-glottal system, a model of the time varying glottal opening area modulating the trans-glottal impedance and a supra-glottal system.

The sub-glottal system incorporates three resonance modes and the glottal impedance contains the complete inventory of inductance, kinetic resistance and frictional resistance. The examples Fig. 1-3 pertains to a supra-glottal configuration of a vowel [ɑ:]. For computational details see [4]. The LF-model was used to control the glottal area function $A_g(t)$. The effects of adding a constant leak are also demonstrated in the three examples, Fig 1-3.

Results

In the zero glottal leak simulation, curve 1 in both Fig 1 and Fig 2, there is an apparent double peak in the positive part of $U_g'(t)$ which is a typical instance of interaction ripple as seen in true speech, similar to what has been discussed in [2]. The main cause of the double peak is the F1 oscillatory component of the trans-glottal pressure. The spectral consequence of a valley and peak around 800 Hz is rather weak. It should be kept in mind that the specific shape of the interaction ripple is highly dependent on the duration of the voice fundamental period which determines the particular phase of the previously excited component upon arrival in the beginning of the next open phase.

Fig. 1 is an example of adding a constant leak to a glottal area function, with return time T_a=0.1 ms. Three values of glottal leak have been simulated: no leak A_{go}= 0, A_{go}= 0.03 cm^2, A_{go}= 0.075 cm^2. The glottal flow derivative $U_g'(t)$ for zero leak displays the typical double peak, which is smoothed out and disappears with increasing constant leak. This is to be expected because of the increased damping and thus the low carry-over of F1 oscillation from the previous period. At the same time we observe an irregularity in the return phase of $U_g'(t)$, a phenomena which has frequently been observed in inverse filtering of real speech and in the simulations, [1], [5]. The mere presence of the leak appears to be sufficient as an explanation.

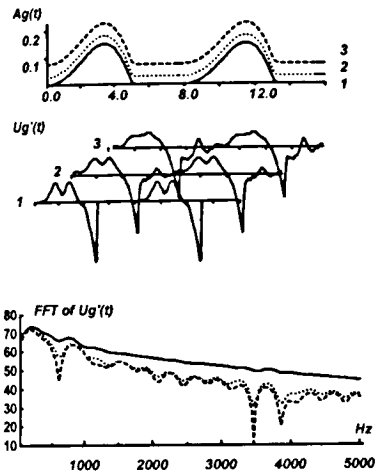


Fig 1. From the top: Glottal area function $A_g(t)$ with a small return time T_a = 0.1 ms and three values of a constant leak. T_0 = 8.0, T_e = 5.0, T_p = 3.5 ms, the resulting differentiated glottal flow and at the bottom the spectrum of the differentiated glottal flow.
Curve 1: without constant leak,
2: constant leak = 0.03 cm^2 included,
3: constant leak = 0.075 cm^2 included.

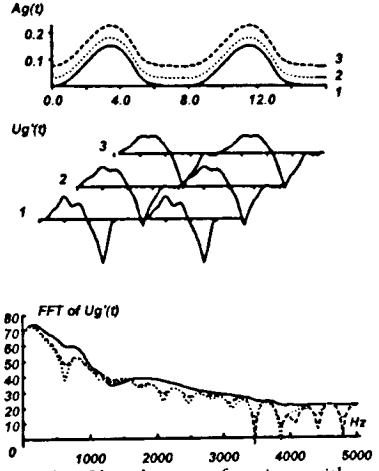


Fig 2. Glottal area function with a moderate return time T_a = 0.5 ms, else specifications identical as in Fig 1.

Another typical effect is the increase in the effective return time T_a in the flow compared to the T_a of the underlying area function. The T_a=0.1 ms of $A_g(t)$ corresponds to a critical frequency F_a= $1/(2\pi T_a)$= 1600 Hz. For the no leak case we observe a T_a corresponding to approximately F_a = 850 Hz, with the small leak F_a=700 Hz and the larger leak F_a= 650 Hz. The corresponding increase in spectral tilt associated with leakage is apparent in the FFT spectrum.

With increasing T_a of the glottal area function and the presence of a constant leak there is a non-uniform shift in the source slope. This is demonstrated in Fig. 2, where T_a of $A_g(t)$ is 0.5 ms. The corresponding T_a of $U_g'(t)$ in Fig. 2 is 0.6 ms with no leak, 1.2 ms for the small leak of 0.03 cm^2 and 1.4 ms for the larger leak of 0.075 cm^2 equivalent to F_a= $1/(2\pi T_a)$ values of 256 Hz, 130 Hz and 115 Hz respectively. The presence of a leak thus causes approximately a doubling of T_a and an octave lowering of F_a. Thus, with a leak present the spectral tilt sets in at a lower frequency. However, instead of a uniform slope there is a recovery above 1000 Hz which restores the spectrum level to nearly the same as the no leak case at 3000 Hz. An additional effect of the constant leakage is a reduction of the maximum negative value of $U_g'(t)$, i.e. E_e by about 3 dB.

These effects are even more apparent in Fig. 3 where T_a=1 ms in the glottal area function. Above 1700 Hz and with a leak the spectral level is restored to about the same or higher than without leak in spite of the fact that the higher T_a (6 dB) and the lower E_e (2 dB) together would have produced a 8 dB lower level in this range. Because of the higher T_a (lower F_a) the presence of a leak causes a steeper spectral slope up to 1000 Hz followed by the restoration. The magnitude is of order 10 dB here.

These observations supporting the earlier findings of [1] and [2], have implication for the overall spectral characteristics of the female voice.

To what extent are these phenomena dependent on the sub-glottal system?

We repeated the experiment in Fig. 3 but with the sub-glottal system short-circuited. The E_e value is not substantially reduced by the introduction of the leakage. The higher E_e without the sub-glottal system is partially compensated by a lower T_a of $U_g'(t)$ and the spectral difference becomes small [4].

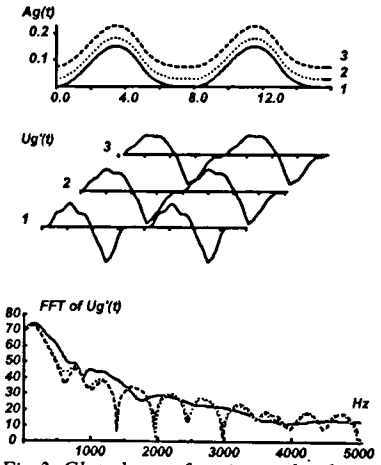


Fig 3. Glottal area function with a large return time T_a = 1.0 ms, else specifications identical as in Fig 1.

PERCEPTUAL TEST

When selecting stimuli for a perceptual test we avoided glottal configurations that produce more complicated interactions. We thus excluded the superimposed constant glottal leak and choose parameters typical for a well developed male voice with normal time constants of the glottal flow return phase. One of the test stimuli was the sentence "Ja adjö", ['jɑ:-a'jø:]. In addition we produced isolated vowels [ɑ:] and [ø:] of about 400 ms duration. These were produced with two different modes of temporal modulation, one with varying F_0 and waveform parameters approximating natural speech, the second one with constant F_0 and a moderately varying amplitude profile at onset and offset which was expected to enhance the perceptibility of the ripple. Each of these three stimuli types,

(1) Vowels with time varying F_0 and waveshape parameters.
(2) Vowels with constant F_0 and waveshape parameters.
(3) The complete sentence, "Ja-adjö".

were compared to synthetic versions produced with optimal LF-source in non-interactive formant synthesis.

Test procedure

The main part of the study was devoted to discriminabilities in AB and ABX tests. The question posed in the AB test was "which do you prefer, A or B?". The ABX-test posed the standard question: "Is X more like A than B?". In addition we ran tests aiming at a categorical rating of perceived differences, within contrasting stimulus pairs of full representations and LF versions.

Results

The AB-tests, Fig 4, show that the majority of the listeners preferred the interactive source in all vowels and in the sentence "Ja-adjö". For vowels the average score was of the order of 70%, i.e. well above chance level, and not significantly different for the time varying and constant parameter settings. In the sentence test the score was higher, 85% which was to be expected in view of other shortcomings in the overall matching to the human reference.

The ABX test, Fig 5, supported the general findings from the AB test. Apart from the low discriminability of the vowel [ɑ:] which could be explained by its initial placement in the test sequence without previous training, the tendencies appear similar to those of the AB test and with higher test scores.

The perceived difference test supported the view that the differences in interactive and LF synthesis are small or very small. Only 13% of the votes were in the category of a large difference. In the sentence test, on the other hand, as much as 55% votes were in the large difference category [4]. However the "large" assignment is a relational rather

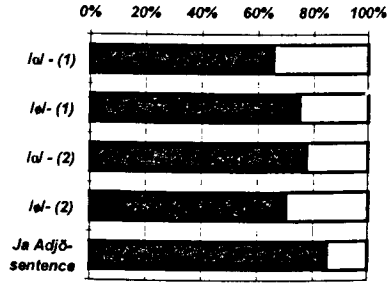


Fig 4. The AB-test, (1) Vowels with time varying F_0 and waveshape parameters. (2) Vowels with constant F_0 and waveshape parameters.

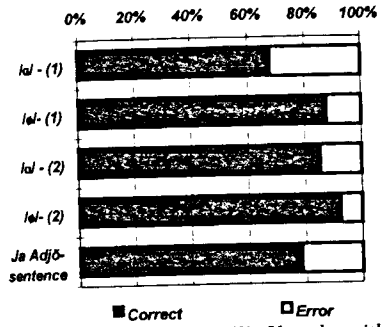


Fig 5. The ABX-test, (1) Vowels with time varying F_0 and waveshape parameters. (2) Vowels with constant F_0 and waveshape parameters.

than an absolute judgement. The synthetic version is a rather good approximation to the human utterance. This overall impression is supported by the no difference votes of approximately 20% in the ABX test.

CONCLUSIONS

Interaction ripple is a non-linear perturbation superimposed on glottal flow. It originates from residuals of vocal tract oscillatory modes, largely those evoked in the preceding fundamental period, adding to the instaneous value of the trans-glottal pressure drop which has a second power relation to glottal flow. The occurrence of ripple does not require a glottal leak or the influence of the sub-glottal system. On the contrary a constant leak adds to the damping of formant oscillations and thus smoothes out the ripple as is encountered in phrase final abduction.

One recurrent observation is that the return time T_a of the glottal flow tends to be larger than the T_a of the glottal area function and that it increases with increasing glottal leakage. A glottal leakage also reduced the excitation amplitude E_e but only when the sub-glottal network is retained.

We have verified the tendency observed in [1] and [2] that a combination of a constant leakage as with a glottal chink, and a finite T_a of the glottal area function causes a high frequency boost that partially inhibits the spectral drop above 1000 Hz. This effect is of the order 6-12 dB at 2500 Hz. It is somewhat reduced when the sub-glottal impedance is short-circuited. This phenomena has implications for the interpretation of female voice source spectra. The presence of a glottal chink may also induce an irregularity in the closing phase as earlier observed in [1] and [5].

Our ABX test provided data on the detectability of interaction ripple while the preference was judged by an AB test. The main outcome is that the presence of interaction ripple adds a weak but detectable quality which is prefered by a majority of the listeners. This was found in both stimulus category (1) and category (2), as well as for the sentence "Ja-adjö". Although the synthetic version of the sentence was a fairly good replica of the spoken version the perceived difference was judged to be substantial.

Our perceptual studies have been directed to one aspect of interactive synthesis only, that of the ripple. A more important object for future research will be to gather experience of quality gains associated with complete articulatory synthesis and thus a more complex source-filter interaction. For this purpose we are now considering a more flexible glottal area function than the LF-model which initially was intended for glottal flow only.

ACKNOWLEDGEMENTS

This work has been funded by ESPRIT/BR project 6975, SPEECH-MAPS, in part financed by NUTEK

REFERENCES

[1] Cranen, B. and Schroeter, J. (1993): "Modelling a leaky glottis". *Proc. Dept. of Language and Speech* 16/17, University of Nijmegen. pp 56-64.
[2] Lin, Q. (1990): "*Speech Production Theory and Articulatory Speech Synthesis*", Ph.D. Thesis, Dept. Speech Com. and Music Acoust., KTH, Stockholm.
[3] Fant, G. (1986): "Glottal flow: models and interaction," *J of Phonetics*, 14 Nos (3/4), pp.393-399.
[4] Båvegård M., Fant G. (1994), "Notes on voice source interaction ripple", *STL-QPSR* 4/1994, pp 63-77.
[5] Karlsson I., Liljencrants, J. (1994): "Wrestling the two mass model to conform with real glottal wave forms", Proceedings ICSLP'94, Yokohama, pp 151-154

PHYSIOLOGICAL CORRELATES OF GLOBAL AND LOCAL PITCH RANGE VARIATION IN THE PRODUCTION OF HIGH TONES IN ENGLISH

Mary E. Beckman[a,b], Donna Erickson[c,d], Kiyoshi Honda[d], Hiroyuki Hirai[d], and Seiji Niimi[e]

[a]Ohio State University, Linguistics, Columbus, Ohio, USA
[b]ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan
[c]Ohio State University, Speech & Hearing Sciences, Columbus, Ohio, USA
[d]ATR Human Information Processing Research Laboratories, Kyoto, Japan
[e]Research Institute of Logopedics and Phoniatrics, University of Tokyo, Japan

ABSTRACT

Subglottal pressure and cricothyroid muscle activity level were measured to explore the physiological bases of the systematic differences in measured F0 peak value that previous studies have shown for high tones in English pitch accents. There was a strong correspondence between EMG level and F0 variation associated with pitch accent type, but subglottal pressure differences were associated primarily with the overall loudness of the utterance.

INTRODUCTION

The modeling of F0 patterns in large sets of utterances elicited under controlled conditions of variation in intonational function and related prosodic structure has made [+Hightone] ("H") one of the best understood phonetic features. Such modeling studies have shown that the relative F0 target values associated with H tones are sensitive to differences in the tones' paradigmatic intonational categories, their position in the sentence or larger discourse segment, and the prominence of the associated word or phrase. Moreover, these relationships are preserved across global variation such as the overall F0 increase associated with a louder voice. For example, Liberman & Pierrehumbert [1] studied nuclear H* pitch accents in two successive intonational phrases in an English declarative sentence produced many times in each of ten different degrees of overall vocal effort and in the context of two dialogues which contrasted the relative informational prominence of the accented words. When the accented word in the second phrase was the more prominent "answer" focus, the target F0 peak for its H* pitch accent was slightly higher than that for the first, but when it was the weaker "background" focus (which repeated information in the context), its peak was substantially lower. These relationships could be modeled by two effects — a "final lowering" of the second peak and a proportional raising of the "answer" peak — both expressed as constant distances above a "reference line" for the variation in the utterance's global pitch range as the speaker increased or decreased overall loudness. Analogous patterns have been found in other similar studies of English and other languages (e.g. [2, 3, 4])

This paper examines the physiological bases of such fundamental frequency relationships for English. We recorded subglottal pressure and electromyographic activity level from the cricothyroid muscle as a talker produced a sentence in three different intonation contours in which we could examine F0 peaks associated with H tones exemplifying prosodic contrasts that have been well-studied in the experimental literature. The talker produced dozens of tokens of each intonation type in each of three overall vocal effort levels. The results suggest that the talker adjusted cricothyroid activity level to control the variation associated with simple versus rising nuclear pitch accents, but the variation in peak height associated with global pitch range seems to originate primarily in the differences in subglottal pressure associated with soft versus normal versus loud voice.

METHOD

The data were originally recorded as part of our ongoing examination of L(ow) tones in English intonation [5, 6]. We re-examined the three intonation contours listed in Table 1, which had two F0 peaks for pitch accents varying in accent type and in associated qualitative stress level. Contours (2) and (3) have the same stress pattern, and differ only in the nuclear accent type, whereas contour (1) has two nuclear pitch accents in successive phrases. (See [7] for our analysis of the qualitatively stronger and weaker stress levels corresponding to nuclear versus prenuclear accentual position within an intonational phrase, and [8] for the inventory of phrase tone and accent types and their meanings.) The talker was the first author, a female native speaker of American English, who could produce the contours consistently many times at each of soft, normal, and loud voice.

Table 1. Discourse contexts and intonation contours, with target tones underlined.

1. Two phrases demarcated by a L- phrase accent, with H* nuclear accent in each.

Do you have any pasta dishes less fattening than fettucine Alfredo?
We have a lean, mini-noodle dish.
H* L- H* L-L%

2. Single intonational phrase, with L*+H type in nuclear accent position, in the canonical contrastive contour.

Do you have any bean dishes other than this couscous thing?
We have a lean mini-noodle with beans.
H* L+H* L-L%

3. Single intonational phrase, with scooped L*+H type in nuclear accent position, in a contour indicating pragmatic uncertainty.

Do all of your rice dishes have this fatty meat sauce?
We have a lean mini-noodle dish.
H* L*+H L-H%

Subglottal pressure (Ps) was recorded from a transducer mounted at the tip of a catheter inserted through the nose, and cricothyroid muscle activity level (CT) was recorded in a separate session, using subcutaneous hooked-wire electrodes. We analyzed about 20 tokens per cell for the Ps dataset and about 24 for the CT dataset. (See [5] for more details concerning recording methods, smoothing, and so on.) There was always a clear peak in the F0 contour and in the smoothed Ps contour that we could measure for each of the two target H tones. The smoothed CT data, however, often showed a series of peaks during the F0 rise, so instead of choosing any single maximum value, we instead took a measure of the average value over the F0 rise, from the minimum in the [h] in *have* (for the first H tone) or from the minimum at the preceding L target (for the second H tone).

RESULTS AND DISCUSSION

Figure 1 plots the value for the second F0 peak as a function of that for the first F0 peak in the Ps dataset. The datapoints for contour (1) cluster just below the x=y line, indicating that the second peak is lowered somewhat relative to the first. Compared to the analogous plots in [1], the lowering is substantially less than it would be if the second phrase were subordinated to the first in an answer-background focus sequence, suggesting that the data in Figure 1 show a pure "final lowering" effect. The datapoints for contours (2) and (3), by contrast, show a second peak that is substantially higher than the first. (The scatterplot for the CT dataset is virtually the same except for a slightly smaller dynamic range over the three voice effort levels.)

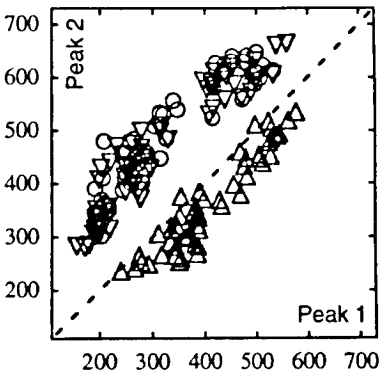


Figure 1. F0 value in second peak as a function of that in first peak for contours (1) Δ, (2) O, and (3) ∇. Values for both axes in Hz.

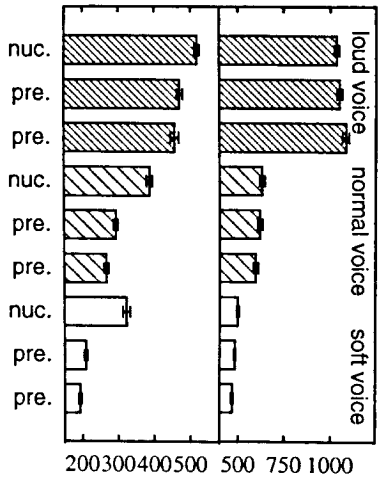


Figure 2. Means for F0 peak (left, Hz) and Ps peak (right, arbitrary units) for the first pitch accent in the Ps dataset, averaged over vocal effort level and over contour, with types (1), (2), and (3) in descending order along the y-axis.

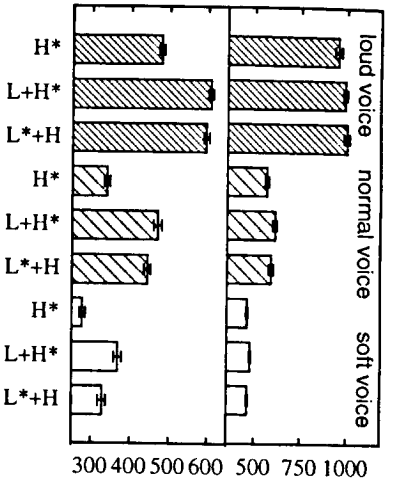


Figure 3. Means for F0 peak (left) and Ps peak (right) for the second pitch accent, arranged as in Figure 2.

Figures 2-3 give mean values for these F0 peaks and for the associated Ps peaks. They show that the difference among contours in Figure 1 is a function both of the second peaks being higher in contours (2) and (3) than in contour (1), and of the first peaks being lower. The higher second peaks in contours (2) and (3) suggests an inherently greater tonal prominence for the rising accents than for the single-tone accent. This is in keeping with the accents' meanings. Both L+H* and L*+H differ from simple H* in explicitly contrasting the focused discourse entity to other values in the presupposition set [8]. The lower first peaks in types (2) and (3) cannot be attributed to accent type, since all three contours have H* here, but it is in keeping with the different associated stress levels. A prenuclear accent is less prominent than a nuclear accent.

The corresponding Ps values in the right-hand panels of the figures show differences in peak Ps means that correspond well to the differences in the F0 means within a loudness level, differences that we have related to the tonal prominence inherent to the accent type (Figure 4) or to the associated stress level of that prosodic position (Figure 3). However, the correspondence is not completely consistent. In the loud-voice productions, the prenuclear H* of contour (3) has a higher mean Ps value than the nuclear H* of type (1). Moreover, the variation among accent types and stress levels is very small by comparison to the enormous increase in Ps in going from soft to normal to loud voice.

Figures 4-5 show the analogous mean values for F0 peak and average CT over the F0 rise for the productions in the CT dataset. The F0 peaks showed the same pattern of effects as in the other dataset, but the CT values showed two different patterns, neither exactly like the pattern of effects for the Ps values. For the first accent peak, the average CT value over the F0 reflected both the F0 variation due to overall pitch range and the variation within a pitch range due to local accentual prominence (although the correspondence to the latter was best at soft voice). For the second peak, however, there was only the variation due to accent type, and virtually no difference in means corresponding to the differences in F0

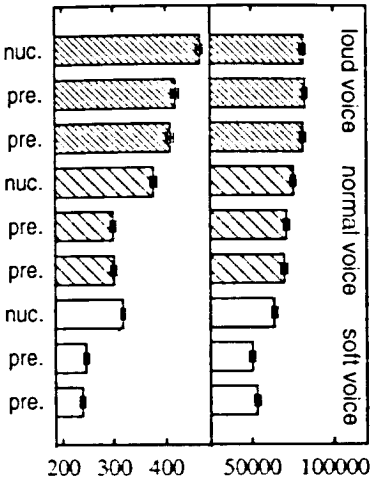


Figure 4. Means for F0 peak (left, Hz) and average CT over the F0 rise into the peak (right, arbitrary units) for the first pitch accent in the CT dataset, contour types as in Figure 2.

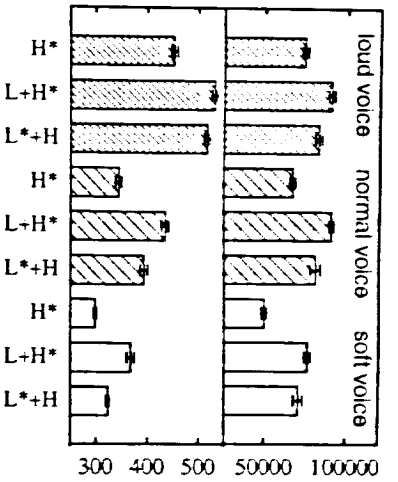


Figure 5. Means for F0 peak (left) and average CT over the F0 rise (right) into the second pitch accent, as in Figure 4.

values across the three overall vocal effort levels.

Although further experimentation is necessary to understand the pattern of CT activity for the first F0 peak, differences in global pitch range for different overall vocal efforts seems to be related primarily to the different associated subglottal pressures. By contrast, the pattern of CT activity for the second F0 peak suggests that at least some aspects of local pitch control can be differentiated physiologically from the more global effect.

REFERENCES

[1] Liberman, M., & Pierrehumbert, J. (1984). "Intonational invariance under changes in pitch range and length," In M. Aranoff & R. Oehrle, eds, *Language Sound Structure*, Cambridge, MA: MIT Press.

[2] Bruce, G. (1982). "Developing the Swedish intonational model," *Working Papers, Lund*, no. 22, pp. 51-116.

[3] van den Berg, R., Gussenhoven, C., & Rietveld, T. (1992). "Downstep in Dutch: implications for a model," In G. J. Docherty & D. R. Ladd, eds, *Papers in Laboratory Phonology II: Segment, Gesture, Prosody*, Cambridge, UK: Cambridge University Press.

[4] Pierrehumbert, J. B., & Beckman, M. E. (1988). *Japanese Tone Structure*, Cambridge, MA: MIT Press.

[5] Erickson, D., Honda, K., Hirai, H., Beckman, M. E., & Niimi, S. (1994). "Global pitch range and the production of low tones in English intonation," *ICSLP'94*, vol. 2, pp. 651-654.

[6] Erickson, D., Honda, K., Hirai, H., & Beckman, M. E. (in press). "The production of low tones in English intonation," To appear in *Journal of Phonetics*.

[7] Beckman, M. E., & Edwards, J. (1994). "Articulatory evidence for differentiating stress categories," In P. A. Keating, ed., *Phonological Structure and Phonetic Form: Papers in Laboratory Phonology III*. Cambridge, UK: Cambridge University Press.

[8] Pierrehumbert, J. & Hirschberg, J. (1990). "The meaning of intonation contours in the interpretation of discourse," In P. R. Cohen, J. Morgan, & M. E. Pollack, eds, *Intentions in Communication*, pp. 271-311. Cambridge, MA: MIT Press.

THE PROSODIC VARIABILITY OF SPEECH IN A NOISY CONTEXT

N. Kadiri, N. Vigouroux, G. Pérennou
Université Paul Sabatier, IRIT ,118, Route de Narbonne, F- 31062 Toulouse Cedex

ABSTRACT

This paper reports on the prosodic variability (phoneme and word duration, fundamental, vocalic release frequency) which occurs in quiet and noisy conditions, with the presence or not of feedback. The study revealed that: 1) the type of noisy condition and the presence or not of feedback changes the prosodic parameters, 2) theses variations are also speaker-dependent, and 3) the vocalic release frequency observed is also function of the nature of the noise/feedback.

INTRODUCTION

In the presence of noise, speech is masked and its production is modified by what is called the Lombard effect [1]. Lane and Tranel [2] showed that the speaker modifies his speech production while speaking in a noisy environment. In particular this study illustrated:

- phonetic cues and features for normal and Lombard speech change,
- prosodic parameters: not only does the loudness (intensity) of the speech increase but also the fundamental frequency and the speech rate.

Such observations are also reported by different authors [3], [4], [5], [6] and [7].

The Lombard effect has been neglected in speech recognition systems until a recent date [4]. In several experimental studies Furui [7] reported that the Lombard effects have a greater effect on speech recognition than does the direct influence of noise entering by microphones, for example.

This paper briefly reports our analysis results observed on the French database BD-BRUIT [9]. Firstly, we will show how the type of noisy conditions and the presence or not of auditory feedback influences the prosodic parameters (phonemes and word duration, pitch) changes occurring in Lombard speech. We will also discuss the vocalic release frequency observed on the corpora as a function of noise condition, speaker and type of phonological endings.

EXPERIMENTAL DATA

The corpora of our study are issued from BD_BRUIT data base [9]. The recordings have been done under five conditions of noise:

- the reference condition without noise, noted REF,
- white noise without audio feedback, noted BB,
- white noise and audio feedback, noted BB_ret,
- cocktail party noise without feedback, noted Amb,
- cocktail party noise and audio feeback, noted Amb_ret.

Five men and five women have pronounced five times the corpus of ten isolated digits from zero to nine, in the five conditions described above. We have carried out two hand-labelling levels: one in phoneme units (1) and the other in word units (2), both on a temporal and a spectral representation of the speech signal.

The labelling (1) was carried out for two repetitions per speaker and per condition, e.g. a total of 4700 phonemes (47 phonemes per repetition).

The labelling (2) has been made for the complete speech data base files (e.g. 2500). At the end of words, a vocalic release appears sometimes. It can be interpreted as the realization of an underlying schwa such as in the word "quatre" [katrə]. In other cases, it is a release appearing with a final obstruent like in the word "six" [sis]. These releases have not been considered as being a part of the final obstruent (plosive or fricative) which precedes them for the duration of phonemes.

METHODS: VARIANCE ANALYSIS

In order to better explain the set of available observations, we have carried out an analysis of the variance which consists in explaining a quantitative variable by a set of qualitative variables called factors. The phoneme and word duration, the fundamental frequency and vocalic release frequency were analyzed by separate repeated analysis of variances (ANOVAs). Noise condition (noted Noise), the speakers (noted Loc) and the auditory feeback (noted Ret) were the factors. This analysis of variances (ANOVA) has been done with "PROC GLM" of the SAS software [10].

The test consists in observing the probability that a "law of Fisher" with the appropriated degrees of freedom exceeds the statistic of Fisher "F value". If we decide to do the test at 5% (respectively at 1%), we say that the variable has an effect if the probability is inferior to 0.05 (respectively to 0.01).

We have preferred to explain the increase percentage of this parameter in comparison with the mean value of it in the reference condition.

The size of this article does not permit to present all results. So, we will report the most significant results.

DURATION OF PHONEMES AND WORDS

PHONEMES' DURATIONS

As suggested by the results of previous works [3], [5] to measure the noise effect on the duration of speech segments, we define two phonetic classes: vocalic and fricatives. The plosives were not considered because the difficulty to define a robust rule about the status of the consonantic release at the end of words like in /sɛ̃k/ or /sɛt/.

Variation of vowels' duration

The variance analysis shows:

- a significant increase of vowels duration in the presence of noise on average with 21% (F Value = 12.45, p<0.0001),
- the significant effect of the speaker (F Value = 9,56, p <0.0001)
- on the other hand, the type of noise and the auditory feedback of speech are not significant.

These results are illustrated in the figure 1 which shows the duration increase in comparison with reference according to the noise conditions for all speakers. We will notify that the type of noise has an important influence for two speakers (130 for the Amb condition for one of speakers 29 on average for the other speakers).

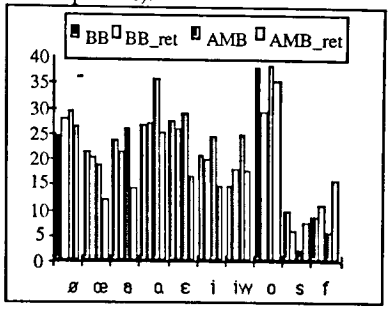


Figure 1: Increase average of the phonemes' duration according the noise conditions (phonemes in API, the code "iw" represents a labialized /i/.

Variation of fricatives' duration

For fricatives' duration, no effect was significant for speech in noise. On the other hand, the increases of duration are significantly affected only by the speaker (F Value = 6.12, p<0.0001). This result is not totally in agreement with other observations, where the consonant durations are decreasing by the effect of noise [7].

The type of noise has some significant effect on the duration of phonemes but these effects are not confirmed for consonantic phonemes in isolated speech. It seems that the speaker factor has also an influence on the phoneme variation.

WORDS' DURATION

The results of variance analysis

The computation of words'duration raises a question: how to determine the end boundary of the word. We have decided to take into account all the words of the corpora, whether they end with a plosive (like /sɛt/) or not. In all cases the final release, voiceless or voiced, after a plosive or after a fricative, has been counted as making part of the word. However, we have separated the results into two groups in order to distinguish the case where the vocalic

release does not appear (words 0, 1, 2 and 3) and where it may appear (words 4, 5, 6, 7, 8 and 9).

We ran variance analyses to explain the variation of the word duration. We added another factor to explain this duration variation: the type of the word. The variance analysis shows:

• significant effects of the noise (F Value = 167.91, p< 0.0001, around 14% on average of duration increase), the speakers (F Value = 81.98, p<0.0001 and the type of words (F Value = 15.51, p<0.0001).

• on the other hand, the type of noise and the presence or not of the auditory feedback have no significant effect on the increase of the word duration.

The results will be presented according the two types of word endings: vocalic and consonantic.

Words with vocalic endings

The words have on average, a duration increase of about 16% in the presence of noise. The type of the noise and the presence of feedback are not significant. The variation of the word duration is also function of the word type. This can be explained by the fact that they do not contain the same number of consonants — these ones do not show a tendency to lengthen under the influence of noise. The speaker factor has sometimes an effect on the variation.

Words with consonantic endings

The results are identical as for the vocalic endings:

• the noise has a significant influence (12,5% on average of increase for all noises merged),

• the type of noise (white noise or cocktail party noise) does not influence significantly the increase of the duration,

• the speaker is also a significant factor.

The analysis of the word duration confirms the study done on phonemes. It shows that speakers have, according to noise conditions, a variable behaviour (but significant, speaker has a typical behaviour).

CONSEQUENCE OF NOISE ON THE VOCALIC RELEASE AT THE END OF WORDS

Here, we will demonstrate if the conditions of speech production have an influence on the phonological strategy of speakers in noisy and/or feedback conditions. In a previous study [5] have observed:

• an increase of the number of omissions of consonants located at the end of the word, more particularly of phonemes /t/, /p/, and /f/

• and an insertion of phonemes in the shape of schwa appearing at the end of words in a ratio of two more times in noised speech than in reference speech.

In this study, we are interested in the observation of the schwa realizations and in its effects on the syllabic organization of the utterance. The vocalic releases at the word end, are characterized by voiced segments similar to weak central vowels. These segments appear in words like /katrə/ as well as in words like /sɛk/ or /sis/ ... without a final /ə/.

The results of the variance analysis

Two factors are added as explacative factors:

• the sex of the speaker

• and the phonological type of the word ending. We define three final types: T: for words ending with a plosive: /sɛ̃k/ /sɛt/, and /ɥit/,

S: for words ending with a fricative: /sis/ and /nœf/

tre: for words ending with the set of phonemes /trə/ in the word /katrə/.

The variance analysis shows the significant influence of the noise condition, the speaker, the speaker sex (216 vocalic releases for men and 133 for women), the phonological type of the words' ending and also the effect of the auditory feedback (131 vocalic release with feedback and 216 without feedback).

A high increase is to be noted at the end of noise. In the absence of auditory feedback, this percentage is multiplied by 4. The feedback reduces in a sensible way this influence. The type of noise does not seem to have a real effect.

The figure 2 shows the effect of noise conditions and of phonological type on the frequency of vocalic release realization. For the whole noise conditions, the phonological type of the ending has an important influence on the rise of the realization percentage: the increase of type is the double of the type S. The type 'tre' gives the most important

increase since it is almost four times more important than that of type S. The percentage of vocalic release of the words ending with a plosive is more important than those ending with a fricative.

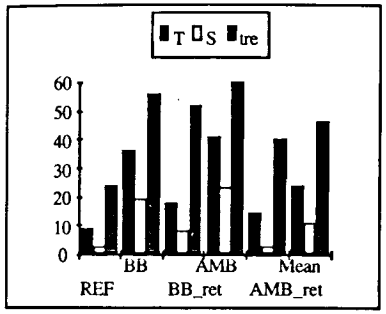


Figure 2 - Percentage of vocalic release.

The feedback has a constant effect on the rise of this percentage. The presence of vocalic release at the end of the word is significantly dependent of the speaker, of the noise conditions, of the auditory feedback, of the speaker sex and of the phonological type of the end of the word. On the other hand, the nature of noise (white noise or cocktail party noise) has no influence on this presence.

NOISE CONSEQUENCE ON FUNDAMENTAL FREQUENCY F_0

[5], [6], and [7] indicate an increase of the fundamental frequency (often correlated to an intensity increase) under the noise. The maximal value of the fundamental frequency (noted F_0) for each occurrence of the word was computed. The study of the F_0 variation to reference conditions shows:

• the significant influence of the speaker factor (F Value = 337.21, p< 0.0001), of the noise factor (F Value 588.58, p< 0.0001, on average 20,25% of fundamental increase for the noise merged), of the feedback (F Value 75.86, p <0.0001).

• the no effect of the type of the word.

Like observed by [6] our analyses of fundamental revealed no further significant differences between conditions (white noise and cocktail party noise) for isolated words. The type of noise is more significant without feeback than with auditory feeback. The variation is also dependent of the speaker.

CONCLUSION

To gain an understanding about the Lombard effect we analyzed the change of prosodic and phonologic parameters occuring in the five conditions of noise. The analyses made on French speech corpora,, to confirm previous studies on English and American, demonstrated the effect of the phonologic strategy of speakers due to the noisy/feedback conditions.

REFERENCES

[1] Lombard E. (1911), "Le signe de l'élévation de la voix", Ann. Maladiers Oreille, Larynx, Nez, Pharynx, 37 pp.101- 19.

[2] Lane H., Tranel, B. (1971), "The lombard Sign and the Role of Hearing in Speech" in Speech and Hearing, 14.

[3] Summers, W., Pisoni D.B., Bernacki R.H., Pedlow R.,Stokes M., (1988) "Effects of Noise on Speech Production : Acoustic and Perceptual Analyses", in J. Acoust. Soc.Am., 84(3) : pp. 917-28.

[4] Hassen J. and Bria O. (1990), "Lombard Effect Compensation for Robust Automatic Speech Recognition in Noise" in Proc. of ICSLP-90, pp. 1125 -8.

[5] Junqua J.C., Anglade Y. (1990), "Acoustic and Perceptual Studies of Lombard Speech: Application to Isolated-Word Automatic Speech Recognition", in Proc. of ICASSP-90, pp. 841-4.

[6] Howell P., Young K. and Sackin B. (1992), "Acoustical changes to speech in noisy and echoey environments", in Proc. of ESCA on Speech Processing in Adverse Conditions, Antibes, pp. 43-51.

[7] Junqua J.C. (1992), "The variability of Speech Produced in Noise", in Proc. of ESCA on Speech Processing in Adverse Conditions, Antibes, pp. 223-5.

[8] Furui S. (1992), "Toward Robust Speech Recognition Under Adverse Conditions" in Proc. of ESCA on Speech Processing in Adverse Conditions, Antibes, pp. 31-42.

[9] BD_BRUIT (1993), "Base de Données Parole, Locuteurs soumis au Bruit", GDR-PRC-CHM et ICP, Février.

[10] SAS/STT (1990), Users Guide, Volume 2 GLM-VARCOMP, version 6, SAS Istitute Inc.

SPEECH DEFICIENCES IN PRIMARY SCHOOLERS WITH DOWN SYNDROME

Iréne Johansson
Department of Education, Karlstad, Sweden

ABSTRACT

A comparative study of primary schoolers with Down syndrome in Sweden, Finland and New Zealand revealed syntactic skills more advanced than speech skills. The intelligibility of the childrens´ speech was very low. The speech difficulties are complex and characterized by variability, asymmetry and overshoot of movements. Even the task to imitate some simple non-verbal movements of the tongue and the lips were very hard for the children.

In 1993 teachers in Sweden, Finland and New Zealand were asked to answer some questions about their pupils with Down syndrome. To the opinions of their teachers about 5% af the primary schoolers with Down syndrome in Sweden, Finland and New Zealand used pre-symbolic communication. 17% of the children in Sweden and Finland but 45% of the children in New Zealand were considered to use one-word utterances. The vast majority of the Swedish and the Finnish children and 50% of the children in New Zealand were considered to express themselves in sentences. Of them only 36% of the Swedish children, 43% of the Finnish children and 21% of the children in New Zealand had started to use hierarchial sentences and morphology.

This is a different description of syntactic development in children with Down syndrome than earlier reports on the subject. E g Schlanger (1) and Schlanger & Gottsleben (2) wrote that many of the children with Down syndrome never acquired their mother-tongue and that the majority of them learned to talk in short one- or two-word sentences at best. Those who acquired spoken language had severe speech disorders. The articulation of vowels and consonants was described as indistinct, nasalized and the quality of the voices was described as aberrant.

The teachers in Sweden, Finland and New Zealand were asked to evaluate speech intelligibility of their pupils as well. To the opinions of their teachers the speech of the vast majority of primary schoolers with Down syndrome in the three countries is impossible or very hard to understand. Only 20% of the children in Sweden and New Zealand and 36% of the Finnish children were considered to articulate well enough for strangers to understand their speech. However, there were even some children who did not use speech at all - 8% of the Swedish children and 5 % of the children in Finland and New Zeeland. 3% of the Swedish non-speaking children "talked" in signed sentences.

The low level of intelligibility may partly be explained by language deficiencies - above all phonological disorders. However, there are neuromotor and neurosensory as well as cognitive deficiences which may have contributed to the prominent oral motor problems.

In another experiment 30 children with Down syndrome (8 years) and 30 children with normal development (8 year) in Sweden were asked to imitate some movements of the tongue and of the lips. The children were allowed to watch the model and to try several times.

In comparison to children with normal development at the same age, the children with Down syndrome did not succeed very well in imitating the movements of the tongue or the lips. It was especially hard for them to imitate movements they could not watch e g to follow the palate with the tongue-tip from the velar region to the teeth, repeated movements eg to move the tongue repeatedly from the left to the right corner of the mouth, and asymmetrical movements e g to raise only the left corner of the lips .

There are typical deficiencies in the speech production skills of the vast majority of a group of observed children with Down syndrome (N<100), producing the low degree of acceptability and intelligibility to their speech. There are eg a strong tendency to shorten the duration of the vowels and to approximate the vowel sounds to a centralized, semi-rounded vowel quality. Front vowels was retracted and back vowels was fronted , narrow vowels was opened and wide vowels was made more narrow . Spread vowels was made non-spread and rounded vowels were unrounded. [i] was as difficult to articulate as [y].

Also, the observed children had problems to find the adequate places of consonant articulation no matter what are the articulation manners. The children rarely used the dento-alveolar place of articulation. The speech of many of the children revealed a preponderance of velar, pharyngeal and glottal articulations.

Articulations at fronted places are interfered by the fact that the relation between the size of the mouth cavity and that of the tongue is different than normal (3). The oral cavity is defined by the reduced mandibular angle, the reduced caudal development of maxilla and a high, arched hard palate (4). The tongue is protruded and it fills the greater part of the front part of the oral cavity. Tongue protrusion in Down syndrome may be a symptom of an airway restriction by enlarged tonsils and/or adenoids.

The manner of articulation was often deviant as well. The stop production was characterized by a very prolonged and variable duration of the occlusion phase, the explosion phase was indistinct and sometimes not released and there was a lack of aspiration. The prolonged occlusion phases of the stops were followed by shortened vowels durations and the temporal structure of the syllable was detoriated. Sometimes the opposition between a stop and a nasal was cancelled due to a low velar resistance associated with velo-pharyngeal inadequacy. This is particularly true when the stop is prenasalized - a feature sometimes used in our observed children.

The children managed the production of a non-specific fricative sound very well. However, it was very hard for them to create adequate relationships between area-functions and attributes of the air-stream to produce different fricatives. Laterals sounds were often fricativized. A tremulant or retroflex sound was not observed at all.

The human tongue is an extremely flexible structure. The shape of the tongue (e.g. the degree of concaveness and convexness or the width of groove) and the position of the tongue (e.g. the degree of retraction, extension or retroflexion) is due to the tongue's internal composition, to the external muscles and to the support of the hard palate stabilizing the production of certain tongue shapes (5). In children with Down syndrome lingual diastatis is common (6) and there is a strong evidence for hypotonia of the tongue muscles in Down syndrome (7).

REFERENCES.

(1) Schlanger,B.B., 1957. Oral language classification on the training school residents. *The Training School Bulletin 53*, 243–247.

(2) Schlanger,B.B. & Gottsleben,R.H., 1967. Analysis of speech defects among institutionalized mentally retarded. *Journal of Speech and Hearing Disorders 22*, 98–103.

(3) Adrian,G.M., Harker,P. & Kemp,K.H., 1972. Tongue size in Down´s syndrome. *Journal of Mental Deficiency Research 16:3*, 160–166.

(4) Fischer–Brandies,H., 1988. Entwicklungsmerkmale des Schädels and der Kiefer bei Morbus Down unter Berücksichtigung der funktionellen Kieferorthopädischen Frühbehandling. *Habilitationsschriften der Zahn–,Mund–und Kieferheilkunde* Quintessenz Berlin; Verlags–HmbH.

(5) Stone, M., 1991. Toward a model of three–dimensional tongue movement. *Journal of Phonetics 19*, 309–320

(6)Limbrock, G.J., Fischer–Brandies,H. & Avalle,C., 1991. Castillo–Morales´ oro–facial therapy; treatment of 67 children with Down´s syndrome. *Developmental Medicine and Child Neurology 33, 296–303.*

(7) Yarom,R., Sherman,Y., Sagher,U., Peled,I.J., Wexler,M.R. & Gorodetsky,R., 1987. Elevated concentrations of elements and abnormalities of neuromuscular junctions in tongue muscles of Down´s syndrome. *Journal of Neurological Sciences 79*, 315–326

ARE PHONOLOGIC PROBLEMS PREDICTABLE FROM PRE-SPEECH VOCALISATION IN CHILDREN BORN WITH CLEFT PALATE?

B. Hutters
Department of General and Applied Linguistics, Copenhagen, Denmark
A. Bau, and K. Brøndsted
Copenhagen Institute for Speech and Hearing Disorders, Denmark

ABSTRACT

Studies on normal children and children at risk of developing language problems suggest that it is possible to predict phonologic problems during their pre-speech stage of development. Also with cleft children it has been assumed that such prediction is possible. The main purpose of the present study is to examine this assumption by comparing pre-speech vocalisation with early consonant patterns produced by a group of children born with cleft palate. A matched group of children born with normal velopharyngeal function was included for comparison. The results indicate that prediction is not possible for cleft palate children in the same way as for non-cleft children.

INTRODUCTION

Studies on normal children and children at risk of developing language problems suggest that quantity and diversity of vocalisations in the prelinguistic period are linked to subsequent speech and language development, and that prelinguistic measures seem to predict subsequent linguistic development including phonologic development [1].

Surgery should provide cleft palate children with a competent velopharyngeal mechanism. Nevertheless, velopharyngeal insufficiency may persist for a varying period of time resulting in speech errors such as nasality and nasal emission of air. However, these children are also at risk of developing phonologic disorders which may be phonetic or phonologic in nature. Errors, which initially occur as a consequence of structural and motor deviations may over time become incorporated into the child's developing phonologic system. The compensatory sounds developing by some cleft children may be considered a special case of this category of phonologic problems. Also, children born with cleft palate may be at risk of developing "true" phonologic disorders related to overall delays in their expressive language. As with other children at risk of developing language problems it has been assumed that also with cleft palate children phonologic problems may be predictable from their pre-speech vocalisation [2]. Therefore, the purpose of the present study is to compare vocalisation produced by one year old Danish children born with and without cleft palate with their consonant phoneme pattern at age three, in order to determine if phonologic problems in the early speech of three-year-olds can be predicted from their pre-speech along the same lines as for their non-cleft peers.

METHOD

Subjects

The subjects were 17 cleft children (12 of these had also cleft lip) and 17 non-cleft children matched for age and sex. The cleft children had surgery of the palate at 22 months of age. At the pre-speech recording the children were one year old (range: 0;11-1;01), and three years at the early speech recording (range: 2;11-3;03). Thus, the first recording took place about one year before surgery, the second one about one year after surgery. The children with cleft palate were recruited through the Cleft Palate Department in Copenhagen. None of the cleft children exhibited other congenital serious anomalies, neurologic impairment, serious sensorineural hearing impairment or intellectual deficits. Some of them had histories of middle ear problems, but had all received otologic mangement since birth. The non-cleft children were recruited through personal contacts and they had no reported history of speech, language or hearing problems, neurologic impairment or intellectual deficits.

Data collection

The one year old children were videotaped in their homes for approximately one hour. The toddler's spontaneous vocalisation was obtained during free play with his og her mother and in some cases also with the father. As to the second recording, the three-year-olds were videotaped at the Cleft Palate Department in Copenhagen. Each child was to name a non-standardized series of pictures (photographs) representing words which include 14 Danish word initial consonants. Each consonant was represented in two different words (photographs). In addition, one picture, which is part of a screening test for four-year-olds, served as basis for conversation with the speech and language pathologist in order to get information on consonants occurring in semi-spontaneous speech. From a standardized test it appears that children aged from three and a half to four years should manage about two thirds of the phonemes in question. Thus, a certain differentiation as to phonologic behaviour should be expected with the cleft as well as with the non-cleft children.

Data analysis

The International Phonetic Alphabet with extensions recommended for transcription of disordered speech was used for the analysis based on the videorecordings.

Concerning **pre-speech**, for each toddler 100 consecutive sequences were independently transcribed by two students in speech pathology under supervision by one of the authors.[1] Non speech-like sounds were excluded. This resulted in 1780 speech-like consonants produced by the 17 cleft toddlers and 1711 speech like consonants produced by the non-cleft toddlers. The transcription of pre-speech was later controlled for certain aspects by the three authors of the present paper. For each toddler in both groups, frequency of occurrence of various characteristics of his or her consonant inventory was analyzed. In the present context we focus on the number of different consonantal sounds (DIF) produced by the toddlers, which reflects diversity of vocalisation, and on the number of sequences produced in 30 minutes (SEQ), which reflects quantity of vocalisation. The sequences were perceptually separated by means of caesura, a phenomenon known from music.

The consonants in **early speech** were independently transcribed by the three authors, who are all familiar with cleft palate speech. When disagreement occurred between the transcribers, the sequence was replayed in an attempt to reach consensus between at least two of the transcribers. With few items no consensus could be reached, and in these cases the sequence in question was omitted from the data set. Also, in case of doubtful lexical meaning the sequence was excluded. Thus, the material for each child informs as to how a given phoneme may be realized, but not as to frequency of occurence of a given realization. The number of identifiable phonemes (ID) was analyzed for each child. It should be noticed that whether a phoneme is considered identifiable or not is based on our impressionistic judgement.

A Mann-Whitney U-test was used to determine level of significance (one-tailed) concerning differences between the groups.

Assuming that degree of **correlation** reflects degree of predictability, Spearman's rank correlation (one-tailed, corrected for ties) was used to determine the degree of correlation between ID and various pre-speech characteristics, in the present context with focus on SEQ and DIF.

Finally, two of the authors made an unformal **qualitative description** based on the video recordings of the one-year old toddlers with focus on social interaction and communication. This description will only be mentioned in the discussion. It should be added, that these two authors had no knowledge of the subsequent speech development, as they were unable to identify the toddlers.

RESULTS

The present paper focuses on the relationship between pre-speech and early speech. Therefore, no general presentation of the childrens' pre-speech vocalisation and early speech will be presented except that in pre-speech, glottal stops and [h] occured more or less with all cleft and non-cleft children, and that significantly more glottal stops were found with the cleft group.

Based on other studies, we considered SEQ and DIF in **pre-speech vocalisation** potential predictors of phonological problems [1]. We find that SEQ is slightly lower with the cleft group, but the difference is not significant ($p > .05$). Contrarily, DIF is on the average considerably lower with the cleft toddlers than with their non-

cleft peers. Median and range are 6 (1-9) and 12 (5-16) for the cleft and non-cleft group, respectively, and the difference is significant ($p<.001$).

As to **early speech** ID was used as a measure reflecting the positive aspect of the phonologic capability of the three-year-olds, assuming a negative relationship between number of identifiable consonants and degree of phonologic problemes. Median and range are 9 (3-13) and 14 (9-14) for the cleft and non-cleft group, respectively, and the difference is significant ($p<.001$). It appears, that none of the cleft children managed all the consonanats, whereas about half of the non-cleft children did.

As to **correlation** between the early speech variable ID and the two pre-speech variables SEQ and DIF, the results are as follows:

Number of sequences(SEQ):

non-cleft group:	rho= 0.55 $p<.025$
cleft group:	rho=-0.01 $p>.10$

Number of different consonants(ID):

non-cleft group:	rho= 0.60 $p<.01$
cleft group:	rho=-0.23 $p>.10$

It appears that only with the non-cleft group we find the expected correlations. However, with the cleft group, a third pre-speech variable correlates negatively with ID:

Frequency of non-initial [h] (% of total number of (a) all consonants and (b) all [h]'s):

cleft group:	rho=-0.45 $p<.05$(a)
	rho=-0.60 $p<.01$(b)
non-cleft group:	rho= 0.28 $p>.10$(a)
	rho= 0.14 $p>.10$(b)

It should be added that with [h] no significant difference was found between the groups, irrespective of position.

Thus, in spite of the fact that none of the significant correlations are very high, it appears that phonologic problems in early speech produced by cleft children do not seem to be reflected in the same pre-speech variables as with the non-cleft children.

DISCUSSION

As to differences between the groups, we did not find a significantly smaller SEQ-score with the cleft group than with their non-cleft peers, which disagrees with our assumption on pre-speech quantity. However, if quantity of pre-speech vocalisation tends to reduce with development of speech, this may explain the non-significant SEQ-difference between the two groups provided that the cleft group are delayed compared to the non-cleft group. This is supported by the qualitative description of the toddlers, from which it appears that all the non-clefts are more or less at transition to early speech as their vocalisation begins to take on the characteristics of speech as a tool for communication, whereas none of the cleft toddlers show this tendency.

Concerning correlation it can be stated that the results with the non-cleft group support the assumption based on other studies, namely that quantity (SEQ) and diversity (DIF) in pre-speech vocalisation seem to be linked to subsequent speech and language development. However, this does not seem to be the case with the cleft group. As mentioned above children born with cleft palate are at risk of developing "true" phonologic disorders as well as various phonologic disorders which are phonetic in nature. Thus, it seems likely that only with cleft children who develop "true" phonologic disorders, significant correlations between the same variables as found with the non-cleft children should be expected. Unfortunately, our material is not suitable to go into that question.

In the present material we found one correlation which only occurs with the cleft group, namely 'frequency of non-initial [h]': the higher frequency of non-initial [h], the lower ID-score, especially with non-initial [h] compared to the total number of [h]. If this turns out to be true, non-initial [h] should be a valid predictor of phonologic disorders. This may seem plausible, as sequence initial [h] and glottal stops may be considered phonation onset rather than speech-like sounds, and therefore common in the non-cleft group as well. And as [h] can be considered a non-active sound from an articulatory point of view, a high frequency of non-initial [h] might reflect level of speech development. The question is if it predicts specific kinds of problems, as for instance development of speech dominated by [h] and glottal stops. In the present material the three highest non-initial [h]-scores are found with one toddler whose early speech is dominated by glottal stops, and two with early speech dominated by [h]. But, two other children whose early speech is dominated by glottal stops and [h], respectively, are at the lower end of the range of non-initial [h] scores as toddlers. So, even though no conclusion can be drawn from these few data, it seems worth while to look closer at this point.

One important question is whether compensatory use of glottal stops are predictable from pre-speech vocalisation. From the present material including two cleft children with glottal compensatory articulation, it appears that quantity of glottal stops in the pre-speech vocalisation of one year old cleft toddlers do not reflect development of phonologic use of glottal stops. In a previous study [3] we hypothesize that "teaching-like mothers" are potential candidates for reinforcing speech with compensatory articulation as one of several factors involved in the development of glottal compensatory articulation. Interestingly, it appears from the qualitative description of the pre-speech recordings that with these two children, the social interaction on the part of the mother and sister, respectively, is clearly teaching-like as to communication behaviour. This has not been observed in the other recordings. This observation support our hypothesis, although it still has to be proved.

The present study illustrates the fact that in logopedic research group analyses may be less suitable due to inhomogeneity within the groups. Thus, before conclusion we shortly present two cases from each of the groups, which illustrate that with cleft as well as with non-cleft children research on predictability based on single cases may not be a simple task either. One cleft and one non-cleft toddler whose pre-speech vocalisation apart from glottal stops and [h] is characterized by non-velar consonants really develop fronting in their early speech. However, two other children from each of the groups whose pre-speech is charcterized by velar consonants likewise show subsequent fronting in their early speech. Finally, it should be mentioned that in the vocalisation of one of the cleft children who develops 'pure' h-speech, few - but non-initial - [h] occur in his vocalisation, while nasals and nasalized sonorants are the dominating consonantal sounds.

CONCLUSION

To conclude, the present study indicates that cleft children with non-active articulatory pre-speech behaviour as reflected in the frequency of occurrence of non-initial [h] may be at risk of developing less identifiable consonants in their early speech. Further, it may be that mothers who are teaching-like in their communication with their cleft toddler are potential candidates for reinforcing compensatory articulation with their child. If these two factors turn out to be essential to the question of phonologic development in cleft palate speech, the clinical consequence could be that more active training with a non-active toddler as a prophylactic intervention may result in development of the active, but very undesirable, glottal compensatory articulation.

The answer to the question if phonologic problemes are predictable from pre-speech vocalisation in children born with cleft palate is a "maybe" judged from the present study, and a considerable amount of complicated research has to be performed before a more final answer can be given.

REFERENCES

[1] Stoel-Gammon, C. (1992) Prelinguistic Vocal Development. Measurement and Predictions. In *Phonological Development. Models, Research, Implications* (C.A. Ferguson, L. Menn, C. Stoel-Gammon, eds.), York Press, Timorium, Maryland.

[2] Russel, J. & Grunwell, P. (1993), 1"Speech Development in Children with Cleft Lip and Palate." In *Analysing Cleft Palate Speech* (P. Grunwell, ed.), 19-47. Whurr Publishers, London.

[3] Hutters, B. & Brøndsted, K. (1993), Preference between Compensatory Articulation and Nasal Emission of Air in Cleft Palate Speech - with Special Reference to the Reinforcement Theory. *Scandinavian Journal of Logopedics and Phoniatrics* vol.18, pp. 153-158.

NOTE

1. The pre-speech data originate from an A.M. thesis by Anja Bau and Ulla Lahti.

ARTICULATORY/ACOUSTIC RELATIONSHIPS IN LATERALISED PRODUCTIONS OF SIBILANT FRICATIVES

*H. Dent, **F. Gibbon, **W. Hardcastle and ***M. Wakumoto
*National Hospital's College of Speech Sciences, London.
**Queen Margaret College, Edinburgh.
***First Department of Oral & Maxillofacial Surgery, Showa University, Japan.

INTRODUCTION

Clients presenting with abnormal productions of the lingual fricatives /s/ and /ʃ/ (the "sibilants") are common in the speech and language disordered population. One form of sibilant misarticulation is the "lateral lisp", (Dagenais, Critz-Crosby and Adams, 1994), which may be encountered not only in individuals who have structural abnormalities such as malocclusion and cleft palate, and those who have a history of delayed or disordered acquisition of speech and language, but also in children with otherwise normal articulatory skill development. It has been suggested that in such cases the speech production difficulty may be due to oral sensory feedback or neuromuscular deficits, or speech discrimination problems, (Wilcox, Daniloff and Ali, 1984). Regardless of aetiological factors, it is generally acknowledged clinically that lateral misarticulations are notoriously resistant to therapeutic intervention employing conventional techniques.

Neither the frequent occurrence, nor the resistance to treatment, of the lateral lisp is particularly controversial given the precise lingual neuromuscular control and the fine co-ordination of the lingual and respiratory systems which are necessary for accurate sibilant production. Furthermore, the visual inaccessibility of sibilants with placement posterior to the alveolar ridge not only means that description of the precise articulatory characteristics of abnormal productions has proved difficult, but also reduces the amount of information available to the clinician and client attempting to modify lingual configuration.

Electropalatography (EPG) is a computer-based technique which records, stores and displays information concerning the timing and location of lingual contact with the hard palate. The sibilants /s/ and /ʃ/ involve a relatively narrow oral constriction, resulting in an identifiable area of lingual-palatal contact, and EPG can therefore be employed to derive objective details regarding articulatory placement during their production. This information can be used in both the assessment and the remediation of atypical speech patterns and the application of EPG to these aspects of intervention has been widely reported, (see Nicolaidis, Hardcastle and Gibbon, 1994).

This paper will illustrate the articulatory patterns seen in lateral lisps produced by five English speaking children and adolescents. These will be related to acoustic characteristics of the noise spectra in an attempt to identify the salient articulatory and acoustic features of this type of distortion. Observations before and subsequent to EPG-based intervention will be compared and discussed with a view to providing some insight into why this type of disorder commonly arises and why it is so resistant to therapeutic intervention.

Table 1. Subject Details

	AGE	SEX	LATERALISED SOUNDS	ASSOCIATED FACTORS
D1	16;03	F	/ʃ/	family history of speech problems; orthodontic abnormalities
D2	11;07	F	/ʃ/	open-bite
D3	12;03	F	/s/ and /ʃ/	early history of conductive hearing loss; orthodontic abnormalities
D4	10;05	M	/s/ and /ʃ/	delayed speech and language development; attended primary language unit
D5	18;06	M	/s/	delayed speech and language development; ongoing disfluency

METHOD

Subjects

The five subjects formed part of a larger group which participated in a research project, described in Dent, Gibbon and Hardcastle (in press). Each had failed for some time to respond to conventional therapy. Auditory judgements of each of the subjects' productions of the sibilants were made (by the authors) at the time of their referral to the project. For each of the five subjects, the lateral lisp was the only presenting speech problem. All had histories of possible associated factors although only for D4 and D5 were these considered to be potentially significant. Table 1 summarises the details of the five subjects.

Instrumentation

The Reading EPG system is based on an artificial palate which contains 62 electrodes arranged in eight horizontal rows and exposed to the lingual surface. When the tongue touches the electrodes, a signal is conducted to an external processing unit and the pattern of lingual-palatal contact displayed on a VDU. Patterns can be stored on computer or printed out for subsequent analysis. (For more details see Hardcastle, Gibbon and Jones, 1991.) A multichannel system for simultaneous acquisition and processing of EPG and acoustic data was used for assessment of the subjects' speech, with sampling rates of 200 Hz for the EPG and 20,000 Hz for the acoustic signal. The system has been described elsewhere, (Hardcastle, Jones, Knight, Trudgeon and Calder, 1989). In addition, DAT recordings were made of each assessment session, using a Sony DAT (DTC-100ES).

Test material

The recordings made were of a word list containing all the consonants of English in single words and short sentences. Also recorded in each case was an additional list which contained further examples of sibilant targets, in items which were potential homonyms in the speech of the subject concerned.

Data analysis

In order to extract the salient articulatory features of the subjects' lateral lisp productions, specific EPG frames were selected at a number of points during each abnormal fricative segment. These points were identified according to criteria relating, for example, to the initiation and cessation of friction, and the frame of maximum

lingual-palatal contact. (The criteria are defined in Dent, Gibbon and Hardcastle, in press.) EPG frames from the annotation points for each production could then be presented in a single sequence. (See Figure 1.)

Acoustic analysis was undertaken using the CSL-4300 system (Kay Elemetrics, USA). A 50 millisecond Hanning window was placed at a point during the period of frication which corresponded to the frame of maximum contact in the EPG data. A spectral envelope was then extracted from the FFT and used to obtain a measure of Consonant Peak Energy Frequency (CPF, Wakumoto 1989).

Therapeutic procedure

The therapy mode of the EPG system was used to provide the subjects with visual feedback of target lingual configurations for the sibilants they were misarticulating, and feedback of their own attempts to approach these. General principles and stages of treatment using EPG are outlined in detail in Hardcastle et al (1991).

RESULTS

Sequences of EPG frames from the annotation points of selected items are presented in Figure 1 to illustrate each of the five subjects' production of target sibilants prior and subsequent to EPG intervention.

Figure 2 shows the change seen in CPF prior and subsequent to EPG intervention, for target /ʃ/ in the case of D2 and for target /s/ in the case of D5.

DISCUSSION

The EPG data reported here reveal a tendency for lateral lisps to involve complete closure between the tongue and the hard palate in the anterior region. The exception to this, D1, has increased contact approaching complete closure in the palatal region. The presence of such (near) closure, together with the lack of bilateral posterior contact seen in each of the patterns here, suggests that the friction for these sounds is produced by directing the airstream laterally behind the dental arch into the buccal cavity. Previous EPG studies have also reported increased tongue-palate contact in lateral

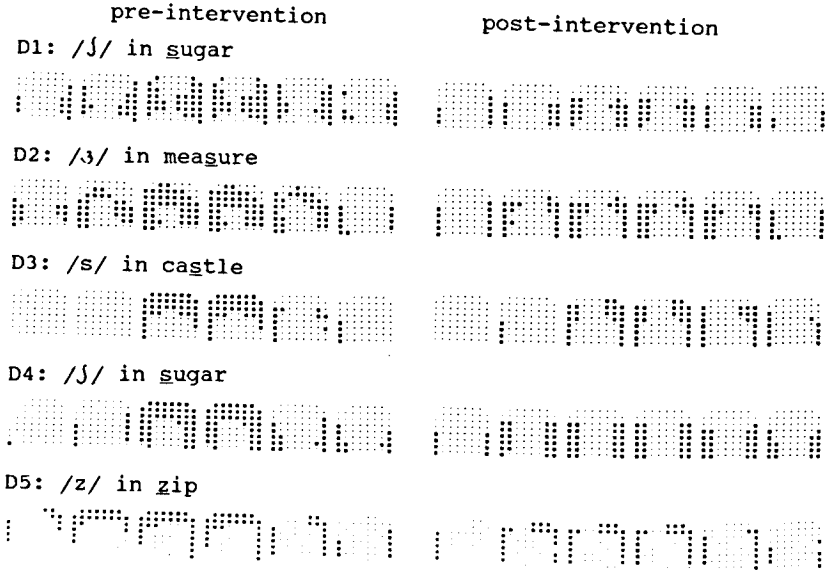


Figure 1. Series of EPG frames representing each subject's production pre- and post-intervention

		castle	saw	search	sue	sip	said	circus1	circus2	average
D5	Session1	4521	3869	3953	4470	4297	4403	3991	4299	4228
	Session2	3804	4810	8943	4503	6380	6362	9453	7543	6475

		wash	fishing	shore	shed	ship	shop	shoe	average
D2	Pretherapy	5183	5553	5237	5476	5357	5171	5271	5321
	Posttherapy	1601	2081	1793	3534	2099	1742	1777	2089

Figure 2. CPF change pre- vs post-intervention

lisp productions, (eg: Dagenais et al, 1994). This may reflect a lack of the fine neuromuscular control necessary to produce accurate sibilants: in the absence of this control, these subjects have adopted the broader, less precise tongue-palate contact typical of plosive articulation and allowed air to escape laterally in order to produce friction.

Following EPG therapy, all the subjects were able to produce sibilants which were perceptually more normal, with minimal lateralisation. The articulatory changes which they made, (see Figure 1), were reflected to some extent in the acoustic results. Figure 2 gives two examples: D5 has, for /s/, CPFs of approximately 4000Hz prior to therapy, and clearly higher peaks, (approximating the more normal 7000Hz, Suzuki et al (in press)), for over half of the items after therapy. The trend for D2's /ʃ/ is clearly towards a lower CPF following intervention, perhaps reflecting the more posterior (and therefore more normal) tongue placement. The results for the other subjects reported here showed similar patterns of CPF values for /s/ and /ʃ/.

ACKNOWLEDGEMENTS

The work described in this paper was supported by the British Medical Research Council (Project no. G8912970N). Thanks are due to Wilf Jones who designed the Reading EPG system and provided technical assistance.

REFERENCES

Dagenais, P., Critz-Crosby, P. and Adams, J. (1994) Defining and remediation persistent lateral lisps in children using electropalatography: preliminary findings. American Journal of Speech-Language Pathology, 3, 67-76.

Dent, H., Gibbon, F. and Hardcastle, W. (in press) The application of electropalatography to the remediation of speech disorders in school-aged children and young adults. European Journal of Disorders of Communication, 30.

Hardcastle, W., Gibbon, F. and Jones, W. (1991) Visual display of tongue-palate contact: electropalatography in the assessment and remediation of speech disorders. British Journal of Disorders of Communication, 26, 41-74.

Hardcastle, W., Jones, W., Knight, C., Trudgeon, A. and Calder, G. (1989) New developments in electropalatography: a state-of-the-art report. Clinical Linguistics and Phonetics, 3, 1-38.

Nicolaidis, K., Hardcastle, W. and Gibbon, F. (1993) Bibliography of electropalatographic studies in English (1957-1992) Parts I-III. Speech Research Laboratory, University of Reading, Work in Progress, 7, 26-106.

Suzuki, N., Dent, H., Wakumoto, M., Gibbon, F., Michi, K. and Hardcastle, W. (in press) A cross-linguistic study of lateral misarticulation using electropalatography. European Journal of Disorders of Communication, 30.

Wakumoto, M. (1989) Quantitative technique for evaluation of Japanese palatalised articulation. Journal of the Japanese Cleft Palate Association (Japanese), 14, 21-43.

Wilcox, K., Daniloff, R. and Ali, R. (1984) Speech sound discrimination in /s/ misarticulating children. In R. Daniloff (Ed.) Articulation, Assessment and Treatment Issues. San Diego: College-Hill Press.

TONGUE MOVEMENTS IN MALOCCLUDED SUBJECTS

A. Giannini, M. Pettorino*, G. Savastano**, C. Cimmino**, G. Della Pietra***
** Istituto Universitario Orientale, Fonetica Sperimentale, Napoli, Italia*
*** Università "Federico II", Chirurgia Maxillo-Facciale, Napoli, Italia*

ABSTRACT

The presence of "egressive clicks" due to the mandibular protrusion seems to be a peculiarity of III class maloccluded subjects. In this experimental research two methodologies are used to assess the relationship between tongue movements and skeletal discrepancy.

PROCEDURE

The aim of this experimental research is to establish a relationship between abnormal tongue movements and skeletal discrepancy in III class maloccluded subjects. In the present work two methods of investigation have been employed: a cephalometric analysis through telecranium in L/L projection and a spectrographic analysis through DSP Sonagraph 5500 KAY.

In the cephalometric analysis not only parameters relative to vertical and sagittal discrepancy were evaluated, but also some parameters indicative of tongue posture. In particular the following parameters were measured: i. the distances of the tongue dorsum from the palatine arch and the vertebral column; ii. the distance between the tip of the tongue and upper and lower incisors; iii. the distances of the hyoid bone from the vertebral column and the mandible.

For spectrographic analysis the Pitch Display program, which monitors both the variations of fundamental frequency by means of a 18.75 Hz bandwidth filter and the formant frequencies of the supralaryngeal cavities by means of a 300 Hz bandwidth filter, was used. The frequency scale has been settled either at 8 KHz or 16 KHz according to the situation.

A corpus of 100 meaningful Italian words and one minute of spontaneous speech have been uttered in a silent room by 12 maloccluded subjects as well as by a control group of 8 normoccluded subjects. All speakers were Italian, males and females, aged from 17 to 22 years. The list of words included all Italian phones that, when inserted in the right context, appeared to be relevant to the definition of the relationship between articulatory movements and acoustic signal.

DISCUSSION

In figure 1 all speech anomalies found throughout the experiment are reported.

As far as the voiceless alveolar fricative is concerned, several speech defects can be detected. These can be linked to the following different types of faulty s-sounds: dental, alveolo-coronal and whistled. These three types are compared in figure 2 with the alveolar fricative uttered by a normoccluded subject. For space saving reasons only the important segment is reported. As it is observed the dental fricative exhibits an intense signal at high frequencies around 10-12 KHz (fig. 2, *b*). Therefore the signal portion below 7 KHz, that is perceptively considered more relevant, has been found to be much less intense. This frequency distribution reflects a punctual articulatory constriction between the edges of the upper incisors and the tip of the tongue.

The production of an alveolo-coronal fricative represents one of the peculiarities of III class maloccluded subjects. As a matter of fact, in normoccluded subjects the blade of the tongue opposes the postalveolar area so that the formed constriction produces a ʃ-sound. On the

FAULTY SPEECH SOUNDS		S 1	2	3	P 4	E 5	A 6	K 7	E 8	R 9	S 10	11	12
[s]	dental	●									●		
	alv.coron.	●		●		●				●		●	
	whistled			●					●				
[k][g]	fricative	●		●						●			●
[r]	fricative	●					●	●		●		●	
	approximant			●							●		
[f]	whistled				●								●
clicks	bilabial			●	●					●			
	alveolar	●	●	●		●		●		●		●	
	palatal	●	●	●		●		●			●		●
	velar	●	●	●			●	●			●		●

Figure 1. Different speech anomalies of the maloccluded sujects.

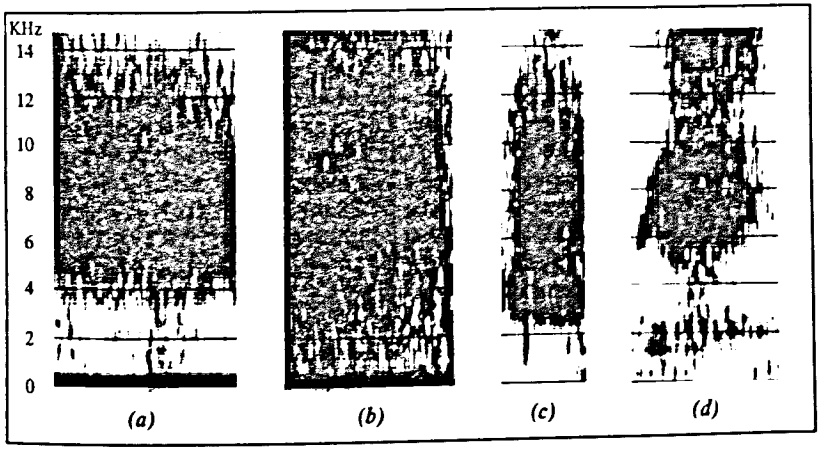


Figure 2. Broad band spectrograms at 16 KHz of s-sounds:
(a) alveolar; (b) dental; (c) alveolo-coronal; (d) whistled.

contrary in III class maloccluded subjects the mandibular protrusion is such that the tongue is in a forward position with respect to the hard palate. As a consequence the blade of the tongue opposes the alveolar area. The acoustic effect of this anomalous contrapposition results in a signal frequency lowering due to a longer articulatory constriction (fig. 2, *c*).

As far as the whistled fricative is concerned the tongue, instead of taking on a convex conformation, flattens down in the middle and arches up on the edges in such a way that it makes a middle channel

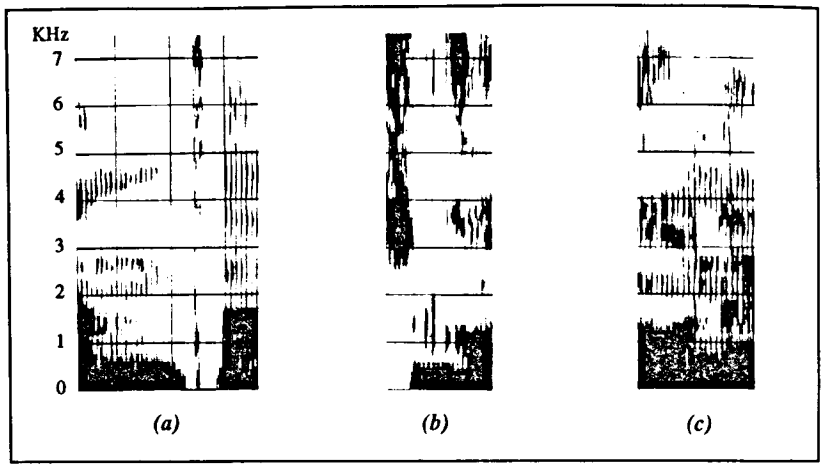


Figure 3. Broad band spectrograms at 8 KHz of different clicks: (a) voiceless; (b) voiced; (c) nasal.

which ends in a constriction between the tip of the tongue and the upper incisors. The resulting acoustic signal is characterized by very high harmonic frequencies (fig. 2, *d*). The same acoustic features are in some cases found also in the labiodental fricatives.

The anomalies found out in the alveolar fricative production were not noticed in the other consonants produced in the same alveolar and postalveolar place of articulation, that is in the alveolar stops and affricates as well as postalveolar affricates. The acoustic signal does not exhibit any anomalous posture as it is to be observed from the behaviour of the F2 transitions of the adjacent vowels. As the silence portion is of capital importance in distinguishing stops and affricates from other consonants, the subject is forced to carry out such closure at alveolar and postalveolar level thus compensating the anomalous tongue forward position with a backward movement of the tip of the tongue. In order to produce such backward movement the subject, in theory, could follow three different articulatory mechanisms: i. draw backward the whole tongue mass toward the pharynx; ii. lift the tongue postdorsum towards the soft palate; iii. flatten and enlarge laterally the tongue dorsum. The first two mechanisms can be excluded since from an acoustic point of view the former would cause an F1 increment and the latter an F2 lowering. Since in none of the cases such behaviour has been found, it is to be concluded that the third mechanism is the likely one.

A confirmatory evidence of such hypothesis comes from the "egressive click" examination (figure 1). With the term "click" we refer to a momentaneous acoustic signal lasting about 10 ms, of strong intensity and variable in frequency depending on the place of articulation where it is generated. This kind of click, from an articulatory point of view, is produced by the formation and rapid release of an occlusion with subsequent air outflow. The release burst is particularly violent and sharp because of the pressure build up behind the occlusion. In fact, most clicks found in the examined subjects, appear in those phones requiring a closure in the oral

cavity. In this case, too, in order to ascertain the site of activation of such mechanism, it is possible to go on by exclusion. A first hypothesis that can be excluded on an experimental basis is that clicks could be produced at glottal level through anomalous openings of the vocal folds. The spectrograms exhibit, in fact, clicks forming both during voiceless phones peculiar of a wide open glottis and during voiced phones (fig. 3 *a*, *b*). In the latter ones one can easily identify clicks that are independent of periodic openings and closings of the vocal folds. A second hypothesis to discard is that correlating the click to a faulty contact between the velum and the naso-pharynx with subsequent air inflow through the nasal cavities. Spectrograms show, in fact, that clicks appear, but with variable intensity, both in nasal phones where the velum is lowered and in the oral ones where no sign of nasalization in the acoustic signal is revealed (fig. 3 *a*, *c*). In three of the examined subjects the presence of clicks during the production of bilabial phones has been detected. In this case the click is due to an anomalous backward movement of the lower lip. This movement causes a defective lateral contact quite similar to that occurring when the articulator is the tongue.

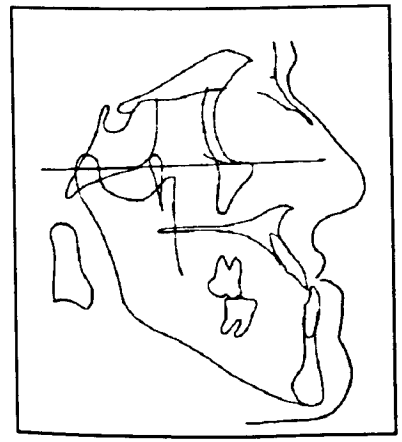


Figure 4. Telecranium of a maloccluded subject.

The data obtained from cephalometric analysis, comparing I class maloccluded subjects with those belonging to III class, show that, in the latter class, there is a tongue backward posture with interposition in lateral sectors. In figure 4 the telecranium of one of maloccluded subjects is reported.

The results are in full accord with the phonetical conclusions.

CONCLUSIONS

All arguments thus discussed allow the following conclusions as far as the movements of the articulators in III class maloccluded subjects are concerned. When a subject has to realize a closure into the oral cavity he has to withdraw his tongue and this is accomplished through a concerted flattening and a lateral enlargement. Thus the lateral contacts do not occur between the tongue edges and the alveoli but between the tongue edges and the molar teeth. This results in faulty movements that cause trills originating sometimes an almost periodical sequence of two or more clicks.

In addition it must be noticed that in the case of the velar stops, case of particular stress for the speaker, quite often the occlusion is replaced by a constriction. Even the alveolar trill undergoes a change in the manner of articulation since it is sometimes accomplished as a fricative and other times as an approximant.

As for the read and spontaneous speeches it is worth emphasizing that the spontaneous speech is marked by a larger number of defects than the read one which is characterized by compensatory strategies.

The spectrographic data agree quite well with the cephalometric data indicative of a lingual backward posture with interposition in the lateral sectors.

FOREIGN ACCENT SYNDROME: AN ITALIAN CASE STUDY.

*G. Denes, **J. Trumper, **M. Maddalon, **L. Romito
Univ. of *Padua and **Calabria

0. Studies of patients affected by Foreign Accent Syndrome are few and not well documented, often anecdotal. Many of the possible cases have been eliminated, since their clinical deficit is explicable in other terms [for overview cf.3: for first systematic studies see 4]. Here we report an Italian case meriting in-depth study, where the patient developed a supposedly strong "English" accent following head trauma: her performance has been studied acoustically against a normal Italian control case and a native English speaker.

1.*Clinical case study.* MCF, a right-handed woman, Ph.D. in geology, at 28 years of age had a climbing accident with severe head trauma. A CT scan showed presence of a left-frontal contusion. Three months later she was assessed neuropsychologically : initially completely aphonic, then dysphonic/ dysprosodic, with a tendency to use ingressive airflow in speech, she had a supposedly foreign accent labelled "English" by most. She was neither aphasic nor apraxic and her general neurological exam showed minimal right hemiparesis. Because of the persistent "foreign" accent she was re-assessed a year later. An MR was performed showing bilateral hypodensity of the frontal white matter.

2. 0 *General linguistic correlates.* The following parameters were used: A. prosody (concept / realization of FO declination); B. vowel quality / length; C. consonant length (Italian fatto vs. fato, palla vs. pala); D. consonant with vowel / sonant coarticulation; E. abnormal consonant / sonant realizations (not length or VOT); F. reductions of syllable structure; G. vowel bleaching vs. non-bleaching; H. VOT.

2.1 *Evaluation methods.* The patient's speech production was first analyzed (a. reading tests: Little Red Riding Hood; b. spontaneous résumé of the same; c. natural conversation; d.word lists) and then compared with her female cousin's identical production (only a + b), lastly with similar tests done with a native female speaker of English with no knowledge of Italian (trained to reproduce the same text : only level a).Our report deals only with the level a tests. Vowel parameters, except for F_0, were measured with a Kay 5500 and ONO SOKKI 930, F_0 with appropriate software (Signalyze by E.Keller). Statistics were elaborated with conventional software (Phonetics Lab. Univ. Calabria).

3.0 *Analysis.* First hypothesis advanced [5] was that FAS mainly involved significant F_0 alterations and NOT other parameters mentioned (B-G). Later research pinpointed the equally meaningful role played by segmental phenomena vs. prosody or combinations of both [10;4;3]. All the phenomena evinced [5] are typical of SOME natural language unlike those in aphasia and other pathologies, the probable reason why medical researchers have labelled it with its current name (FAS). We have sofar investigated all parameters given in the a test: VOT alteration (G), bleaching (F) and some aspects of (D,E) will not be reported on here. Our results thus concern: A, B, C, D (only stop- sonant sequences [pr, br, tr, dr, gr]) , E (only stop lenition and strengthening, aspiration and glottalization of stops / continuants, continuant affrication), F (vowel / syllable reduction). Though many phenomena present in the patient's speech occur in SOME natural languages, their sum does not characterize any SINGLE one. Some are conflicting (stop lenition vs. strengthening). Sometimes in the same utterance we find phenomena typical of both stress-timed and syllable-timed languages (most Italian varieties are syllable-timed).

3.1 *Results.* Prosodic analysis used 4 complex sentences: "Vado a trovare la mia nonna e le porto una focaccia e un vasetto di burro, preparati per lei dalla mamma"/ "Il lupo non ci mise molto ad arrivare alla casa della nonna"/ "Cappuccetto rosso tirò il chiavistello e la porta si aprì"/ "Il lupo tirò il chiavistello e la porta si aprì", comparing F_0 contours for the patient, her cousin and the English speaker. First impressions do not favour an absence of intonation contour hypothesis in the patient's case. A declination line is observable in the last 3 cases and a clear fall- rise pattern strategy corresponds to major syntactic breaks (patient + cousin). In the first the patient has a series of F_0 peaks which do not tie in with any syntactic foci. After a false start (F_0 = 262 Hz) a new peak is begun and followed by 4 peaks higher or as high as the initial one [Pi= 280 Hz trovarE, 275 Hz portO, 290 Hz focacciA, 280 Hz burrO, 278 Hz leI). All fall on unstressed vowels, contrary to expectations and are examples of overshooting , as in sentence 2 Pi = 320 Hz lupO, sentence 4 Pi = 280 Hz. lupO. Sentence 1 is anomalous resetting due to sentence length / complexity, while 3 sentences out of 4 show obvious overshooting: four cases are noted in sentence 1 (cf. figs.1, 2 in that order: sentence 1 for patient + cousin). Since the literature claims overshooting to be due to imperfect synchronizing of laryngeal muscular tension and subglottal pressure, there are obvious neurophysiological problems to be looked into. Furthermore, in the patient's case there is no biunique correspondence between amplitude peaks and stressed vowels, as would, instead,be expected in Italian, where stress correlates not only with segment length but also with intensity and pitch [2; for different combinations in different regional varieties cf. 7]. The cousin presents increased intensity and amplitude corresponding to vowels that carry dynamic sentence stress.

The patient and her cousin differ little in the length of vowels with dynamic stress: CLOSED SYLLABLE Patient $\bar{x}$ = 86 ms (N=77, σ=17), Cousin $\bar{x}$ = 88 ms (N=49, σ=13); OPEN SYLLABLE Patient $\bar{x}$ = 132 ms (N=56, σ =27), Cousin $\bar{x}$ = 144 ms (N=34, σ=22). Using a two-tailed t- test we find length differences significant at the 98% level in open but not in closed syllables. The patient, nonetheless, observes the general trend, her behaviour matches that of numerous normal speakers tested for Italian syllable- timing. Vowel realizations, however, are meaningfully different. The cousin's 7 vowel areas have been plotted in terms of F1- F2 (areas correspond better to Tuscan use of acoustic space than that of Veneto and Campania speakers, cf. relevant areas in 9). The patient's values superimposed (see fig.3 with cousin's vowel areas and patient's values superimposed) show serious overlaps between /i/ and /e/, /ɛ/ and /a/, /o/ and /ɔ/. Dispersion in terms of σ values is extremely high: 'maximal perceptual contrast' and 'maximal dispersion in vowel space' are here meaningless concepts. High-Low or Back-Front dimensions are not exploited for maximal contrast: we even have minor overlaps between /i/ and /u/, major ones for /ɛ/ and /a/. F1-F2 target space is insufficiently mastered; perhaps there is a neurological problem to be investigated. The problem seems both phonological (programming vowel contrasts) and phonetic (controlling articulatory movements). Though consonant length is phonological, most Northerners variably shorten [social determinants] . Large Veneto populations show shortening: Males $\bar{x}$ = 46.07% (pop.N=93,σ=15), Females $\bar{x}$ = 35.3% (pop.N=27, σ=14); compare with Patient's 21.43% (non-sonorants 24 /112), Cousin's 31.75% (id. 40 /126). We have calculated shortening with merger when duration of a long consonant is < 100 ms, since long consonants have length variation with range 105-200 ms, short consonants 60-95 ms. The cousin enters into known variability ranges, the patient is significantly more accurate, an accuracy linked with the excessive number of pauses she makes: tendentially her story style = word-list style. The cousin has no problems with [pr,br,tr,dr,gr] stop + sonant groups. Loose Romance coarticulation seems to present a problem for our patient, who sometimes subtitutes with close coarticulation of the Germanic type: in the case of [pr] we have instances of vowel insertion [pə̆ɹ], [pə̆ɾ] or r-simplification [pw] (4/13), [br] shows approximant assimilation as [bʊ] (2/2), [tr] close Germanic coarticulation [$t^h\underline{ɻ}$] (1/9), as does [dr] > [$t^h\underline{ɻ}$] (1/3), the [gr] cluster simplification as [ʔʀ], [ʏ] (4/6).

Many such problems are being investigated.

Here we only list consonant realizations different from usual Italian ones and which characterize the patient but not her cousin. For usual Northern intervocalic / intersonant lenition of the type /b/ > [β], /d/>[ð, δ], /g/> [ɣ, ɰ] we have North-Eastern female pop. (pop. N=27) /b/ $\bar{x}$ = 10% (σ=7), /d/ $\bar{x}$ = 29% (σ=13), /g/ $\bar{x}$ = 43% (σ=21); our patient has /b/=11%, /d/=5%, /g/=29%, her cousin /b/=47%, /d/= 13%, /g/= 18%. In the case of the most frequent voiced stop /d/ our patient is the most accurate female speaker. Cases of non- intervocalic anomalous lenition and strengthening in the patient's speech also noted were: Case 1 /p/> [ɸ,f] 7 / 94=7%, /t/ > [θ] 7 / 94 =7%, /k/ > [h, x] 7 /111=6%, μ= 7%; Case 2 /p/>[p̪] 3 / 94= 3%, /t/> [t̪] 4 / 94= 4%, /k/>[g]3 / 111=3%, μ=3%; Case 3 /b/ > [b̪],[p̪] 9 / 31= 29%, /d/ > [d̪],[t̪], 21 / 86 = 24%, /g/ > [g̪],[k̪], 3 / 14 = 21%, μ=25%. Cases 1-2 have percentages not significantly different from 0% at the 99% confidence level, while the same intervall for Case 3 ($\bar{x}$ ± 2.57 σ) is significantly so (min. 15%, max. 35%). This and other problems connected with voicing and VOT in the production of voiced segments tend to represent an overall "Germanic" effect for Romance speakers. Other problems regard the glottalization of /k,g/ as [ʔ], laryngealization of words or entire phrases ("un tantino la voce", "nella foresta", "spaventata"), aspirating fricatives /f,s/, preconsonantal loss of white noise components in /s/ > [θ],[h], word-initial voicing or affrication of /s/,i.e. "sotto" ['tsɔt:ɔ], "Sissignore" [zizi'ɲoɾe], /r/ produced as a large variety of approximants or even fricative [z], /l/ realized as [d], [ð], [δ], [l,], [ʎ] , though the more complex /ʎ/ is always [ʎ].

Vowel / syllable erosion occurs as if Italian were stress- timed, e.g. p(a)reva, v(e)stita, ch(e) (a)veva (a)ppena, (a)bbracciarti, (a)bitava, p(e)ric(o)loso, repeated examples of f(o)caccia. Post-stress cancellation is rarer but present (sure cases are sub(i)to, stav(a)no). There are cases of badly executed consonant clusters with apparently cancelled segments and occasional stress shifts which drastically modify stress patterns in phrases ("si míse a córrere"> "si míse a corrére", "la piccína seguìva l'áltra" > "la píccina seguìva l'áltra").

There is still a major problem to be investigated: the patient's use of pauses as compared with that of normal speakers (cousin). This has to be reconsidered from the viewpoint of intonation contours. At the moment it suffices to say that the patient presents ca. 50% more pauses in the same text than her cousin (78 vs. 53): such pauses are, on average, considerably longer (average 598 ms vs. 449 ms for her cousin) and in some cases they even break up MINOR syntactic constituents.

4. *First conclusions.* From the point of view of her basic units the patient does not seem to have lost elements of her linguistic SYSTEM, her intonation contours seem to correspond to major syntactic constituents dealt with by usual strategies, though in CERTAIN respects she is more accurate than normal speakers.In other words, her native competence is not impaired: the English speaker, instead, shows NO competence of Italian prosodic strategies, with both UNNATURAL prosodic and syntactic breaks. The patient seems to be segmenting in terms of very small units at all levels of her linguistic organization, with a resulting overaccuracy in a number of phenomena (lenition, close coarticulation of certain segments in clusters: some general anti-Romance trends), while widely missing her targets in other cases. From this latter point of view she presents serious production disturbances: too frequent pauses, inability to exercize a consistent control of physical parameters over long periods or wherever there is linguistic complexity. She has obvious respiration and timing problems; variations in subglottal pressure during performance ought also to be measured, since this may be related to F_0 alteration, overshooting and resetting.

REFS. [1] C.Avesani (1987), *Dalla parola al discorso.* Verso un modello della 'declinazione' intonativa in italiano, Ph.D. thesis, Scuola Normale Superiore, Pisa.

[2] P.Bertinetto (1976), L'accento secondario.....Analisi teorica e sperimentale, in AA.VV. *Studi di fonetica e fonologia*, Rome, 189-236.

[3] S.E.Blumstein et al. (1987), On the Nature of the Foreign Accent Syndrome: A Case Study.*Brain and Language* 31,215-44.

[4] N.R.Graff-Radford et al. (1986), An unlearned foreign 'accent' in a patient with aphasia, *Brain and Language* 28, 86-94.

[5] J.C.L.Ingram et al. (1992), Phonetic analysis of a case of foreign accent syndrome, *Journal of Phonetics* 20,457-74.

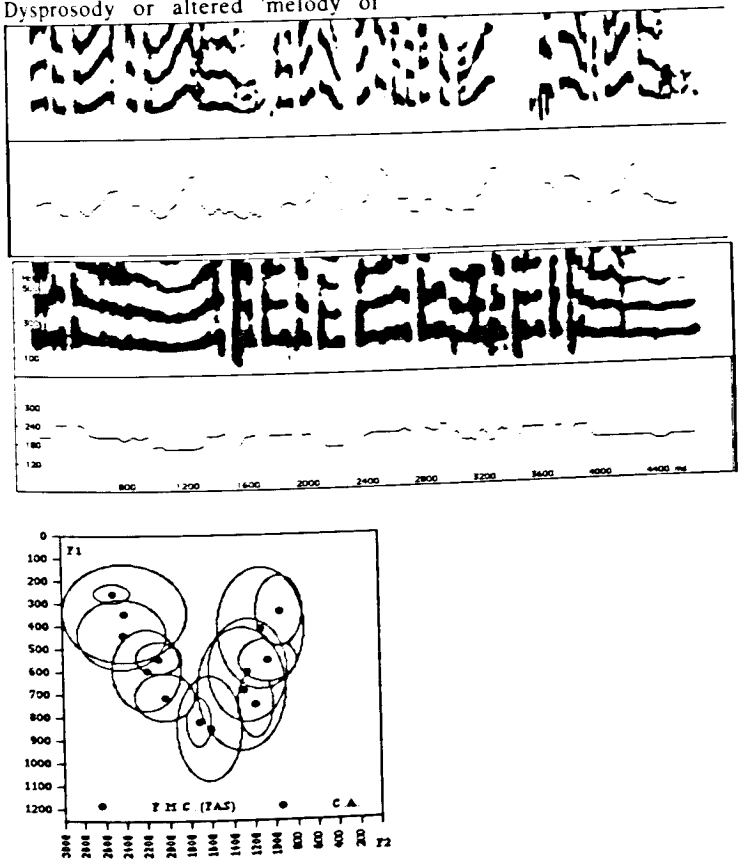
[6] G.H.Monrad-Krohn (1947), Dysprosody or altered 'melody of language', *Brain and Language* 70, 405-15.

[7] L.Romito (1994), Cenni sui correlati elettroacustici dell'accento..... in P.L.Salza (ed.), *Gli aspetti prosodici dell'italiano*, AIA 21, 107-19.

[8] J.Trumper et al. (1990), Il problema della varietà..., in A.M.Mioni et al. (eds.), *L'italiano regionale*.SLI 25, Rome, 159-91.

[9] J.Trumper et al. (1991), Vowel systems and areas compared.... in E. Caldognetto Magno et al. (eds.), *Interfaccia tra Fonologia e Fonetica*, Padua.

[10] H.Whitaker (1982), Levels of impairment in disorders of speech, in R.N.Malatesta et al. (eds.), *Neuropsychology and Cognition 1*, The Hague.



PERCEPTION OF FORMANT TRANSITIONS IN SYNTHESISED VOWEL PAIRS

W.A.Ainsworth
Dept. of Communication and Neuroscience, Keele University, Staffordshire, UK

ABSTRACT

Speech analysis shows that the second formant transitions in vowel-vowel utterances are not always of the same duration as those of the first formant transitions nor are they always synchronised. Moreover the formant transitions often move initially in a different direction from their final target. In order to investigate whether these deviations from linearity and synchrony are perceptually significant a series of listening tests have been conducted with the vowel pair /ɑ/-/i/.

INTRODUCTION

Speech is produced by a series of articulatory gestures which give rise to formant transitions in the spectrograms of the resulting sounds. It has long been known that that formant transitions are important cues for the perception of many speech sounds, especially voiced consonants [1,2]. These early perceptual studies provided the basis for speech synthesis-by-rule systems [3,4].

These systems were based on formant synthesisers incorporating rules involving simultaneous formant transitions. More recent synthesis systems based on vocal tract models [5,6] generate sounds whose formants change in a nonlinear manner. Careful analysis of natural speech also demonstrates nonlinear and nonsynchronous formant transitions.

The question arises as to whether these departures from linearity and synchrony are perceptually significant. In order to investigate this question a series of experiments have been performed to measure the perceptual tolerances of formant transitions.

ANALYSIS OF VOWEL-VOWEL TRANSITIONS

The six vowels /i, ɛ, ɑ, o, u, ɜ/ were spoken as vowel pairs V_1V_2 in all possible combinations by a male speaker. With many of the formant tracks F1 and F2 changed simultaneously. These give rise to fairly linear transitions in the F1-F2 plane. Other formant tracks, however, show systematic departures from linearity. In the case of /ɛ/-/o/, for example, the path first moved parallel to the F1-axis then to the F2-axis.

Such departures from linearity can arise in two ways. In the case of /ɛ/-/o/ the transition of the first formant is completed in the first 130 ms of the sound whilst the transition of the second formant begins at about this point. In other cases, /ɛ/-/ɑ/ for example, both formants begin and end at approximately the same time but for a time they move in a different direction from their final target.

EXPERIMENTAL PROCEDURE

All the experiments were carried out with the vowel pair /ɑ/-/i/.Four listeners took part in these experiments. Four series of experiments were carried out: the first three involving temporal properties of the transitions and the fourth involving the shapes of the transitions.

The stimuli were two-formant synthetic sounds generated by passing a sequence of pulses with a repetition frequency of 120 Hz through digital resonators.

The stimuli were stored in digital files and presented to listeners via headphones by means of a 16-bit PC sound system. Nine sounds were generated for each

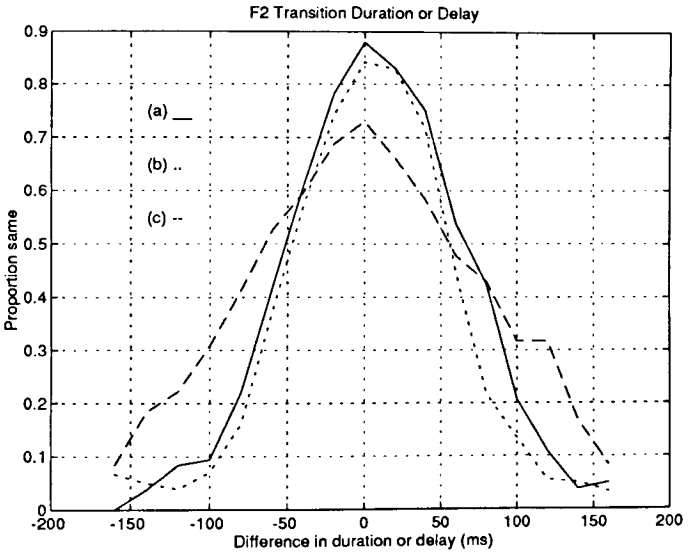


Figure 1. Proportion of stimuli judged the same as a function of the difference between (a) the duration of the F2 transition for simultaneous starts or ends, (b) the delay of the F2 transition with respect to the F1 transition and (c) the duration of the F2 transition for symmetric transitions with respect to the transition of F1 of stimulus B and the duration or delay of the F2 transition of stimulus A.

experiment. They were presented in pairs five times in random order with an ISI of 2 s. Listeners were asked to press 'S' on a keyboard if the two sounds of the pair were judged to be the same and 'D' if they sounded different.

EXPERIMENTS

Transition duration

The stimuli in the first experiment consisted of the /ɑ/-/i/ sounds with an initial steady frequency of F1 of 900 Hz for 100 ms, a linear transition for 100 ms, then a final steady frequency of 250 Hz for another 100 ms. The second formant frequency, F2, was 1100 Hz for 100 ms, then a transition whose duration varied from 20 ms to 180 ms in 20 ms steps, followed by a final steady frequency of 2500 Hz for sufficient time to make the total duration of the sound 300 ms.

It was found that when the F2 transition durations differed by less than about 70 ms half or more of the sounds were judged to be the same.

In the first experiment all the transitions began 100 ms from the beginning of the sound. A further experiment was conducted in which the stimuli were the same as those in the first experiment except that the second formants all ended at the same point as the end of the F1 transition. Similar results were found. The combined results are shown in Figure 1.

Transition delay

The experiments on transition duration confounded two possible cues: the length (or slope) of the F2 transition and the synchrony, or lack of it, between the F2 transitions and the start or end of the F1 transition. In order to explore the effects of this synchrony further experiments were performed in which the durations of the transitions remained constant but the delay between the start of the F2 transition and the F1 transition was varied.

In the first experiment F1 consisted of three segments: a steady segment, a transitional segment and a further steady segment all of 100 ms duration as before. F2 also consisted of three segments with the middle transitional segment remaining constant at 100 ms duration. The first segment, however, varied from 20 ms to 180 ms and the last was adjusted to make the total duration 300 ms.

The average results are also shown by in Figure 1. It can be seen that if the delay between the start of the F2transition and that of the F1 transition was less than about 50 ms half or more of the sounds were judged to be the same.

The experiment was repeated with the durations of both the F1 and F2 transitions at 50 ms or 150 ms. It appeared that there were no systematic differences so that two sounds have a 50% or greater chance of being judged the same if their formant transitions are synchronised to within 50 ms.

Symmetric transitions

In the next series of experiments an attempt was made to measure the effect of transition duration or slope independent of transition synchrony. This was done by generating a set of stimuli whose F2 transition durations varied but whose mid-point remained at the mid-point of the F1 transition.

In the first experiment F1 consisted of three 100 ms segments. The middle F2 segment consisted of the transition and varied from 20 ms to 180 ms. The initial and final F2 segments were equal and chosen so that the total duration of the sound was 300 ms.

This experiment was repeated twice: once with an F1 transition duration of 50 ms embedded between two 100 ms segments and once an F1 transition duration of 150 ms between two 50 ms steady segments. The duration of the F2 transition duration varied from 20 ms to 180 ms as before and the durations of the initial and final segments were chosen to make the total duration 250 ms.

The results of these experiments are also shown in Figure 1. Once again it appears that the duration of the F1 transition, at least in the range 50-150 ms, has little effect on the averaged judgements.

Transition shape

Finally the effect of F2 transition shape was examined. F1 consisted of three 100 ms segments. However the F2 transition was divided into two 50 ms segments with the frequency of F2 at the boundary between them varying from 1000 Hz to 2600 Hz in 200 Hz steps. At the centre of this range the mid frequency of F2 is 1800 Hz giving a linear transition which matches those employed in the previous experiments.

The results showed that a deviation of about 750 Hz in the mid point of the F2 transition can occur before a listener reliably distinguishes between two sounds with this structure.

DISCUSSION

Experiments have been performed to estimate the tolerance of the perceptual system to the duration and delay of transitions in vowel-vowel utterances. The effect of transition shape has also been examined.

A lead or lag of some 50 ms is required for two sounds to be reliably distinguished. Greater differences are required for transitions of different durations. If the F2 transitions are symmetric with respect to the F1 transitions a difference of about 80 ms is required, but if the F2 transitions begin or end simultaneously with the F1 transitions a difference of only about 70 ms is required.

There is evidence that formant transition duration and shape are important in consonant-vowel transitions [1,7]. It therefore remains to be seen whether similar perceptual tolerance values apply to consonant sounds.

CONCLUSIONS

A number of experiments have been performed to estimate the difference limens for the duration, synchrony and shape of formant transitions in a two-formant synthesised vowel pair /ɑ/-/i/. It seems unlikely that the deviations from linearity and synchrony observed in natural vowel-vowel pairs have any perceptual significance.

ACKNOWLEDGEMENTS

The work was supported by EC Science Contract SC1-CT92-0786.

REFERENCES

[1] Liberman, A.M, Delattre, P.C., Gerstman, L.J. & Cooper, F.S (1956) Tempo of frequency change as a cue for distinguishing classes of speech sounds, J.Exptl.Psych., 52, 127-137.

[2] Lisker, L. (1957) Minimal cues for separating /w,r,l,y/ in intervocalic position, Word, 13, 256-267.

[3] Holmes, J.N., Mattingly, I.G. & Shearme, J.N. (1964) Speech synthesis by rule, Language and Speech, 7, 127-143.

[4] Klatt, D.H. (1980) Software for a cascade/parallel formant synthesiser, J.Acoust.Soc.Am., 67 (3), 971-995.

[5] Maeda, S. (1990) Compensatory articulation during speech; evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model, in Speech Production and Speech Modelling (W.J.Hardcastle and A.Marchal, eds.), NATO ASI Series, Kluwer Academic Publisher, Dordrecht.

[6] Mrayati, M., Carré, R. & Guérin, B. (1988) Distinctive regions and modes: a new theory of speech production, Speech Communication, 7, 257-286.

[7] Ainsworth, W.A. (1968) First formant transitions and the perception of semivowels, J.Acoust.Soc.Am., 44, 698-694.

A PERCEPTUAL STUDY OF REDUCED VOWELS IN CLEAR AND CASUAL SPEECH

Moon S-J[1], Lindblom B[2] and Lame G[3]
[1]Department of English, College of Liberal Arts, AJOU University, Suwon 442-749, Korea
[2]Department of Linguistics, Stockholm University, Stockholm S-10691, Sweden
[3]Eloquent Technology Inc, 24 Highgate Circle, Ithaca, NY 14580, USA

ABSTRACT

Data are presented on the perception of vowel formant patterns in [w_l] syllables. Perceptual judgements depended on extent of formant transitions, vowel duration and speaking style indicating that listeners do expect undershoot in syllables of this type and that they expect less of it in clear than casual speech. Recognition scores were highest for the most extensive formant movements, particularly in the clear speech condition.

PROBLEM

In a recent study, English vowels in [w_l] syllables were found to be longer, have faster formant transitions and show more peripheral formant patterns in clear speech than in casual style[1]. These changes had the effect of reducing context and duration dependent "formant undershoot" thus shifting formant values closer to ideal "context-free" values. Such decrease in context-dependence makes intuitive sense perceptually, since the phenomena of undershoot tend to reduce intervocalic contrast and therefore create potential difficulties for the listener. However, firm conclusions can only be drawn given data from native listeners. The following experiment was done to shed some light on the perceptual function of the observed formant variations.

EXPERIMENTAL PROCEDURES

The English vowels /i/, /ɪ/ and /ɛ/ were synthesized and embedded in one casually and one clearly spoken version of "Wheelingham", a possible place name. The reason for choosing that context was that, previously[1], undershoot effects had been found to be particularly marked in trisyllabic words. This hybrid synthesis was implemented using KLSYN88 and other software on one of the Vaxstations in the University of Texas Phonetics Laboratory.

The vowel formant patterns (F1, F2, F3) of the synthetic stimulus portions were derived from stylized values based on average data of five speakers[1] plotted in Figure 1, comprising formant values sampled at the frequency maximum of F2 in the [w_l] context. A relatively higher position on the chart implies a larger upward movement of F2 from its low starting point in [w], and a correspondingly larger downward shift back to the low F2 of [l]. Similarly, a relatively higher F1 value means a bigger frequency excursion relative to its location in [w] and [l]. The F3 values are not shown, but were varied according to a similar, but more limited, excursion pattern.

The following aspects were also defined in terms of averages calculated from the earlier speech sample[1] and were obtained for both styles by pooling across tokens, vowels and speakers: timing of formant frequency maxima, contours of F0 and overall amplitude.

To generate the transitions to and from the vowel pattern, smooth cosine functions were used. Great care was taken to ensure continuity at segment

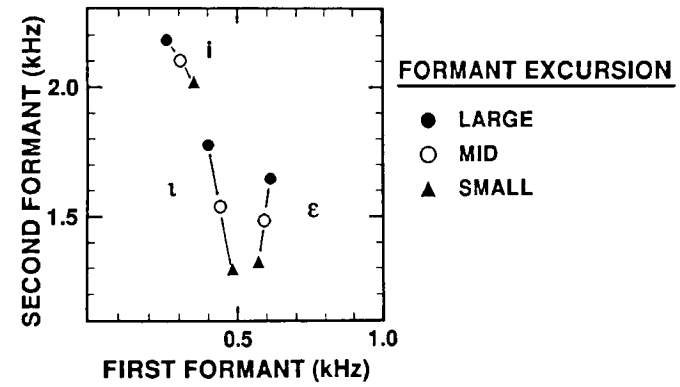


Figure 1. Formant values at approximately midpoints of synthetic vowels. Smooth cosine functions were used as transitions between these patterns and the adjacent [w] and [l] "loci". The entire vowel segments were spliced into a clear and a casual variant of the word ['wɪlɪŋhæm].

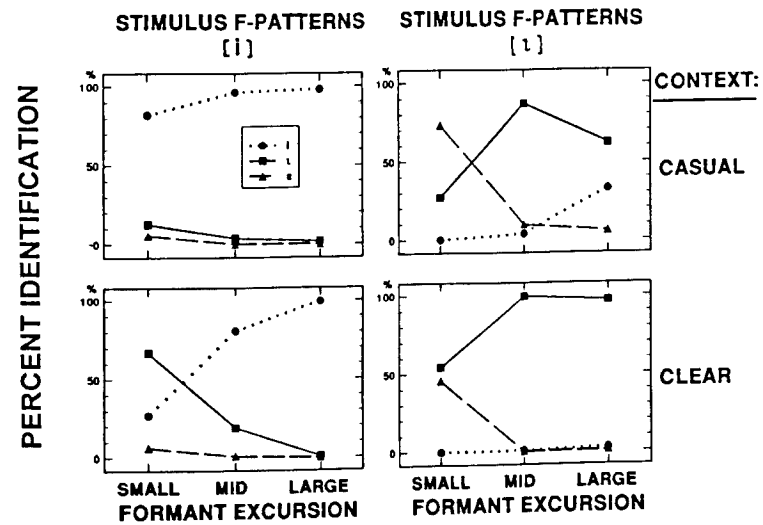


Figure 2. Identification results for stimuli derived from /i/ measurements in the left column and for /ɪ/ in the right. The vertical axis percent identification. Along the x-axes extent of formant transitions. Solid dots, squares and triangles refer to /i/, /ɪ/ and /ɛ/ responses respectively.

boundaries in all parameters including overall amplitude, F0 and voice source characteristics (open quotient and spectral tilt). The synthesized vowel was spliced into the natural speech using cosine tapering at the edges.

A total of 72 stimuli were produced combining 2 contexts (clear and casual), 3 vowels (/i/, /ɪ/, /ɛ/), 3 degrees of formant excursion (or undershoot) and 4 vowel durations. Each vowel had all combinations of three central formant patterns (small, medium and large undershoot) and four durations, 136, 177, 212 and 266 msec chosen to reflect the observed effects of the clear-casual and the open-close conditions.

Five listeners participated. They were graduate students selected as a homogeneous group of American English speakers without marked dialectal features. Each subject participated in three experimental sessions in which they were asked to idetify the stimulus as "Wheelingham", "Willingham" or "Wellingham". The test tapes contained a total of 576 stimuli distributed over the three sessions in randomized sets of 216+216+144 stimuli. There were eight repetitions of every stimulus which means that forty responses per stimulus were collected. The percentages to be reported were calculated with n=40.

RESULTS

Our findings indicate that listener judgements varied as a function of all the variables investigated. Vowel duration, context (casual or clear style of the natural frame), extent of formant transitions and the intended vowel category were all seen to have an effect on the responses.

If "correct identification" is defined in terms of the vowel categories from which the stimuli were derived, we find that all patterns intended as /ɛ/ were identified at a level near 100%. On the other hand, errors were numerous in the labeling of the /i/ and /ɪ/ stimuli. Figure 2 presents the identification results for /i/ in the left column and for /ɪ/ in the right. The vertical axis in all four panels is percent identification. Along the x-axes extent of formants is shown. Here "small", "mid" and "large" excursion corresponds to the triangle, open circle and filled circle respectively of Figure 1. A small formant movement implies a large "undershoot" effect, a large movement means small "undershoot". Data points pertain to percent identification averaged across all four durations. Solid dots, squares and triangles refer to /i/, /ɪ/ and /ɛ/ responses respectively. The two panels of the upper row show how the stimulus F-pattern was classified in the context of the casual frame. Those of the bottom row give the results for the same stimulus pattern in the clear frame.

First note that, for large formant excursions, the recognition score for /i/ is high in both styles. For /ɪ/, identification is near perfect in clear speech but less accurate in the casual context. Note that comparing the upper and lower panels we examine the effect of the style of the frame. Evidently, since for any given vertical position, the top and bottom panels refer to identical synthetic vowel segments, the non-identity of the panels indicates that context has an effect on the responses.

As for small formant excursions, we note that /i/ in casual style is identified resonably well, but in clear speech its score is drastically reduced. Instead /ɪ/ responses are favored. Figure 1 suggest how this effect could be accounted for. It implies that, in the clear context, the small excursion for /i/ (=large undershoot) is interpreted as an instance of /ɪ/ with a large formant movement (=small undershoot). Note the proximity

of the /i/-triangle and the solid /ɪ/-dot of Figure 1. Thus listeners did not expect /i/ to exhibit that much undershoot in clear speech.

That interpretation also clarifies the results for /ɪ/ in the casual context where /i/ responses are seen to increase at the expense of /ɪ/. That the large formant excursions of /ɪ/ get associated with /i/ to some extent would seem to suggest that listeners did not expect an /ɪ/ to show that little undershoot in the casual context.

Looking at the /ɪ/ with a small excursion we find that it is identified as an /ɛ/, especially in the casual context. Figure 1 shows that the /ɪ/ and /ɛ/ stimuli with small excursions are very similar. F1 of this /ɪ/ variant is high.

The strongest duration effects are seen in the responses to the /ɪ/ stimuli. For the /i/ and /ɛ/ stimuli they are more or less absent. As the duration of /ɪ/ is increased, /ɛ/ responses are more and more favored. This pattern is particularly evident in the case of the /ɪ/ variant with small formant excursion. As indicated in Figure 1 this stimulus is very close to the /ɛ/ with the least extensive formant transitions. These findings are compatible with the expectation that, being a more open vowel, /ɛ/ should also be longer. Apart from this /ɪ/-/ɛ/ interaction, duration effects are small indicating that, for this experiment, they were not strong enough to overrule the apparently more important information on formant pattern and the style of the frame.

SUMMARY

The undershoot phenomena evident in previously reported acoustic analyses[1] show up also in the present investigation. Error patterns clearly suggest that listeners expect these effects in [w_l] syllables of the present type. The expectation is that there will be more undershoot in casual than in clear style.

As formants move further and further away from [w] towards an underlying hypothetical "target pattern", recognition scores improve *provided that* the contextual information is consistent. In other words, formant transitions are not judged in absolute terms. Their extent is determined relative to the context. Accordingly, a given pattern can be judged as extensive in casual style but as reduced in clear.

Finally, we note that the vowels /i/ and /ɪ/ when clearly spoken were least confused in the clearly spoken context.

CONCLUSION

Do the present results support the suggestion that the patterns of undershoot observed in clear and casual [w_l] syllables[1] are to be explained as instances of "undershoot compensation" and "facilitation of listener's task"? The present results appear compatible with such a claim, since identification was highest for the most extensive formant movements, particularly in the clear speech condition.

REFERENCES

[1] Moon S-J and Lindblom B (1994): "Interaction between duration, context and speaking style in English stressed vowels", *J Acoust Soc Am* 96(1):40-55.

ACKNOWLEDGEMENTS

The present research was supported by four sources: (i) A grant from the Advanced Research Program of the *Texas Board of Coordination*; (ii) grant No. BNS-9011894 from the *National Science Foundation*; (iii) *Projet Sciences Européen* (ERB4002PL910339); and (iv) by project F 770/93 sponsored by HSFR, *Humanistiska Samhällsvetenskapliga Forskningsrådet* of Sweden.

IMPLICIT MEMORY FOR SILENT-CENTER SYLLABLES

Susan L. Hura
Department of Audiology & Speech Sciences, Purdue University

ABSTRACT

Recent work has shown that specific characteristics of spoken words are encoded in memory. Exposure to a particular token of a word improves listeners' ability to identify the word later . This study extends this paradigm to silent-center (SC) stimuli, in which vowel nuclei are attenuated to silence. Results address the claim that CV transitions in SC syllables are privileged in perception.

INTRODUCTION

A growing body of evidence suggests that information about specific voices and specific exemplars of words is encoded in memory [1]. The effects of these individual memory traces are best observed in implicit memory tasks in which listeners perform some perceptual classification of stimuli. These findings are in opposition to traditional theories of lexical representation which rest on the assumption that there is a single, abstracted entry for each item in the lexicon. Moreover, the implicit memory findings speak against the traditional belief that individual characteristics in the speech signal are a hindrance to perception, in that they will be difficult to match up with abtstract lexical or phonemic catgegories.

LEXICAL ACCESS AND IMPLICIT MEMORY

Current theories of lexical representation posit a single, abstracted phonemic representation for each item in the mental lexicon (e.g.[6]). Recognizing spoken words is thus a process of matching the variable incoming signal with stored canonical representations. Describing the process by which this match takes place has been the goal of theories of speech perception. In many cases, some invariant cue associated with each phoneme is proposed. One of the most recent claims, put forth by proponents of the theory of dynamic specification, is that dynamic vowel information contained in consonant-vowel transitions may be an invariant cue for vowel recognition [4].

Contrary to the traditional view, recent work in the field of implicit memory suggests that there may be multiple entries in the lexicon for each item. There is evidence that detailed information about specific voices and exemplars of words may be encoded in memory and used in speech perception. Goldinger [1] showed that previous exposure to particular tokens of words significantly improved listeners ability to accurately identify those words later. This advantage decreased over time, but was still observable after a delay of up to a week. These results suggest that listeners retain information about particular perceptual events and use this information in later perceptual judgments. Goldinger's study is part of the larger field of implicit memory [3, 5] which suggests that individual episodic memory traces are formed for many perceptual events.

DYNAMIC SPECIFICATION OF VOWELS

Results from Hura [2] suggest that implicit memory effects apply to silent-center (SC) versions of syllables in which vowels have been attenuated to silence. When listeners were exposured to a mixed set of unaltered Full syllables and the corresponding SC syllables, identification performance on the SC stimuli was significantly better than when listeners heard SC syllables alone. This implies that individual episodic traces of Full syllables were encoded in memory and accessed in identification of SC versions. The current study seeks to replicate this finding in a more controlled fashion.

The theory of dynamic specification characterizes vowels as gestures having intrinsic timing parameters. This theory rests on studies of silent-center stimuli, in which syllable nuclei, including static formant patterns of vowel targets, have been removed. The remaining dynamic information often enables listeners to identify vowels in SC syllables as accurately as vowels in full syllables. Consequently, dynamic portions of the speech signal are seen as privileged in perception. Moreover, hybrid SC syllables, which are composed of the initial CV transition from one talker and the final VC transition from another talker, may in some cases yield very accurate vowel perception [4]. This suggests that dynamic vowel information may be speaker-independent, thus greatly simplifying the problem of matching the incoming speech signal with stored canonical representations.

The study reported here is an initial step towards extending implicit memory paradigms to silent-center syllables. Listeners were presented with a set of unaltered CVC syllables for identification. After a variable delay, listeners were tested on a set of silent-center syllables, half of which were part of the set of Full syllables heard originally. Performance on previously heard syllables is compared with new syllables. Effect of length of delay between the two sets of stimuli is also considered.

METHOD

Stimuli

The stimuli for this study began with a set of 10 /bVt/ syllables spoken by an adult male talker. Nine of the ten syllables are real English words: /bit/, /bɪt/, /bet/, /bɛt/, /bæt/, /bɑt/, /bət/, /bot/, and /but/; the tenth syllable, /bʊt/ is a phonologically-possible nonword. Syllables were produced in isolation, and recorded on audio tape using a Sony TC-FX420R cassette recorder and an Audio-Technica ATM11 unidirectional microphone. A single version of each syllable was low-pass filtered at 8 kHz, and digitized at 20 kHz using Kay CSL equipment.

For each syllable, Full and SC stimuli were created. Full syllable stimuli consisted of whole, unmodified /bVt/ syllables, including prevoicing of the /b/ and release burst of the /t/, if present in original recordings. Syllable duration, for the purposes of generating SC stimuli, was measured from the release of initial /b/ closure to onset of final /t/ closure. Silent-center stimuli were created by attenuating to silence all but the first three and last four pitch periods of each syllable.

Each SC syllable was then embedded in white noise, with a 50 msec noise lead before onset of the syllable to guard against overshoot of masking. The amplitude of the noise used with each SC syllable was adjusted to match the energy level of the corresponding Full syllable, yielding a 0 dB S/N in each case. Due to naturally-occurring differences in the overall amplitude of vowels, SC-in-noise stimuli were not at equal dB values across the stimulus set.

In all 20 stimuli were constructed, a Full and SC-in-noise version for each of 10 /bVt/ syllables. Two listening tests were constructed for Full syllable stimuli. Each was composed of 20 repetitions each of 5 of the 10 Full syllable stimuli presented in random order. Test Full 1 contained the syllables /bit/, /bɛt/, /bɑt/, /bot/, and /bʊt/; test Full 2 contained the syllables /bɪt/, /bet/, /bæt/, /bət/, and /but/. Syllables were separated such that tense and lax versions of vowel pairs would occur on different tapes (e.g. /bit/ versus /bɪt/), and such that each tape would contain an equal number of tense and lax vowels overall. Items were presented in blocks of 10,with 3500 msec ISI and 7 sec between blocks. There were a total of 100 items on each tape, which lasted approximately 9 minutes.

A single listening test was constructed for SC stimuli. Twenty repetitions of each of the 10 stimuli occurred in random order. As in the Full syllable tests, stimuli were presented in blocks of 10, with 3500 msec ISI, and 7 sec between blocks. The SC tape lasted approximately 18 minutes. All three listening tests were recorded onto digital audio tape using a Sony DCT-690 DAT deck.

Subjects

Thirty-one undergraduate students enrolled in a speech acoustics course at Purdue University served as listeners. All were native speakers of English and none reported a history of speech or hearing problems. They received partial course credit for their participation.

Data from 28 listeners was included for analysis. The remaining listeners were excluded because they produced greater than 10% errors on the initial Full syllable test.

Procedures

Listeners were tested individually ina laboratory setting. Before the initial test, listeners read a set of printed instructions and were allowed to ask questions. Stimuli were presented via Sony MDR-V400 dynamic stereo headphones at a comfortable listening level. Listeners performed an identification task, and responded by pressing a single key on a computer keyboard which was marked with the /bVt/ word. In this initial phase, listeners heard either tape Full 1 or tape Full 2.

A delay of 5 minutes or 1 day elapsed between the intial Full syllable testing and testing on SC stimuli, with half the subjects in each delay condition. Listeners again read a set of printed instructions before the test, and also heard a short introduction tape. This tape contained 10 Full syllables spoken by an adult female talker, each embedded in wide-band noise. The purpose was to familiarize listeners with the sound of the noise. After asking any questions they had, each listener identified the stimuli on the SC tape, making their responses as for the Full syllable test.

RESULTS

Listeners' responses on both tests were scored for percent errors. Results of the Full syllable tests were used as a criterion for subject inclusion: data from any listener who made more than 10% errors identifying unmodified Full syllables was excluded.

Listeners' responses on the SC test were also scored for percent errors, which averaged 38.5% overall. Listeners were treated as two separate groups, depending on which Full syllable test they heard. Full 1 listeners averaged 39.3% errors; Full 2 listeners averaged 37.2%. Listeners were also classified according to delay between tests: listeners in the 5 Minute condition averaged 37.4% errors; listeners in the 1 Day condition averaged 39.7% errors. There were also differences in error rate across individual vowels, ranging from a high of 92% on /u/ to 1.4% for /æ/. In all subsequent analyses, results reported are collapsed across vowels.

A 2 X 2 X 2 repeated measures analysis of variance was conducted, with Full syllable tape and Delay as between subjects factors and repetion (i.e. Old versus New) as a within subjects factor. There was no significant main effect of Full syllable tape ($F(1,24) = .98$, NS), or for Delay ($F(1,24) = .77$, NS). Repetition, however, was significant ($F(1,24) = 6.73$, $p>.02$). There were significantly fewer errors on previously heard (Old) stimuli than on New stimuli (35.8% versus 41.3%, respectively).

There was only one significant interaction, that of Full syllable tape by Delay ($F(1,24) = 24.5$, $p>.01$). That is, the effects of repetition of stimuli differed for the two Full syllable tapes. Table 1 lists the average percent errors for the four Full syllable tape by repetition conditions. These data show that the repetition effect is stronger for Full 2 stimuli than for Full 1.

Table 1. Mean percent errors for Old and New stimuli for each Full syllable tape.

	Old	New
Full 1	42.3	37.3
Full 2	29.2	45.2

Although the higher order interactions do not achieve significance, trends for effects of Delay can be observed. Specifically, the difference in error rate between Old and New stimuli is greater for the 5 Minute delay than for the 1 Day delay. Table 2 lists the average percent errors for Old and New stimuli at each delay.

Table 2. Mean percent errors for Old and New stimuli for each Delay.

	Old	New
5 Minutes	33.3	41.5
1 Day	38.2	41.1

DISCUSSION

These results provide support for the claim that previous exposure to particular versions of spoken words can improve a listener's ability to correctly identify the word later . There is a strong tendency in these data for previously heard words to be more accurately recognized than words that had not been previously heard. Recall, however, that nine of the ten /bVt/ syllables tested are reasonably common words, with which all subjects were familiar. Therefore, the difference between "old" and "new" stimuli in this study is achieved solely by exposure to the words during the Full syllable test. This is strong evidence for the claim that listeners are accessing a memory trace particular to each token, rather than simply matching a single canonical representation in the lexicon.

Moreover, this study shows that when a memory trace is set up for a Full syllable, it is later accessed in identification of a SC version of the syllable. This is evidence that the portion of the syllable maintained in SC's retains some of the distinct, individual characteristics of the Full syllable. That is, the partial pattern of an SC syllable appears to be easily related to the complete Full syllable pattern for particular tokens. The question that remains, however, is whether any other partial stimulus might be as easily related to the complete memory trace. This is one question for future research in this area.

Another important future question is whether memory traces may be set up for SC syllables. That is, if a listener first hears an SC syllable, will there be an advantage in later perceptual classification of the corresponding Full syllable? Such data would allow us to make a more definitive statement about the dynamic portions of the syllable contained in silent-centers.

REFERENCES

[1] Goldinger, S. (1992), *Words and voices: Implicit and explicit memory for spoken words*. (Technical Report No. 7). Bloomington, Indiana: Indiana University, Speech Research Lab.

[2] Hura, S. (1994), Dynamic aspects of vowel perception. Doctoral dissertation: University of Texas at Austin.

[3] Jacoby, L., Marriott, M., and Collins, J. (1990), The specifics of memory and cognition. In T. Skrull and R. Wyer (Eds.), *Advances in Social Cognition, Vol. 3* , Hillsdale, NJ: Erlbaum.

[4] Jenkins, J., Strange, W., and Miranda, S. (1994), Vowel identification in mixed-speaker silent-center syllables. *Journal of the Acoustical Society of America*, vol. 95, pp. 1030-1043.

[5] Schacter, D. (1987), Implicit memory: History and current status. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 13, pp. 501-518.

[6] Studdert-Kennedy, M. (1976), Speech perception. In N. Lass (Ed.), *Contemporary Issues in Experimental Phonetics*, New York: Academic Press.

EVIDENCE FOR THE PERCEPTUAL RELEVANCE OF VOWEL-INHERENT SPECTRAL CHANGE FOR FRONT VOWELS IN CANADIAN ENGLISH

Terrance M. Nearey
University of Alberta, Edmonton, Canada

ABSTRACT

This study assesses the importance of F1, F2 and vowel-inherent spectral change (VISC) in the perception of front vowels in Canadian English using synthetic stimuli with linear F1-F2 trajectories. Results support earlier findings from our laboratories suggesting that VISC plays an important role in the phonetic specification of these vowels.

BACKGROUND

Previous studies in our laboratories [1, 2] confirmed that formant movement plays a role in the specification of phonetic diphthongs such as /e/, which has a diverging F1-F2 pattern (falling F1, rising F2). Results further indicated that the front lax vowels /ɪ/ and /ɛ/ show a converging F1-F2 pattern (rising F1, falling F2), consistent with the targets suggested by Klatt [3] for synthesis of American English vowels. The present study extends our previous findings using synthetic isolated vowels.

PROCEDURE

A four-factor continuum was constructed spanning the front vowels / i, ɪ, e, ɛ, æ/. The variable stimulus factors were onset F1, onset F2, offset F1 and offset F2 frequencies. Onset F1 ranged from 375 to 525 Hz and F2 ranged from 1700 to 2250 Hz (each in four steps). Offset F1 frequencies were manipulated to produce falling (-70 Hz), flat or rising (+70 Hz) linear trajectories for each onset target. F2 trajectories were generated analogously but with 130 Hz rising and falling excursions from the onset values. Vowels were 300 ms in duration with a linearly falling F0 contour (112 to 95 Hz.) Amplitude rose from 35 to 60 dB in the first 60 ms, declined linearly to 58 dB at 150 ms and tailed off to 45 dB at 300 ms. F3 and F4 were fixed at 2500 and 3500 Hz, respectively.

Eleven Canadian English speakers (with some phonetic training) listened to each of the 120 resulting stimuli six times. (Vowels with falling F1 and falling F2 were not synthesized, nor were vowels with the highest level of F1 and F2 onset.) On each trial, listeners identified the category of the vowel and simultaneously assigned a rating (on a scale of 0 to 9) of the "goodness" of each token as a member of the chosen category.

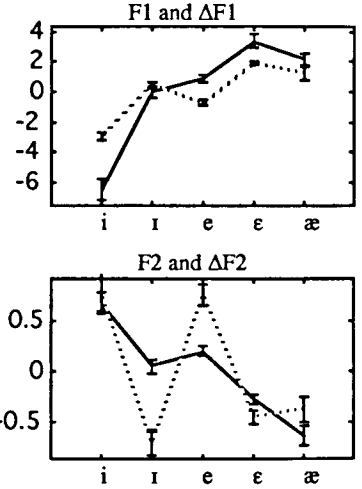


Figure 1. Magnitude (and standard errors) of F1 and F2 coefficients. Solid lines indicate coefficients for onset frequencies, dashed lines for ΔF1, ΔF2.

CATEGORIZATION

Categorization was analyzed using logistic regression, which bears many points of resemblence to the analysis of covariance [4]. Stimulus properties were coded as frequencies (in Hz) of F1, F2, ΔF1 and ΔF2 (corresponding respectively to F1 onset, F2 onset, change in F1 and change in F2 from onset to offset). The coefficients of interest in assessing the relative weight of stimulus properties correspond to vowel-by-stimulus interaction terms. These are plotted as Figure 1. The difference in coefficient values between any two vowels indicates the relative importance of that cue in separating those vowels.

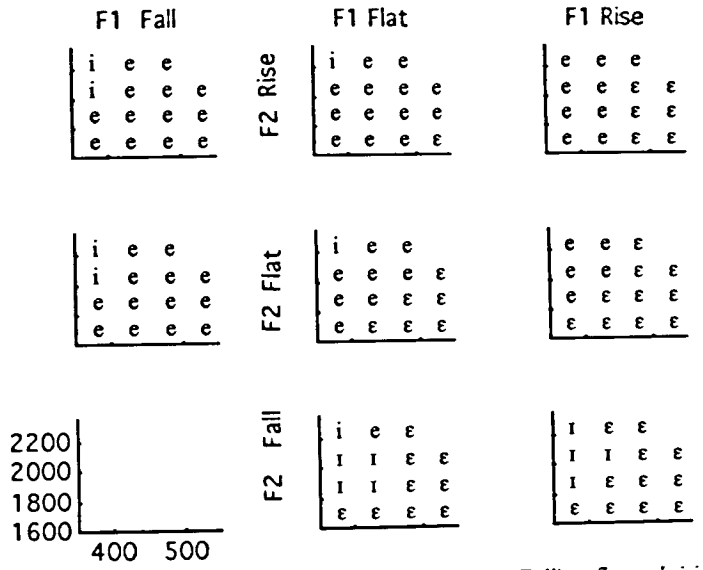


Figure 2. Territorial map for vowels. Columns (left to right): Falling, flat and rising F1 trajectories. Rows (bottom to top): falling, flat and rising F2 trajectories. Stimuli in the lower left corner would possess [w]-like offglides and were not synthesized. Within each cell, the horizontal axis represents F1 (Hz) and the vertical F2 (Hz) of the initial targets of the vowels with frequency values as indicted on the axes of the empty cell.

Random-coefficients regression tests of [5] were applied, testing for the significance of variation in coefficients across vowels compared to inter-speaker differences. Results showed highly significant effects ($p < .0005$) for all stimulus-by-vowel interactions [VxF1: $F(4,7) = 276.2$; VxF2: $F(4,7) = 39.44$; VxΔF1: $F(4,7) = 44.87$; VxΔF2: $F(4,7) = 22.17$.] These tests indicate for F1 and F2 both onset frequencies and changes in frequency are important determinants of listeners' choice of category.

Figure 2 is "territorial map" showing the most frequent response to each stimulus. As expected, formant patterns with falling F1 and rising F2 favored tense /i, e/ responses. There is also a preference for lax /ɪ/ and /ɛ/ with ris-ing F1 and falling patterns. Compatible trends can be observed in the logistic coefficients shown in Figure 1. Vowel categories with relatively higher (positive) coefficient values are favored by a high value of the stimulus property in question over those with lower values. The coefficients of Figure 1 are redisplayed in a 2-dimensional "comet" plot in Figure 3. In this plot (and in Figures 4 and 5) onset frequencies properties are marked with "*" (head of the comet) and offsets by lines (tail). In Figure 3, average offset coefficients have been calculated by adding the ΔF coefficients to the onset coefficients.

The general pattern of the logistic coefficients can be compared to plots like those used in [1] to represent vowel inherent spectral change. Figure 4 shows a comet plot of data from Table II of [1] which is based on the means of production of 10 Canadian English speakers (5 male and 5 female.) Although there are exceptions (discussed further below), the general pattern of means of the onset frequencies is similar as is the direction of movement of formants.

GOODNESS RATINGS

As noted above, "goodness ratings" were also collected for each response. Of primary interest are the locations of patterns where the best tokens of the

vowels occurred. A simple estimate of this can be obtained by computing a "center of gravity" of total goodness votes for each vowels in the F1-F2 stimulus space.

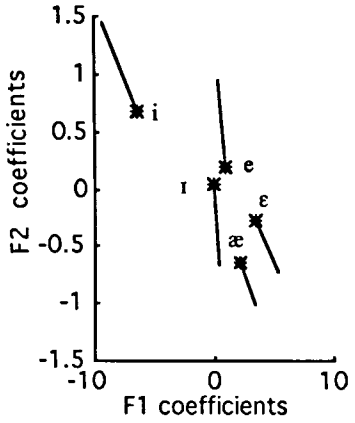


*Figure 3. Plot of mean coefficients from Figure 1 after conversion of ΔF1 and ΔF2 coefficients to frequency offset coefficients. Onset coefficients are marked with *, offsets are unmarked.*

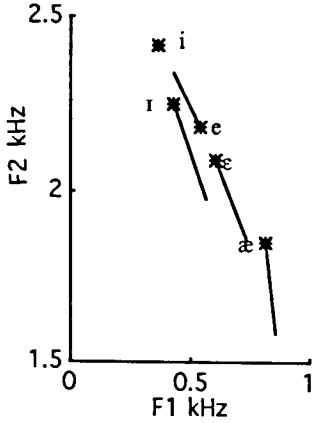


Figure 4 Plot of mean F1 F2 onsets and offsets from Table II of [1]. Labeled ends of lines are onset frequencies and unlabeled tails are offsets.

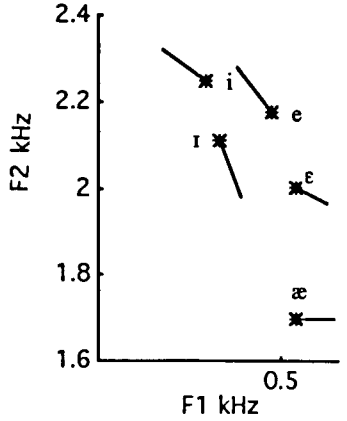


Figure 5 Plot of center of gravity of "goodness votes" for each vowel.

For each of the vowels, the following score was computed:

$$m = \sum_i x(i)g(i) / \sum_i g(i),$$

where *m* is the moment score (in Hz), *x(i)* is the frequency of the stimulus measure, and *g(i)* is the total goodness vote for the vowel in question for stimulus *i*. This total goodness vote is the sum of the total ratings (0-9 scale) across listeners and repetitions. (In order to focus on relatively good stimuli, for each vowel the summation was limited to those stimuli that received a total goodness rating at least 80% of the best rated stimulus for that vowel.)

DISCUSSION

One noticeable discrepancy is that the F1 onset logistic coefficient in Figure 3 for /æ/ is actually somewhat lower than that of /ɛ/, contrary to the trends of the production data in Figure 4. In addition, the vowel /i/ shows fairly large movement toward lower F1 and higher F1 offsets in the logistic coefficients than in the production means.

It should be noted that the onset F1 formant range explored in the synthetic study generally does not extend to those found for /i/ and /æ/ for males. However, these two vowels accounted for only about 7% and 6% of the total responses pooled over subjects and stimuli. Furthermore /æ/ does not show up as the modal response for -- and in fact does not ever receive more than 30% minority response to-- any of the stimuli (see Figure 2). It would seem unwise, therefore to put too much stock in the results for /æ/ without more data in a more appropriate frequency range.

Although /i/ does not account for many responses overall (7%), it does show up as the modal response for 7 of 120 stimuli (see Figure 2). It receives as high as an 85% response for one stimulus (that with lowest falling F1 and highest rising F2) and over 50% for four others. The /i/-pattern for Figure 3 (and 5) is qualitatively compatible with [ij]-like diphthongization often taken to characterize many dialects of North American English (including Canadian dialects [6]) by phonologists. Though we have not found any evidence for such patterns in production of citation forms of Western Canadian English [1, 2], it is not surprising that diverging F1-F2 patterns with low onset F1 and high onset F2 are taken as /i/ by our listeners. Nor is it surprising that the best-rated stimulus patterns for this vowel are those with diverging formants, since the nearest steady-state values are not as extreme as those of production values for male [i].

But a similar argument might be framed by a skeptic who accepts the traditional view that the North American tense vowels are phonetically diphthongal [ij] and [ej], but who believes that /ɪ/ and /ɛ/ are prototypically steady-state. This skeptic might reasonably argue that the categorization results are an artifact of the forced-choice experiment. More specifically, stimuli with converging F1-F2 trajectories are given lax vowel labels because they are more similar to steady-state lax-vowel prototypes than they are to diverging-formant (diphthongal) tense-vowel prototypes.

Even the pattern of results for the goodness data of Figure 5 is not immune to related criticism. However, for one of the lax vowels, /ɛ/ (which received 34% of total responses), there is sufficient relevant data to test the relation between goodness judgments and formant change directly. The stimuli selected for this test were those with flat or converging formants which were also identified by a subject as /ɛ/ six times out of six trials. For eight of the eleven subjects there were sufficiently large numbers of stimuli meeting these conditions that regression equations could be estimated for their responses individually. A (random-coefficients) multiple regression analysis shows a highly significant positive relationship [$t(7) = 4.81$ $p<.01$] between ΔF1 and the goodness rating for /ɛ/ controlling for all other factors. ΔF2 was not significant, though the effects in were in the expected direction ($t(7) = -0.6677$, $p>.5$). Thus there is positive evidence at least for the vowel /ɛ/ that a rising F1 pattern is preferred over a steady state, even when attention is limited to vowels that were consistently identified as /ɛ/. Additional experiments are being planned to explore goodness judgments more thoroughly in local regions of the stimulus space appropriate for each vowel.

ACKNOWLEDGMENT

This work was supported by a grant from SSHRC to the author.

REFERENCES

[1] Nearey, T., & Assmann, P. (1986). "Modeling the role of inherent spectral change in vowel identification." *JASA*, vol. 80, pp. 1297-1308.

[2] Andruski, J., & Nearey, T. (1992). "On the sufficiency of compound target specification of isolated vowels and vowels in /bVb/ syllables." *JASA*, vol. 91, pp. 390-410.

[3] Klatt, D. (1980). "Software for a cascade/parallel synthesizer." *JASA*, vol. 80, pp. 971-995.

[4] Nearey, T. M. (1990). "The segment as a unit of speech perception." *J. Phonetics*, vol. 18, pp. 347-373.

[5] Gumpertz, M., & Sastry, G. (1989). "A simple approach to inference in random coefficient models." *American Statistician*, vol. 43, pp. 203-210.

[6] O'Grady, W.,& Dobrovolsky, M. (1992). *Contemporary linguistic analysis*. Toronto: Copp Clark Pitman.

EFFECTS OF FORMANT FREQUENCY MODULATION ON VOWEL IDENTIFICATION

Peter J. Bailey, Kim Bevan and Tracy Burr
Department of Psychology, University of York, York YO1 5DD, U.K.

ABSTRACT

Two experiments are reported concerned with the perceptual effects of modulating the frequency of formants in synthesised vowels. The first experiment showed that modulated vowels tended to be identified in noise more accurately than unmodulated vowels, and that this tendency was more marked for hearing-impaired listeners than for those with normal hearing. The second experiment explored this finding with a different set of vowels and modulation conditions.

INTRODUCTION

Sinusoidal modulation of the centre frequency of a formant-like spectral peak in a periodic signal can lead to a significant reduction in the difference limen for peak frequency, compared to the difference limen for an unmodulated peak. This effect is particularly marked when the modulated spectral peak is partially masked by the presence in the stimulus of a second lower-frequency spectral peak. [1].

We have inferred from this increase in discriminability found for modulated peaks that when the energy that contributes to a spectrum envelope peak is modulated independently of other peaks and background noise, the perceptual salience of the peak is enhanced. We take this to be analogous to the disambiguating consequences of object movement for figure-ground segregation in visual scenes.

One motivation for measuring the effects of spectral peak modulation was an interest in signal-manipulation strategies that might increase the discriminability of speech in noise for listeners with a hearing impairment. Difficulties with speech perception in noisy or reverberant environments are commonly reported by people with sensori-neural hearing impairments, and are probably due in part to the impairments in spectro-temporal resolution that often accompany loss of auditory sensitivity. Amplification can improve recognition of speech in quiet but leads to relatively less benefit for speech in noise [2]. Enhancement of spectral contrast can lead to reliable increases in speech recognition accuracy in noise, but the improvements are modest [3].

It appeared meet, therefore, to explore the efficacy of modulation as an additional enhancement procedure that might be combined with others designed to augment the salience of informationally-rich parts of the speech signal. Here we report tests of the hypothesis that the accuracy of identification of vowels whose discrimination was dependent on resolution of spectral detail might be improved by formant frequency modulation.

EXPERIMENT 1

We have argued that peak frequency modulation increases discriminability by facilitating perceptual segregation of the modulated peak from other spectral peaks. Complete segregation could well turn out to be non-optimal for signals like speech, in which relationships between spectral prominences are important for segment identification. Rather than increasing speech identification accuracy, modulation of formant frequency could be counter-productive if its major effect is to disrupt the perceptual coherence of the formants. However, modulation may prove to be of overall benefit if it can be applied to a formant frequency so that the formant becomes perceptually more salient without there being a concomitant exclusion of the formant from the overall format pattern. In the first experiment we measured identification accuracy for a small set of synthesised vowels presented in noise to normal-hearing and hearing-impaired listeners. The independent variable was the amount of modulation applied to the frequency of the second formant, selected because of its key role in many phonetic constrasts.

Stimuli

Three-formant tokens of the four vowels /a/, /ɔ/, /u/ and /i/ were created using a parallel-formant synthesiser. The synthesis parameters are given in Table 1.

Table 1. Synthesis parameters for the vowels in Experiment 1. Formant frequencies (F) and 3 dB bandwidths (B) are given in Hz, formant amplitudes (A) are given in dB relative to the maximum.

	/a/	/ɔ/	/u/	/i/
F1	750	475	325	285
F2	1100	750	935	2375
F3	2600	2625	2325	3540
B1	65	75	70	80
B2	70	90	110	160
B3	80	140	150	60
A1	0	0	0	0
A2	-10	-13	-20	-20
A3	-15	-40	-45	-30

The vowels were 540 msecs in duration, with amplitude envelopes shaped at onset and offset by 60 msec linear ramps (except for /a/, which had a 90 msec offset ramp). The fundamental frequency was constant at 125 Hz.

Three versions of each vowel were synthesised: the second formant frequency was either unmodulated, or modulated at 5 Hz or at 10 Hz. When modulated the modulation depth was set to give a fixed frequency excursion of 150 Hz centred around the nominal formant frequency.

Subjects

Data were collected from four normal-hearing listeners and three listeners with mild to moderate unilateral hearing impairments.

Procedure

On each trial in the experiment subjects identified a single presentation of one of the vowels by circling one of four alternatives on a response sheet. In each modulation condition (unmodulated, 5 Hz modulation & 10 Hz modulation) subjects were presented with 20 practice trials, during which they heard each of the four vowels 5 times in random order, followed by 40 experimental trials involving a further 10 presentations of each vowel in random order. As far as possible the order of the three conditions was counterbalanced across subjects.

All vowels were presented monaurally via headphones at 70 dB SPL in a continuous broad-band background noise set individually for each ear at a level corresponding to a 10 dB sensation level for the vowels. Noise levels were in the range 71-79 dB SPL for normal-hearing subjects and 60-73 dB SPL for hearing-impaired subjects. Stimuli were presented to the right ear of the normal-hearing subjects. Hearing-impaired subjects ran through the procedure twice, first using their "good" ear, and again in the same session (with a different order of conditions) using their impaired ear. All listening tests were carried out in a sound-attenuating room.

Results

The overall numbers of errors in the three modulation conditions are shown for normal-hearing subjects in Table 2, and for hearing-impaired subjects in Table 3.

Table 2. Normal-hearing subjects: number of errors out of 40 presentations per vowel per condition.

RIGHT EAR	/a/	/ɔ/	/u/	/i/	total
unmodulated	10	17	18	23	68
5 Hz modulation	7	15	14	15	51
10 Hz modulation	8	11	15	15	49

Table 3. Hearing-impaired subjects: number of errors out of 40 presentations per vowel per condition per ear.

"GOOD" EAR	/a/	/ɔ/	/u/	/i/	total
unmodulated	5	6	10	13	34
5 Hz modulation	1	4	6	10	21
10 Hz modulation	1	2	5	11	19

IMPAIRED EAR	/a/	/ɔ/	/u/	/i/	total
unmodulated	7	6	13	11	37
5 Hz modulation	0	0	6	6	12
10 Hz modulation	0	0	4	9	13

Discussion

Taken together the data suggest that: (i) modulation of the frequency of the second formant tends to decrease the error rate below that found for unmodulated vowels, (ii) the effects of modulation, when present, are not strongly dependent on modulation rate, and (iii) to the extent that the perceptual coherence of vowel formants is indexed by identification errors, there is no indication in the data that modulating a

single formant causes it to be excluded from the vowel percept. The trends seen in the means were evident in the data of all subjects, except for one of the hearing-impaired listeners whose mean identification accuracy was apparently not affected by modulation.

A comparison of the data from good and impaired ears of the hearing-impaired subjects suggests that the tendency for modulation to reduce identification errors may be more marked for the impaired ear than for the good ear. The finding that the error rate for hearing-impaired listeners was lower overall than that for normal-hearing listeners is attributable to the technique used for setting the level of the background noise. Noise levels at masked thresholds were lower for hearing-impaired listeners, with the consequence that in the main part of the experiment the spectrum level of the noise was lower relative to those of the spectral peaks in the vowels than was true for the normal-hearing listeners.

It appears that, at least for these vowel stimuli at the unfavourable signal-to-noise ratios we have used and for these subjects, second formant frequency modulation tends to improve vowel identification performance.

EXPERIMENT 2

In an attempt to establish the reliability of the trends demonstrated in the first experiment we carried out a second experiment, using a different and larger set of subjects, vowels and modulation conditions.

Stimuli

Three-formant tokens of the six vowels /ɪ/, /ɑ/, /ɔ/, /ɪ/, /u/ and /ɛ/ were created using a parallel-formant synthesiser. The formant frequencies are given in Table 4.

Table 4. Formant frequencies (Hz) for the vowels in Experiment 2.

	F1	F2	F3
/i/	280	2250	2890
/ɑ/	710	1100	2540
/ɔ/	590	840	2540
/ɪ/	325	1920	2560
/u/	310	870	2250
/ɛ/	550	1770	2490

Formant bandwidths for the first, second and third formants were constant at 50 Hz, 70 Hz and 100 Hz respectively for all vowels, and formant amplitude parameters were set equal for all formants in all vowels. Vowels were 250 msec in duration, with 5 msec linear onset ramps and 25 msec offset ramps. Pitch fell linearly from 130 Hz to 110 Hz through the vowels. The overall levels of the vowels varied by less than 2 dB.

Eight versions of each vowel were synthesised: one was unmodulated, three had one formant modulated (F1, F2 or F3), three had two formants modulated (F1&F2, F2&F3 or F1&F3), and one had all three formants modulated. When modulation was present the modulation rate was 10 Hz and the peak-to-trough modulation depth was 10% of the nominal formant frequency. When more than one formant was modulated all modulations were in phase.

Subjects

Data were collected from 20 normal-hearing listeners.

Procedure

The task for the subjects was similar to that in Experiment 1, except that there were six response alternatives. Each experimental session involved 12 practice trials (two presentations of each unmodulated vowel), followed by 5 blocks of 48 experimental trials. In each block each vowel was presented in each modulation condition once, in random order. The stimuli were presented from tape binaurally over headphones in a continuous broad-band noise set to a sound pressure level 8 dB greater than the peak level for the least intense vowel (/u/). Overall presentation level was set to be comfortable for the subjects (approximately 75 dB SPL). Listening tests were carried out in a quiet (but not sound-attenuating) room.

Results and Discussion.

Overall errors are shown in Table 5 for all vowels and modulation conditions. The data show large differences in the mean errors across vowels, but little systematic trend in the mean errors for the modulation conditions. These observation were confirmed with a two-way within-subjects analysis of variance on the error data which showed a significant main effect of vowel

Table 5. Overall percentages of identification errors for Experiment 2.

conditions	vowel						
(formants modulated)	/i/	/ɑ/	/ɔ/	/ɪ/	/u/	/ɛ/	MEAN
none	18	2	20	56	38	5	23.1
F1	24	1	25	57	42	6	25.8
F2	29	1	19	60	35	5	24.8
F3	26	3	21	49	47	7	25.5
F1&F2	25	3	19	55	40	7	24.8
F2&F3	22	2	18	55	49	6	25.3
F1&F3	20	3	25	50	40	5	23.8
F1&F2&F3	27	0	20	53	48	3	25.2
MEAN	23.9	1.9	20.9	54.4	42.4	5.5	24.8

[$F_{(5,95)}$=20.64, $p<0.01$], but no main effect of modulation condition [$F_{(7,133)}$=0.53] and no interaction between vowel and modulation condition [$F_{(35,665)}$=1.18]. Although formant frequency modulation had no systematic effect on the accuracy of vowel identification, it is noteworthy that for three of the four vowels for which substantial numbers of errors were made, the unmodulated condition did not show the lowest mean error rate. Analyses of the confusions between vowels revealed a strong tendency for /ɪ/ to be heard as /ɛ/ or /i/, and for /u/ to be heard as /ɔ/.

There is one aspect of the modulation conditions involving simultaneous modulation of more than one formant that deserves comment. Given that all modulations were at the same rate and in the same phase, the perceptual coherence of formants accruing from common modulation may have acted in opposition to any perceptual salience that modulation may have conferred on individual modulated formants.

There were several differences between the stimuli and procedures in the two experiments that might have contributed to the differences in outcome. A likely candidate was the differences in spectrum level of the formants relative to the spectrum level of the background noise. This was particularly true for F2 and F3, whose spectrum levels for all vowels were below the spectrum level of the noise in Experiment 1, but above it in Experiment 2. We draw the tentative conclusion, to be examined in further experiments, that effects of formant frequency modulation may only be manifest when the signal-to-noise ratio is particularly unfavourable.

CONCLUSION

The limited set of data reported here suggest that in some, but not all, circumstances, modulation of formant frequency may be able to improve the discriminability of speech contrasts that are dependent on resolution of spectral detail. If the precise conditions under which such improvements are found reliably can be established, formant frequency modulation would seem to have potential as an additional item in the armoury of the hearing-aid designer.

ACKNOWLEDGEMENTS

We are grateful to Mr G. Hope and Mr. A. Grace for their help in locating suitable hearing-impaired subjects, and to the subjects themselves for their cooperation. The UK Science and Medical Research Councils provided financial support.

REFERENCES

[1] Bailey P.J. & Bevan K.M. (1991), "Frequency modulation of formant-like spectral peaks", Proc. XIIth Int. Congr. of Phon. Sci. (Aix-en-Provence), vol. 4, 46-49

[2] Plomp R. (1978) "Auditory handicap of hearing impairment and the limited benefit of hearing aids.", J. Acoust. Soc. Amer., vol. 63, 533-549

[3] Baer T., Moore B. C. J. and Gatehouse S. (1993) "Spectral contrast enhancement of speech in noise for listeners with sensorineural hearing impairment: effects on intelligibility, quality and response times.", J. Rehabil. Res. and Devel., vol. 30, 49-72

MORAIC STATUS AND SYLLABLE STRUCTURE IN SPEECH PERCEPTION

Takashi Otake and Kiyoko Yoneyama***
*Faculty of Foreign studies, Dokkyo University**
*Graduate School of Dokkyo University***
1-1, Gakuen-cho, Soka-shi, Saitama-ken, 340, Japan

ABSTRACT

This paper investigates the role of syllable structure and allophonic variations of a nasal in Japanese speech perception. Two experiments are designed regarding recognition of the allophonic variations of a nasal in reference to syllable position. In the first experiment stimuli were designed in such a way that a moraic nasal preceding three stop consonants, /p, t, k/ was spliced and embedded in an onset of the first syllable of a word, /natsu/. In the second experiment stimuli were designed in such a way that a nasal from the five contexts was spliced and embedded in the coda position of three CVN words.

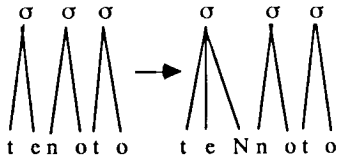
The results of these experiments suggest that moraic status is determined by position in a syllable.

1. INTRODUCTION

A number of studies in speech production have revealed that the mora is the basic production unit in Japanese. Recent studies in speech perception have also demonstrated that the basic segmentation unit in Japanese is a mora [1][2]. The mora *per se* is usually defined by its unique durational characteristics, namely a constant duration, although the existence of an absolute durational unit is not necessarily agreed upon in the literature. Our recent investigations, however, have suggested that recognition of moraic consonants in speech perception is influenced by syllable structure as proposed by phonological analysis [3], as well as by duration [4] [5].

In the earlier works [4][5], we altered the duration of a nasal which was embedded both in a coda and an onset in a syllable. Unlike the findings in the study of production [6], even one tenth of the original nasal duration was perfectly recognized as a moraic nasal as long as it was embedded in a coda position. On the other hand, when the nasal duration in an onset was increased to as long as 1.5 times the original one, a resyllabification process was observed (See (1)).

(1)



Thus, the added portion of the nasal was recognized as a moraic nasal in the preceding syllable. The results of these two experiments led us to conclude that determination of moraic status in speech perception may be highly related to position as well as to duration. These studies have led us to posit another interesting question concerning the relationship between moraic status and syllable structure. This paper reports the results of two experiments investigating the relationship between the allophonic variations of a Japanese nasal and syllable position.

2. EXPERIMENT I

In Experiment I, three variants of a nasal in coda position of a CVN syllable preceding three consonants (p, t and k) were spliced from three words (teNpo, teNto and teNko) and embedded in the onset of words (natsu, nasu, naku, nata). These stimuli were presented to Japanese listeners who were asked to transcribe them.

2.1 Method

2.1.1 Materials

Twelve stimulus words were made from two groups of words, each of which contains a nasal in an onset or a coda. These are /natsu, nasu, naku, nata/ and /teNpo, teNto, teNko/. All these words were recorded at a normal tempo by a male native speaker of Tokyo Japanese. In order to avoid the effect of pitch accent, no accent was assigned. After recording, three stimuli for each word which has a nasal in an onset (natsu, nasu, naku, nata) were made by Kay Sona Graph 5500 in such a way that the nasal in onset position was cross-spliced with three allophonic nasals ([m], [n], [ŋ]), each of which occurs in the three nonsense words (teNpo, teNto, teNko), respectively. The duration of the spliced nasals in the stimulus words was the same as the original duration of the nasal in the original words (natsu, nasu, naku, nata). The portion of the vowel which exhibited nasalization was removed. Each of the twelve spliced words and the original four words (natsu, nasu, naku, nata) were recorded twice onto a tape with a two second inter-stimulus interval in random order.

2.1.2 Subjects

Subjects were 30 students of Dokkyo University. The majority of the students were majoring in English who have basic knowledge of phonetics and phonology.

2.1.3 Procedure

Each subject was instructed to listen to the stimuli, which were repeated twice, and to write down what they heard in Roman alphabet on a test sheet. They were instructed to write moraic nasals with a capital N. The stimuli were presented to each subject individually over headphone in a quiet room.

2.2 Results

The results of the experiment are shown in Fig. 1. As can been seen, the nasal spliced from the coda preceding /t/ which was embedded in the onset of nV-CV words was recognized as an alveolar nasal ($\chi^2[1] = 120$, $p<.001$). The nasal spliced from the coda preceding /p/ embedded in the onset was recognized as a bilabial nasal ($\chi^2[1] = 90.1$, $p<.001$). However, when the nasal spliced from the coda preceding /k/ was embedded in the onset, it was simply recognized as an alveolar nasal ($\chi^2[1] = 104.5$, $p<.001$).

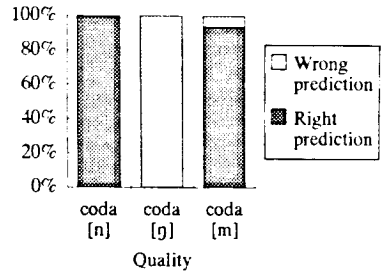


Figure 1. Identification functions for spliced nasal consonants in an onset position that differ in quality.

2.3 Discussion

Studies of Japanese phonology have claimed that as long as a nasal is placed in a coda position, it is automatically defined as moraic [3]. This may imply that moraic status is given to any nasal which is embedded in the coda position even in speech perception. Since the results of the present experiment have shown that the three different nasal qualities embedded in the onset were recognized either as an alveolar or a bilabial nasal, while if they were placed in the coda they were recognized as a moraic nasal, this phenomenon can be understood only if we assume that a nasal's moraic status is determined by its coda position. The reason why a nasal preceding a velar consonant embedded in the onset was reported as an alveolar nasal was simply because velar nasals do not occur in onset position in Japanese and there is no way to represent it in Roman alphabet.

3. EXPERIMENT II

In Experiment II, we investigated how a nasal in the coda in a CVN syllable preceding three different contexts (p, t and k) and a nasal in the onset (n and m) in the coda were recognized in order to confirm the findings of Experiment I. If coda

position is a decisive factor for a moraic status, all the nasals in the present experiment must be recognized as moraic.

3.1 Method

3.1.1 Materials

Fifteen stimulus words were made from two groups of words, each of which contains a nasal in a coda or an onset. These are /teNpo, teNto, teNko/ and /natsu, matsu/. All these words were recorded at a normal tempo by a male native speaker of Tokyo Japanese. In order to avoid the effect of pitch accent, no accent was assigned. After recording, five stimuli for each word which has a nasal in a coda (teNpo, teNto, teNko) were spliced with a Kay Sona Graph 5500 in such a way that five allophonic nasals differently taken either from a coda (teNpo, teNto, teNko) or from an onset (natsu, matsu) were inserted into /teNpo, teNto, teNko/, respectively. The portion of the vowel which exhibited nasalization was removed. The duration of nasals from a coda (teNpo, teNto, teNko) was about 60 ms., and that of nasals from an onset (natsu, matsu) was not altered. Each stimulus word was recorded twice onto a tape with a two second inter-stimulus interval, in random order.

3.1.2 Subjects

Subjects were the same as in Experiment I.

3.1.3 Procedure

The procedure was the same as in Experiment I.

3.2 Results

The results in the three spliced words /teNpo, teNto, teNko/ are shown in Fig. 2, Fig. 3, Fig 4., respectively. As can been seen in the figures, nasals in coda position in the three spliced words were reported as moraic ($\chi^2[1] = 208.3$, $p<.001$ for teNpo; $\chi^2[1] = 284.2$, $p<.001$ for teNto; $\chi^2[1] = 276.5$, $p<.001$ for teNko). Each of five variations of a nasal was significantly more often judged as moraic ($\chi^2[1] = 179.1$, $p<.001$ for coda [m]; $\chi^2[1] = 160.6$, $p<.001$ for coda [n]; $\chi^2[1] = 156.8$, $p<.001$ for coda [ŋ]; $\chi^2[1] = 138.7$, $p<.001$ for onset [m] and [n]. There was no significant difference between the five nasal variation conditions.

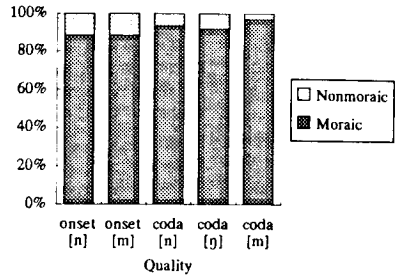


Figure 2. Identification functions for spliced nasal consonants in coda position that differ in quality in "teNpo".

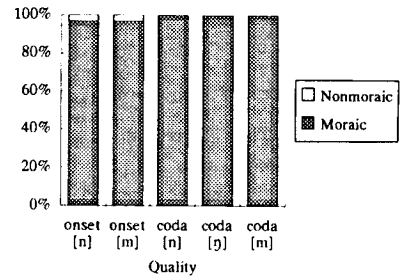


Figure 3. Identification functions for spliced nasal consonants in coda position that differ in quality in "teNto".

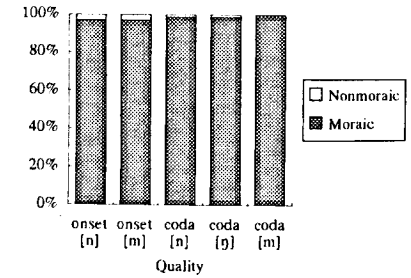


Figure 4. Identification functions for spliced nasal consonants in coda position that differ in quality in "teNko".

3.3 Discussion

The results in the present experiment have further confirmed the findings of Experiment I. Even though the nasals in the five different contexts must have different phonetic qualities, all were judged as a moraic nasal in coda position. It is obvious that syllable position plays an important role for recognition of the mora in Japanese.

4. GENERAL DISCUSSION

The two experiments in this study have clearly shown that recognition of a moraic nasal in speech perception in Japanese is dependent on syllable position. The results of Experiment I have shown that the three allophonic variations of a nasal, [m], [n] and [ŋ] were recognized as different in syllable initial position, while they were simply recognized as a moraic nasal if they were placed in the coda position.

The results of Experiment II have also confirmed that when the two nasals [m] and [n] in the onset and the three nasals in the coda in the CVN syllable were embedded in the coda of the initial syllable in /teNpo, teNto and teNko/, they were judged as an undifferentiated moraic nasal.

In previous studies in the literature of speech production and perception, moraic status has been investigated with respect to duration. On the other hand, in phonological analysis a mora is defined as a subunit of a syllable. The relationship between these two units has been intensively investigated by phonologists. If a mora is defined as a subunit of a syllable, it may be very likely that both mora and syllable structure play a role in speech perception. Our earlier studies have demonstrated that a speech segmentation procedure during on-line speech perception in Japanese must be based upon morae rather than on syllables [1][2]. However, a speech segmentation with an off-line speech perception task may be more complicated. Although we need further intensive examinations to clarify the problem, knowledge of syllable structure must be involved in speech perception.

5. CONCLUSION

In this paper, we have conducted two experiments in order to examine the relationship between the allophonic variations of a nasal and moraic status. We have claimed that moraic status is highly dependent upon syllable structure. In other words, moraic status is determined by a coda position. The results of this study have confirmed the findings in our previous studies, claiming that syllable structure plays an important role in speech perception in Japanese.

ACKNOWLEDGMENT

This study was partly supported by a grant of the Ministry of Education (Grant number is 06610475). We would like to express our gratitude to A. Cutler, who has read the earlier manuscript and gave us many invaluable comments, and to G. Hatano and all other people participating in the experiments. We further thank K. Kurisu and M. Komada for their critical comments.

REFERENCES

[1] Otake, T., Hatano, G., Cutler, A., and Mehler, J. (1993), Mora or syllable? Speech segmentation in Japanese, *Journal of Memory and Language*, 32, pp. 258-278.

[2] Cutler, A. and Otake, T. (1994), Mora or phoneme? Further evidence for language specific listening, *Journal of Memory and Language*, 33, pp. 824-844.

[3] Ito, J. (1988), *Syllable Theory in Prosodic Phonology,* Garland Publishing Inc.

[4] Otake, T. and Yoneyama, K. (1994a), A moraic nasal and a syllable structure in Japanese, *Proceedings of ICSLP 94*, Yokohama, 3, pp.1427-1430.

[5] Otake, T. and Yoneyama, K. (1994b), A geminate consonant and a syllable structure in Japanese, *Dokkyo Studies in Data Processing and Computer Science*, 12, pp.55-64, Dokkyo University.

[6] Sato, Y. (1993), The durations of syllable-final nasals and the mora hypothesis in Japanese. *Phonetica*, 50, pp.44-67.

ACQUISITION OF LISTING INTONATION IN ENGLISH CHILDREN

Louise H. Coward
Department of Linguistics, University of Manchester, UK

ABSTRACT

This paper examines the intonation of lists (enumerations) in two British children at ages 2;0.3 and 2;4.27. The patterns found are compared to standard reports in the adult intonation literature for British English. Theoretical proposals are then advanced as to how such patterns are constructed and employed in both child and adult language.

INTRODUCTION

This research forms part of a doctoral study examining functions of intonation in the speech of in five English children between 1;10.3 and 2;8.14. Data were recorded naturalistically in the children's homes with at least one parent and one researcher present (recordings and orthographic transcriptions were kindly supplied by Dr. E. Lieven). Intonation was transcribed auditorily by the present author by means of an interlinear tonetic notation (instrumental analysis was attempted but proved impractical due to recording conditions).

This paper specifically considers the intonation of counting sequences (a type of listing activity) produced by two of the subjects of the larger study. The author is aware of only one published mention (in passing) of listing intonation in the acquisition literature [1] but there are a number of references to lists or enumerations, albeit usually brief, in standard accounts of adult British English, and it will be helpful to consider some of these first.

ADULT LISTING PATTERNS

It is common in the literature to make a distinction between *incomplete* and *complete* lists, in this paper hereafter referred to as *open vs. closed* lists. Crystal describes this as "the distinction between a limited or an unspecified number of alternatives" [2]: where there is a fixed number of items, he gives the pattern as rising nuclei on prefinal items with a falling nucleus on the final item as in his following example (here, relevant pitch accents are shown schematically immediately prior to the accented syllable):

(1) Would you like gin or whisky or tea?

This is in contrast to a realisation with a final rise, where the number of possible items in the list is not limited to the three actually given, thus:

(2) Would you like gin or whisky or tea?

Both patterns are commonly given in the literature, but a variety of other patterns are also reported. (3) below represents a summary of the patterns described by Crystal [2], O'Connor & Arnold [3], Halliday [4], Kingdon [5], Couper-Kuhlen [6] and Schubiger [7].

(3)

(a) Closed Listing Patterns (Adults)

(i) rise+rise+...rise$_n$+fall

(ii) rise+rise+...rise$_n$+fall-rise

(iii) rise+rise+...rise$_n$+rise-fall-rise

(iv) fall-rise+fall-rise+..fall-rise$_n$+rise-fall

(v) rise-fall-rise+r-f-r+...r-f-r$_n$+rise-fall

(vi) level+level+...level$_n$+fall

(vii) level+level+...level$_n$+fall-rise

(viii) fall+fall+...fall$_n$+fall

(ix) fall+...fall$_n$+rise+fall

(b) Open Listing Patterns (Adults)

(i) rise+rise+...rise$_n$+rise

(ii) fall+fall+...fall$_n$+fall

Working on the assumption that these patterns give an accurate account of adult British English listing intonation, we can make the following generalisations. A *closed listing pattern* must finish on a fall, or in certain cases on a complex rise (fall-rise or rise-fall-rise); it must not finish on a simple rise. Prefinal tones are generally of a uniform type and will typically be rises (whether simple or complex) or levels (which we will treat as a subtype of rise, following Cruttenden [8]); exceptions are where the entire list is realised on a string of simple falls, or where only the penultimate item bears a rise, flanked by simple falls on all other items. An *open listing pattern* will have uniform tones throughout (final and prefinal) and will typically involve either simple rises or simple falls, the simple fall sequence thus being ambiguous between the two categories.

It should be observed, however, that the remainder of this paper will be primarily concerned with *counting sequences*, and whilst it seems highly reasonable to treat these as a form of listing activity (*cf.* [3] pp.58-9), intuitively the author feels that Kingdon's patterns ending in a fall-rise or rise-fall-rise (*i.e.* (3a ii, iii and vii)) would only be expected to be appropriate where some sort of *correction* of a previous miscount is involved. Leaving these three patterns aside produces a neater generalisation about the remaining closed listing patterns: they all end in a variety of fall (for the grouping of nuclear contours into falls and rises according to final pitch movement, see *inter alia* [8]).

The schematic overview in (3) above masks the fact that these adult listing patterns can be considered in two ways: as *compositional* sequences of concatenated nuclei; or as *holistic* contours. Both views are implicit in the literature, although rarely elaborated. O'Connor & Arnold [3] and Crystal [2] imply the compositional view by inserting tone unit boundaries after each item in the enumeration. Halliday [4], Kingdon [5], Schubiger [7] and (following Schubiger) Couper-Kuhlen [6] suggest the holistic view for some of the closed listing patterns. Halliday does this notationally by treating the prefinal pitch accents as constituting a listing pretonic segment, with the whole pattern as a single tone unit; likewise Kingdon [5] states that enumerations form "single sense groups" (all his examples are *closed*). Schubiger [7] and Couper-Kuhlen [6] are most explicit of all, arguing that the patterns in (3a i) and (3a ix) are planned as a unit from the start, whereas the pattern in (3a viii) is planned item by item, and "each might be the last," [7]. Virtually nothing is said in the literature, however, about the planning of *open* listing patterns, with the exception of the use of compositional notation in Crystal [2] and O'Connor & Arnold [3]. Nonetheless the use of rising tones to signal *continuity* is frequently reported in adult intonation literature (*e.g.* [3], [8], [11]), and this would seem to suggest that rises within both open and closed lists are introduced compositionally and not as part of an overall holistic pattern. We shall return to this holistic *vs.* compositional hypothesis and to the question of continuity later in the paper.

CHILD LISTING PATTERNS

We now turn to an examination of some of the counting sequences occurring in the child language data for this study. Examples (4), (5) and (6) below were produced by child AT at age 2:0.3 (in all excerpts shown here, M designates the child's mother and C the child). Here we see three counting sequences produced by child AT; two (those in (4) and (6)) are elicited by a prompt from AT's mother; that in (5) however is a spontaneous production. At first glance, all three patterns look very adult-like, and on the basis of the adult patterns given in (3a & b) we can hypothesise that (4) contains a closed list (pattern (3a i)) whilst (5) and (6) contain open lists (pattern (3b i)). In view of this hypothesis, let us now consider these examples more closely. In (4), we know that there are four jigsaws to be counted. M prompts C to start counting, and C counts to five. Whilst the number is inaccurate, the decision to realise the sequence on a series of prefinal rises plus a final fall is appropriate, since the situation involves a specific limited set of items. It is clear that this is not a case of imitation, despite M's

prompt, since at the lexical level different numbers are inserted by C and at the prosodic level it is her own decision to employ the final fall. What we do not know here, unfortunately (due to insufficient contextual information), is whether C is associating each number with a visible referent or whether she is uttering the counting sequence in abstract. More will be said on this point later.

(4)
M: How many jigsaws did you do?
One —
C: two three four five
M: No, you only did four, didn't you?

(5)
M: Are you going to put all these bits in? You've got to count them. We need twelve, don't we?
C: seven eight nine ten
M: Eleven. ... Let's have a look: two — four, six ...

(6)
M: We need four of these bits — so one —
C: two three four
M: ... We'll do that one again before we put it away, shall we?

In (5), we know that there are twelve items to be counted, and since C stops her sequence at ten, the final rise may be an appropriate indication that the list is unfinished (sadly, again, we do not know whether each number is being associated with a referent here). However, if we consider the similar case in (6), we find that C is in fact not making the open-closed distinction consistently: in this case we know that there are four items to count, and whilst C counts to four, the fourth and final item receives a rise, which makes this an inappropriate use of an open listing pattern on a closed list. So we can conclude here that child AT has mastered the intonational *form* of counting sequences but not yet their *function*.

We can now look briefly at data from one other child. Examples (7) and (8) are from child AH at age 2;4.27. Both these examples are unelicited (and probably uttered referent-free):

(7)
C: one two three four five

(8)
C: one two three four five

(7) seems to be an archetypal closed listing pattern as in (3a i), and we might wish to conclude from this that child AH has mastered at least the form of this closed listing pattern. However, the pattern in (8) deviates to some extent from the adult forms; it introduces high shallow falls in penultimate and antepenultimate positions. This may nonetheless not be too far from adult realisations: it is sometimes claimed that overall high pitch and suspension above the baseline are functionally equivalent to an actual rising movement (*e.g.* [9]). There may also be related regional dialectal considerations in the case of this particular child (Glaswegian and Mancunian influences; *cf.* [10]). The final fall in these examples is however not subject to variation, so this child is probably well on the way to mastering the formal characteristics of the closed list.

THEORETICAL DISCUSSION

The majority of the listing patterns considered in this paper and in the literature involve rising tones in prefinal position. It is tempting to view the use of rises in listing sequences as fulfilling a continuity function, whether in the case of prefinal rises which are then indeed followed by further material, or in the case of final rises in open lists, where there is the *potential* for further material to follow. Indeed, this interpretation of adult listing intonation is implied by O'Connor & Arnold [3], and suggests a compositional view of listing patterns.

However, in view of the acquisitional data, the issue seems more complex than this. We have already made a distinction between acquisition of the *form* of listing patterns and acquisition of their *function* (*i.e.* appropriate choice of variant). In the data we have considered, the children do not seem to be constructing a series of continuity rises one by one and indicating completion with a final fall; rather, they seem to have a holistic pattern, or *template*, in mind for the sequence as a whole, for

example: rise+rise+...rise$_n$+fall, into which varying lexical material can be inserted in varying amounts, but which must retain certain intonational characteristics across all instantiations.

The question then arises: what are adults doing when they produce listing sequences? Are they constructing these unit by unit on the basis of continuity rises? It seems perfectly possible that, adults too have a 'stretchable' template in mind in advance for the entire sequence, and need only really consider the question of continuity with regard to the final item; this decision itself might be made in advance in selecting a closed rather than an open template before the sequence is started, or it might be made towards the end of the production of the sequence. In the latter case, either two templates, an open one and its closed equivalent, would have to be available in parallel; or else the first part of the pattern only would be holistically planned, and the last one or two items added semi-compositionally. This template hypothesis does not deny that the use of rises in listing patterns *originated* (historically) in a genuine continuity function, but such patterns seem so ingrained as speech habits that it seems more likely that they are stored and activated as prefabricated templates than that they are built, as it were, brick by brick each time, even where they *are* being used functionally. It is also important to observe that even adults treat counting sequences in two ways: as a series of labels applied to successive related referents (a *functional* use), and as an abstract prememorised recitation sequence (a predominantly *formal* use — as in response to: "Close your eyes and count to ten"). Although it is the abstract recitation form which a child acquires first, the fact of this dual adult usage means that *unlike* in many other cases of early item-learning [12], the child must still retain and employ the unanalysed recitation sequence even *after* he/she has learnt the functional use of its analysed components. It would thus seem more reasonable that as an adult and even in functional contexts a speaker bases the form of his/her counting sequences on the analogy with the prememorised sequence rather than constructing it from scratch on each occasion.

ACKNOWLEDGEMENTS

The author expresses gratitude to the Humanities Research Board of the British Academy who funded this research and this paper (award BA92/1291), to Dr. E. Lieven for supplying data, and to Dr. A. Cruttenden for general supervision and specific advice.

REFERENCES

[1] Carlson, P. & M. Anisfeld (1969), "Some observations on the linguistic competence of a two-year-old child," *Child Development*, vol. 40, pp. 569-75.
[2] Crystal, D. (1969), *Prosodic systems and intonation in English*, Cambridge: CUP.
[3] O'Connor, J.D. & G.F. Arnold (1973), *Intonation of colloquial English*, London: Longman.
[4] Halliday, M.A.K. (1967), *Intonation and grammar in British English*, The Hague: Mouton.
[5] Kingdon, R. (1958), *The groundwork of English intonation*., London: Longman.
[6] Couper-Kuhlen, E. (1986), *An introduction to English prosody*, London: Arnold.
[7] Schubiger, M. (1958), *English intonation: its form and function*, Tübingen: Niemeyer.
[8] Cruttenden, A. (1981), "Falls and rises: meanings and universals," *Journal of Linguistics*, vol. 17, pp. 77-91.
[9] Hirst, D. (1993), Lecture given at the ELSNET Summer School on Prosody held at University College, London, 12–23 July 1993.
[10] Cruttenden, A. (1994), "Rises in English," in J.W. Lewis (ed.), *Studies in general and English phonetics: Essays in honour of Prof. J.D. O'Connor*, London: Routledge, pp.155-173.
[11] Crystal, D. (1975), *The English tone of voice: Essays in intonation, prosody and paralanguage*, London: Arnold.
[12] Cruttenden, A. (1981): "Item-learning and system-learning," *Journal of Psycholinguistic Research*, vol. 10, pp. 79-88.

ON VARIATION IN EARLY SPEECH PRODUCTION

B. Krüger
English Department, Kiel University, Germany

ABSTRACT

The Kiel Project on Early Phonological Development focuses on data from German children during their initial stages of phonological development. Based on this new evidence it will be shown that not all types of phonological variation and the children`s deviations from target words can be accounted for in terms of production difficulties. Certain kinds of variation require either perceptual explanations or have to be discussed with regard to the maturation of mental processing abilities [1].

DATA COLLECTION

Eight children, growing up in L1 monolingual German-speaking environments, were visited at home at weekly intervals. The investigation began when the children produced their first words. The children`s utterances were recorded with a wireless beyer dynamic microphone on Uher 4000 IC tape recorders connected with a bd NE 42 receiver.

The goal was to document the children`s early phonological and lexical development by eliciting as many tokens per lexical item as possible in each session.

The children`s utterances were transcribed using the IPA symbols (1993). In this paper the data from four subjects (Isabell, Julian, Simon and Till, see Table 1.) provide the basis for a typology of phonological variation and phonetic deviations from source words. For each child there exist lists of all reproductions (tokens) per word narrowly transcribed plus remarks regarding session number, age, situation and the child`s mood.

Table 1.

Child	Sessions	words	Age
Isabell	40	351	1;00,30-2;00,15
Julian	32	128	1;00,11-1;09,22
Simon	54	211	0;10,07-1;10,10
Till	22	79	0;11,24-1;05,25

RESULTS

Except for one group of replica which displayed target-like phonological forms from the beginning, the children`s reproductions can be divided into three types:

(a) words whose phonetic structure is often stable but not target-like;

(b) words whose phonetic structure is similar to the target, the lack of accurate articulation, however, leading to particular types of minor deviations from the target.

(c) words whose reproductions varied greatly, especially in early periods of development and words that deviate from an established routine (see *(a) leveling of gestures)* in later developmental stages.

The data analysis and interpretation is based on the assumption that the child`s mental word representations reflect the input, i.e. the representations are assumed to be "fuzzy". The storing of particular sound chains over time is based on an extremely variable input and must be linked to information about situations and moods. This means that the child`s representations cannot be as accurate as his/her ability to distinguish between phonemes in a test paradigm (e.g. Categorical Perception investigation,[2]), might suggest.

The time span up to when word representations become accessible for reproduction differs greatly from child to child as pointed out by many researchers [3]. Our data suggest that the frequency of access to mental representations has an effect not only on the automization of nerve impulses to certain muscles in the vocal tract but also on the ability to distinguish between entities of mental representation.

Articulatory difficulties and articulatory simplifications are explained in terms of Gestural Phonology [4]. Phoneme-sized entities of aquisition are not postulated, although our data indicate that they may exist in some early word production. But the ability to consistently reproduce sound features in the same position in a word develops differently from child to child and from word to word. Therefore the size of phonological entities may differ at a given stage of aquisition and cannot at all be assumed consistent.

(a) Leveling and Elision of Consonant Constriction Gestures per Word

Similarity between gestures or a low number of gestures per word specify types of variation that display phonologically simplified tokens. The resulting output is either stable in at least three tokens produced in a row, all tokens in one session or over a longer period of time or they reflect a unique phonetic form in an unstable stage of word development as described in the last part of this paper.

Words, for instance, that require dorsal gestures and labial gestures were reproduced *either by dorsal gestures or by labial gestures:*

SIM *Frauke (first name)* [ˈkn̥ŋkə]
ISA *Schiff (ship)* [ʃɪɛ]
JUL *Koffer (case)* [ˈpɔpɐ]
komm (come) [pɔm]

Words that require labial gestures and tip of the tongue gestures were reproduced *either by tip of the tongue gestures or by labial gestures:*

JUL *baden (to take a bath)* [ˈdạdn̥]
TIL *bitte (here you are)* [ˈdzɪzɪ]
JUL *baden (to take a bath)* [ˈb̥ạpm̥]

Words that require dorsal gestures, labial gestures and tip of the tongue gestures were reproduced *by dorsal gestures and tip of the tongue gestures:*

ISA Löffel (spoon) [ˈdɔkəl]
[ˈçœkəl]

If the representation of a particular source word is activated frequently and generates phonologically stable reproductions over several weeks, a kind of allocation automatism has apparently developed. This is particularly true in cases where children rely extensively on a small set of articulatory patterns [5] to render different words.

The data show individual preferences for either certain consonantal constrictions or consonant-vowel-combinations that shaped the idiosyncratic phonological development of each child.

Isabell for instance reproduced words with labiodental and dorsal fricatives primarily with dorsal constrictions.

Table 2. ISA`s preference for dorsal gestures instead of labial gestures.

Word	tokens	age	%
Schlafsack	07	1;10,07-1;11,17	71
Schleife	10	1;10,07-1;11,17	70
Fische	09	1;10,27-1;11,10	56
Käfer	26	1;07,22-2;00,15	39

Examples:

Schlafsack (sl.bag) [ˈʂlɑ̝ɻ̊ʂʌk]
Schleife (bow) [ˈʂlaɛ̆ɻ̊ə]
Fische (fishes) [ˈʂɪɻ̊ə]
Käfer (bug) [ˈkʰɪɻ̊ɛ]

Another sort of output simplification is the elimination of certain gestures or the reduction of complex sequences of motion to single components. The position of the eliminated gesture in the target word as well as the number and complexity of different gestures determine the child`s reproduction.

For instance, final gestures of constriction as well as initial fricatives were often eliminated. Plosives, however, tended to be preserved in initial position.

Elision of final gestures

JUL heiß *(hot)* [hạːe]
Baum *(tree)* [ɓaə]
Deckel *(lid)* [ˈd̥ɛkə]
ISA Schiff *(ship)* [ʃɪ̈]
Gockel *(cock)* [gyˈgaː]

Elision of medial gestures

JUL Teddy *(teddy bear)* [ˈtʰeɪ]
ISA Pullover *(pullover)* [pʊˈlʊə]

Elision of initial gestures

JUL Reiter *(horseman)* [haɛ̆tɑ̝]
ISA Vogel *(bird)* [ˈhoːçe]

Instead of a fricative in initial position children reproduced glottal gestures relatively often, e.g. **ISA`s**[ʔaɛ̆ç]: *weich (smooth)*; **SIM`s**[ʔɑ̝ŋ̊ŋ̊]: *Wagen (car)*, **JUL`s** [haen]: *rein (to put sth. in sth.)*, **ISA`s** [ʔɪtʰi]: *Zwietschi (toy duck).*

More rarely reproductions started with the following vowel of the source word, e.g. ISA`s [ˈyʂə]: *Füße (feet).*
We assigned this kind of variation to *Elisions of gestures,* because the glottal gestures were interpreted as varying reflexes of the necessity to avoid different oral constrictions per word. The variation between glottal stops, glottal fricatives and a smooth beginning of vocal fold vibrations could be explained in terms of unstable voicing.

Final cluster reductions

JUL	Kind *(child)*	[tɪn]
ISA	Milch *(milk)*	[mɪç]

Medial cluster reductions

JUL	Katze *(cat)*	[ˈkaʂej]
ISA	Flugzeug *(plane)*	[ˈʃuʃɷkˀ]
SIM	Knöpfe *(buttons)*	[gnœˈpɛ]

Initial cluster reductions

JUL	trinken *(to drink)*	[ˈd̥ɪtn̥]
ISA	Schwanz *(tail)*	[fa̰ts]
SIM	Frauke (first name)	[ˈʁa̰ɷə]

(b) Variation due to Lack of Articulatory Precision

In this group variation types were included whose phonetic forms deviate from their targets because either the coordination of gestures is unstable, certain constrictions lack one or more gestures, or the gestures match the direction of the target constriction but vary in their amplitude. The child`s abilities at this point to make precise articulatory adjustments do not allow for more stable productions. In general the resulting output varies at a particular constriction of the source word, whereas the other sound components are reproduced fairly targetlike. Moreover, tokens that show these phonetic deviations appear next to target-like tokens, i.e. they cannot be assigned to a particular stage of word development except for the fact that motor control is still developing. The following examples illustrate variation that is caused by unstable timing between vocal fold vibration relative to the movements of the articulators and varying amplitude of gestures.

Voicing

JUL

Baby [ˈb̥eb̥ɪ],[ˈb̥epiː],[ˈpʰypʰɪ]

Teddy [dɪhˈd̥ɪh],[ˈtɛt̬æ],[t̬ʰɪrɪː]

ISA

weich *(smooth)* [vaĕç],[faĕç]

Stopping

ISA	Wasser *(water)*	[ˈba̤ʂə]
SIM	auf *(open!)*	[q̥ɷ̃p]
JUL	Haus (house)	[haq̈ˀ]

Continuation instead of friction noise

ISA	Möve *(seagull)*	[ˈmywə]
SIM	weg *(out of sight)*	[weʔ]

(c) Variation That Cannot Be Attributed to Articulatory Constraints

The most interesting part of our data corpus is represented by those replica whose deviations from their targets cannot be explained sufficiently by articulatory constraints.
Particulary during their first attempts to reproduce a word children try to match the perceived sound features. As long as no articulatory routine can be established these tokens are phonetically variable, because the child may be liable to match different sound features each time as:

TILL: *Tasse (cup)* (1;00,10)
1.[aːb̥ë], 2.[b̥z̥ˈʐ̥q̥], 3.[d̥a̤hə̃],
4.[d̥aəː], 5.[dq̥ˈdq̥ʐ̥ə̃]

The fact that each token differs from its predecessor to a certain extent suggests that TILL fails to generate the appropriate motor commands, i.e. he is not able to imitate the perceived sound chain yet. It remains unclear, however, wether the word representation is indistinct or wether the transmission of afferent and/or efferent nerve impulses is still immature or both?
The example in Table 4. illustrates the role that ("fuzzy") word representations and immature mental processing might play in relation to output variation: ISA`s *weg (sth. or sb. is to disappear).*

Table 4. Variation of initial and final strictures of weg, Nr. of tokens=24, age: 1;08,30-2;00,15

Initial gest.	Final gest.	Example	%
palatal constr.	velar clos.	[çɛk]	50
		[jɛg̊]	
alveolar clos.	velar clos.	[dɛkʰ]	33,3
or constr.		[ʂɛkʰ]	
		[ʔneːk]	
labiodental	velar clos.	[vɛkʰ]	16,6
constr.	pal. constr.	[vɛç]	

ISA`s tokens of *weg* varied greatly in their initial sounds up to token no. 15 at 2;00,08. The following tokens displayed the form [çɛk] relatively stable.

Her input of *weg* had been standard German [vɛk] and colloquial North German [vɛç]. One explanation of ISA`s favorite output form could be that her representation of the end of the word includes not only a velar closure but also a palatal constriction. On the other hand, one can argue that [çɛk] reflects a leveling of dorsal gestures for the purpose of simplification. Since palatal constrictions belong to ISA`s preferred articulations and serve as substitutes in positions where labiodental constrictions are required in other words, too, (see Table 2.), this argument seems powerful.

But ISA had no difficulties in pronouncing an initial labiodental constriction as the tokens no.3 [vəg], no.8 [v̆vɛğ], no.9 [vɛç] and no.23 [vɛkʰ] illustrate. Due to the fact that palatal constritions occur much more frequently in ISA`s repertoire a more spontaneous and faster access to the appropriate mental processes may be assumed. Still the first question of how the palatal constriction is stored in ISA`s representation of *weg* remains unanswered, not to mention those renditions of the source word that display apical gestures in initial position (e.g. [ʂɛkʰ]).

Although conclusive explanations cannot be given, the data indicate that the initial sound of *weg* either has no distinct mental representation at all, or it may simply indicate, that access to a given representation is still insufficient. This results in the inability to imitate this sound structure in a consistent way, i.e. the child varies between dorsal, apical and labiodental articulations.

In contrast, the stable reproductions of the final sound suggest a distinct representation of a velar closure. However, if one takes into account that ISA`s caretakers provided a considerable amount of those variants that display a final palatal fricative in *weg*, one cannot be so sure about that.

OUTLOOK

Although only a few instances could be discussed and related to different types of phonological variation they suggest that early phonological variation cannot be related to articulatory restrictions as a general and exclusive explanatory framework. Constraints in terms of perceptual and mental processing abilities at a given aquisitional stage must be taken into consideration for an adequate account of the "inconsistency" in early speech production.

REFERENCES

[1] Wode, H. (1994), "Speech perception and the learnability of languages", International Journal of Applied Linguistics, vol.4,No.2, 143-168

[2] Jusczyk, P.W. (1992), "Developing phonological categories from the speech signal", Ferguson, C.A., Menn, L. & Stoel-Gammon, C. (eds.), *Phonological development: Models, research, implications.* Timonium, MD:York Press, 17-65

[3] Ferguson, C.A. (1978), "Learning to pronounce: The earliest stages of phonological development in the child", Minifie, F.D. & Lloyd, L.L. (eds.). *Communicative and Cognitive Abilities - Early Behavioral Assessment.* Baltimore: University Park Press, 237-297.

[4] Studdert-Kennedy, M. & Goodell, E.W. (1992), "Gestures, Features and Segments in Early Child Speech", *Haskins Labs. 1992,* SR-111/112, 89-102.

[5] Piske, T.(1995), "Articulatory Patterns in Early Speech Production", *this volume.*

ARTICULATORY PATTERNS IN EARLY SPEECH PRODUCTION

T. Piske
English Department, Kiel University, Germany

ABSTRACT

At the beginning of their lexical development children do not acquire segments or phonemes as claimed by e.g. Jakobson [1], Ingram [2], and others. Instead they operate with a limited inventory of articulatory patterns usually extending over more than one segment of traditional segmental phonology. The evidence is based on a longitudinal study of 8 L1 German monolingual children.

INTRODUCTION

It has often been reported that several of a child's first words may be highly similar to each other in terms of their phonetic structure and that the pronunciation of these words is fairly stable. This observation has led to the assumption that the mental 'construction' of phonetically similar forms is based on the same pattern or protoform that may be 'exploited' for the imitation of several, sometimes surprisingly different adult words. In the literature these patterns have been termed differently, e.g. prosodic schemata [3], vocal motor schemes [4], or gestural routines [5]. Here they are referred to as articulatory patterns since the structure of each pattern is obviously due to certain articulatory movements a child is able to control and to coordinate from a very early stage onwards. Although a number of studies have described such patterns, their true status in early child speech has not been clarified sufficiently yet, since they have primarily been discussed only on the basis of individual words from individual children. No attempts have been made, however, to interpret those findings within the framework of a comprehensive set of developmental data which cover the range of early vocabulary items of a larger number of subjects. This paper looks at a) the extent to which children at the onset of speech really make use of articulatory patterns and b) the origin, nature and development of these patterns.

METHOD

Leon is one of 8 L1 German monolingual children whose data were collected within the Kiel Project on Early Phonological Development [6]. In the following sections his early utterances serve to illustrate some typical phenomena occurring at the onset of speech. Leon's linguistic development was followed in weekly recording sessions from age 0;11,22 to 2;03,26. During the time of investigation he produced 148 different lexical items. Their identification and differentiation was aided by his parents' recognition. For the 61 audio taping sessions, lasting for about an hour each, Leon wore a beyerdynamic TS 42 radio microphone concealed in a small pocket. The recordings were carried out with a Uher 4000 IC Report Automatic tape recorder connected to a beyerdynamic NE 42 receiver. The Kiel Project particularly focuses on the types and the extent of phonetic variation at the onset of speech. During each recording session the experimenters therefore tried to elicit as many tokens as possible of Leon's first words by the use of picture books and things familiar to him (food, toys, tools etc.). At the same time commands and questions such as "Please say X!" or "How do cats, dogs etc. go?" were avoided. Immediately after each recording session the experimenters transcribed the tapes, using the symbols of the 1993 revised edition of the IPA. A second 'control transcription' was carried out after the data collection had been finished.

THE STRUCTURE OF ARTICULATORY PATTERNS

It is hardly possible to determine the status of a chain of sounds in early child speech on the basis of its first occurrence. So when is it reasonable to suggest that a sound sequence represents an articulatory pattern? We start from the assumption that a phonetic structure represents an articulatory pattern only if it is used for a certain time to render at least two distinct lexical items. Moreover, it only appears legitimate to suggest that the pronunciation of an individual word is based on a specific pattern if the majority of its tokens over some period of time reflect a certain phonetic structure that is characteristic of other words as well. Following these two criteria, it is possible to assign 126 of Leon's first 148 words, that is about 85% of his total early lexicon, to 11 different patterns, some of which seem to be responsible for the structure of up to 26 different words. (See Table 1 for examples.)

What are the structural characteristics of the 11 articulatory patterns Leon relied on? A comparison of the words

Table 1. 11 articulatory patterns determining the structure of 126 of Leon's first 148 words. OF = onomatopoeic form, PW = protoword (word without model in adult lg.).

Pattern refl. from reflected by X words	Structure of Pattern (short segmental description, IPA-symbols)	Examples: *target word* [target structure] (meaning)	Examples: examples of Leon's tokens
Pattern 1 0;11,22 8 words	[n ~ ɲ ~ ŋ] + [a ~ æ ~ ɛ] or [ae], from about 1;09,14 also + [ə ~ ɐ ~ ɪ]	*nein/nee* [naen ~ neː] (no) *PW 'hurry up'* *nicht* [nɪçt] (not)	[nɛ ~ ŋa ~ nːaŋ] [næˈn̪a ~ naŋnaŋ] [ʔəˈnəç ~ ʔɛ̆hɛ̆ˈnəç]
Pattern 2 0;11,22 22 words	[t̪ ~ t ~ d̪ ~ d] + [a ~ æ ~ ɛ] or [ə ~ ɪ ~ ʏ ~ ɷ], from about 1;08,10 also + [i ~ e] and [ɔ]	*da* [d̪a] (here/there (it is)) *'scheiß Katze'* ['ʃaesˈkʰa̤tsə] *zwei* [tsv̥ae] (two)	[d̪a ~ d̪æ ~ d̪ɛ ~ ta] ['d̪tʰɛ ~ 'd̪ad̪aˈtʰɛ] [d̪æ]
Pattern 3 0;11,22 25 words	glottal activity before V-syll., i.e. [ʔ ~ h ~ ɦ ~ ø] + V., e.g. [a ~ æ ~ ɛ], [ə ~ ɷ], [ae ~ eɪ]	*ey/hey* *Affe* ['ʔafə] (monkey) *blau* [b̥laɷ] (blue)	[ʔeɪ ~ heɪj ~ ʔɛɪː] ['ʔɑpa ~ 'ʔabɐ] [ʔɷlʲ ~ ʔɷʌ ~ ʔɷl]
Pattern 4 1;00,00 26 words	[p ~ pʰ ~ b̥ ~ b] + [a ~ ɑ], [ə ~ ɷ] or [ae], from about 2;01,22 also + [e ~ u] and [ɪ]	*Ball* [b̥al] (ball) *Papa* ['pʰapʰa] (daddy) *Bauer* ['b̥aɷ ɐ] (farmer)	[pʰa ~ baeːˑ ~ paɪ] ['baba ~ 'papʰə] ['bɷtʰæːˑ ~ bɷːˈda]
Pattern 5 1;00,14 22 words	[m] + [a ~ ɑ ~ ɛ], from about 1;11,20 also + [ə ~ ɪ ~ ʏ ~ ɷ]	*Mama* ['mama] (mum) *Mann* [man] (man) *Müll* [mʏl] (garbage)	[mɑmː ~ 'mama] [man ~ mɑn] [mɷl ~ ʔɛ̆'mɐʌ]
Pattern 6 1;00,14 4 words	[v ~ w] + [a ~ ɑ], [ə ~ ɷ], from about 1;11,13 also + [i ~ ɨ]	*PW 'turning/rotating'* *OF 'dog'* *OF 'flying etc.'*	[ə̆'vɑw ~ ʔəˈvɑwə] [wəwə ~ wɷw] [wi ~ 'wiɷ.wɨ]
Pattern 7 1;01,11 18 words	[h] + [a] or [ae], from about 1;06,23 also + [ə ~ ɪ], [u ~ y ~ i] and [ia]	*Vogel* ['foːg̊l] (bird) *hallo* ['halo] (hello) *zwei* [tsv̥ae] (two)	[huˈhu ~ 'ɸuhu] [hae ~ 'hae.a̤] [hae]
Pattern 8 1;03,17 5 words	[g] + homorganic syllabic nasal [ŋ̍] (velar plosive released through the nose)	*PW 'pointing at s.o./s.th.'* *Andrea* [ˌʔanˈdɐɛːɐ] (a name) *OF 'church'*	[gŋ̍ ~ gn̩ gŋ̍] ['gŋ̍.ɑ ~ 'gŋ̍.ga] [gŋ̍.gŋ̍ ~ gn̩.gə̃]
Pattern 9 1;03,24 6 words	coronal fricative, i.e. [θ ~ s̪ ~ s ~ ð ~ z̪ ~ ɕ ~ ʃ], often very long	*OF 'insects'* *OF 'merry-go-round'* *waschen* ['vaʃn̩] (wash)	[dɪɕː] [ðːː ~ z̪ ːː] [ɕ]
Pattern 10 1;08,10 6 Wörter	[aɷ ~ aɔ], at first often preceded by [ʔ] or [h], from about 2;02,27 also preceded by [m]	*auf* [ʔaɷf] (on/onto) *Haus* [haɷs] (house) *Maus* [maɷs] (mouse)	[ʔaɔfʰ ~ ʔaɔ̆f] [haɷfː ~ haɷɸ] [mːaɔf ~ maɔɸː]
Pattern 11 1;10,14 6 words	[ʃ ~ ɕ] + [i ~ y ~ u ~ o] or [ɪ ~ ə ~ ɷ], from about 2;02,27 also + [ae]	*tschüs* [tʃys ~ ʃys] (bye-bye) *Kerstin* ['kʰɛɐstin] (a name) *Fleisch* [flaeʃ] (meat)	[ʃu̯ ~ ʃyʃː ~ ʃiʃ] ['ɕi.ɐ ~ 'ʃi.ɐ ~ 'ʃɪ.ɑ̤] [ɕaɪɕ ~ ʃaɪʃ ~ ʃaeʃ]

listed under 'Pattern 5' in Table 1 shows that the production of many (22) of Leon's first words was obviously based on an articulatory pattern whose most characteristic element was a bilabial nasal [m] that was at first preferably combined with open or mid-open vowels, i.e. [a ~ ɑ ~ ɛ]. Later, from about 1;11,20 on, however, the bilabial nasal was also produced in combination with central or centralized vowels, i.e. [ə ~ ɪ ~ ʏ ~ ɷ]. Leon's imitations of words that were produced at a later stage thus indicate that in the course of time the patterns become more flexible with regard to their structural properties. The word *Mama*, which was first registered when Leon was 1;00,14, seemed to function as a kind of trigger for pattern 5 because he soon started to reproduce several target words in a form very similar or even identical to his renditions of *Mama*. All these words were apparently coded on the basis of the same articulatory pattern or protoform.

The second column of table 1 offers a short segmental description of the structure of each individual pattern. Some of the typical characteristics of these patterns should be noted here. Each of the 11 patterns was obviously based on those articulatory movements Leon could already control in a more or less safe way at a very early stage. It is striking that some time or other almost every early word was integrated into one of the 11 patterns. In other words, at some point Leon tried to produce almost every word by relying on preferred articulations. On the whole he operated with a rather restricted inventory of articulatory routines. He clearly preferred coronal articulations as far as consonantal elements are concerend. Plosives and nasals were much more often produced than fricatives, approximants or trills. At least until age 1;8 front vowels, i.e. open and mid-open unrounded front vowels, clearly dominated over close and back vowels. The majority of Leon's patterns can be described as a combination of a rather stable consonantal element and a more variable vocalic element, because many of his early words involved sequences like [na ~ nə], [da ~ də], [ba ~ bɷ] or [ma ~ mɛ]. But even structures that were not CV, e.g. [gŋ̩], took over the function of patterns, cf. e.g. the examples listed under patterns 8 and 9 in Table 1.

THE DEVELOPMENT OF PATTERNS

If the production of a word is based on articulatory movements a child is able to control and to coordinate in a more or less consistent way, the pronunciation of this word will of course be fairly stable. This phonetic stability provides the grounds for guaranteeing a crucial and indispensable feature of communication: intelligibility. The more uniform and consistent the use of patterns, the easier the task for the interlocutor to get involved in communicating with the child. If a child's pronunciation of a word varies too much, the people he or she talks to will have difficulties in identifiying the objects of reference. Yet phonetic stability also competes with the necessity to reproduce a word in a fairly target-like way. Consequently, the early articulatory patterns will have to undergo further development so that structurally more complex words can be produced. Leon's data and the data of the other Kiel children indicate that there are basically 5 ways in which the original patterns change over time.

1) All patterns that are characterized by an initial consonantal element reflect a kind of 'internal' development. While the consonantal element serves as a kind of basic element within the pattern and remains fairly stable, the vocalic element tends to be subject to a higher degree of variation after a given time. One effect of the internal change of a pattern is that the degree of homophony among the words based on the same pattern decreases gradually.

2) From about 1;08,04 on Leon obviously started to 'combine' already established patterns with each other. This has different effects: a) patterns with a fairly simple structure such as CV are expanded into more complex patterns that allow the production of forms that may, for example, have a $C_1V_1C_2V_2$ structure; b) 'integrating' a word into a pattern or a combination of patterns may also lead to a stabilization of its pronunciation. Initially Leon's forms for the word *Bauer*, for example, varied drastically. Later, however, he produced very stable renditions of this word and these were obviously based on a combination of patterns 2 and 4. (See Table 1 for examples.)

3) The development of patterns through combination does not only involve the combination of patterns as a whole. Patterns may also be expanded by adding just components of other patterns. This facilitates the production of words with a C_1VC_2 structure. Such a development is, for example, illustrated by Leon's forms for *Mann* (c.f. Table 1) and *gemein*. In this case the typical structure of pattern 5 [ma] is expanded by adding an alveolar nasal [n] in final position. From the first recording session onwards [n] was the most characteristic element of pattern 1.

4) Leon also combined already established patterns with new components, i.e. elements that are not characteristic of any other pattern. Between 2;01,22 and 2;06,06 he, for example, started to expand patterns 3 and 5 by producing several words, e.g. *blau* and *Müll* (c.f. Table 1), that showed the typical structure of these two patterns but that additionally had a lateral approximant [l ~ ʎ] in final position. Before that period of time lateral approximants had hardly been produced at all.

5) Several words also changed from one pattern to another. In most cases the words changed from a rather early pattern to one that only developed at a later stage. Such a development was, for example, reflected by Leon's forms for *zwei*. Initially they reflected the structure of pattern 2 and later that of pattern 7 (cf. Table 1). It seems as if by changing the patterns the child tries to develop a more target-like but at the same time still stable pronunciation of a word.

SOME CONCLUSIONS

On the whole an analysis of Leon's first words shows that his early phonological development cannot adequately be described as the acquisition of individual phones or phonemes acquired within a given developmental sequence. Leon rather seemed to operate with a limited inventory of articulatory patterns. The more Leon's motoric and organizational abilities developed, the more complex these patterns became. Moreover, his data suggest that it is not quite adequate to assume that initially phonological organization is strictly based on the individual lexical item as claimed in most studies on early phonological development. Leon's early utterances rather show that articulatory patterns which have provided the framework for a fairly stable pronunciation of one word are 'exploited' for the reproduction of other words. The patterns seem to enable children to organize the phonological information specifying a word in such a way that a fairly stable pronunciation of this word is facilitated. At the same time they seem to establish the first systematic links between children's early words. Consequently, patterns represent a first possiblity for children of going beyond the domain of the individual lexical item when they organize their early speech production.

REFERENCES

[1] Jakobson, R. (1941), *Kindersprache, Aphasie und allgemeine Lautgesetze*, Uppsala: Almquist & Wiksell.

[2] Ingram, D. (1989), *First language acquisition: Method, description and explanation*, Cambridge: University Press.

[3] Waterson, N. (1971), "Child phonology: A prosodic view", *Journal of Linguistics*, vol. 7, pp. 179-211.

[4] McCune, L. & Vihman, M.M. (1987), "Vocal motor schemes", *Papers and Reports in Child Language Development*, vol. 26, pp. 72-79.

[5] Studdert-Kennedy, M. & Goodell, E.W. (1992). "Gestures, features and segments in early child speech", *Haskins Laboratories Status Report in Speech Research 1992*, SR-111/112, pp. 89-102.

[6] Wode, H. (1994), "Speech perception and the learnability of languages", *International Journal of Applied Linguistics*, vol. 4, No. 2, pp. 143-168.

An Investigation of Rhythmic Processes in English-Speaking Children's Word Productions

M. Kehoe and C. Stoel-Gammon
Department of Speech and Hearing Sciences, University of Washington, Seattle, USA

ABSTRACT

This study examines whether English-speaking children's productions of multisyllabic words are consistent with metrical constraints or perceptual biases. Children (aged 22-34 months) produced three-syllable words which varied across stress pattern and segmental content. Overall results indicate a complex interaction between metrical and syllable-level constraints which change with development.

INTRODUCTION

There is very little research on children's development of rhythm and stress in English. There are reports, however, indicating that children delete syllables and add syllables in certain positions more than others, and alter stress patterns in systematic ways. Some investigators account for these patterns in terms of metrical constraints; other investigators propose perceptual biases as the underlying mechanism.

Proponents of the metrical constraint view argue that children have difficulty producing utterances that do not conform to a strong weak (SW) metrical pattern [1],[2]. The SW pattern denotes a trochaic foot, a unit of stress in metrical phonology. Therefore, children produce *MONkey* correctly because it conforms to a SW pattern but they delete the initial syllable of *giRAFFE* (WS) producing *RAFFE*, because it does not conform to a SW pattern. Proponents of the perceptual salience view argue that children produce stressed or word-final syllables more frequently than other syllables because of their perceptual salience [3]. Therefore, children produce *MONkey* correctly because they perceive the stressed and final syllable, but they produce *giRAFFE* as *RAFFE* because they do not perceive the unstressed non-final syllable. The predictions of these two approaches result in similar error patterns for two-syllable words, which has been the main focus of current research.

It should be noted that the perceptual salience hypothesis accounts only for children's truncation patterns. It has less predictive power when used to explain processes such as stress shift and epenthesis. However, because there is little documentation of these processes in English, their relative frequency cannot be ascertained. Investigations on the acquisition of stress in Dutch-speaking children indicate that stress shift and epenthesis also provide strong evidence of metrical templates [4]. This study examines truncation, stress shift, and epenthesis patterns in English-speaking children's productions of three-syllable words with the aim of separating out perceptual salience and metrical factors.

Predictions

The study is based on the following perceptual and metrical predictions: If a perceptual constraint is operating, children's productions of three-syllable words with the stress patterns: ŚWS (e.g., *DInoSAUR*); SWŚ (e.g., *KANgaROO*); SWW (e.g., *Elephant*) and WSW (e.g., *toMAto*) should show similar truncation patterns. Children should reproduce the stressed and final syllable equally frequently in these sets of words. If a rhythmic constraint is operating, truncation rates in these three-syllable words should vary. Truncation rates should be greatest in SWW and WSW words because application of a SW or trochaic foot results in deletion of one weak syllable. In contrast, application of a SW foot will not necessarily result in deletion of a weak syllable in ŚWS and SWŚ words, because the weak syllable is contained within the trochaic template. The pattern of truncations predicted by metrical constraints is shown in Figure 1. In terms of error patterns, different results are expected for SWW words. The prediction of the perceptual salience hypothesis is that the stressed and final syllable will be reproduced; the prediction of the metrical hypothesis is that the first weak syllable will be reproduced.

The metrical hypothesis also predicts that stress errors should be more frequent in SWŚ words, because main stress on the final syllable is an exception to the English stress rule, and that epenthesis should be associated with the stress

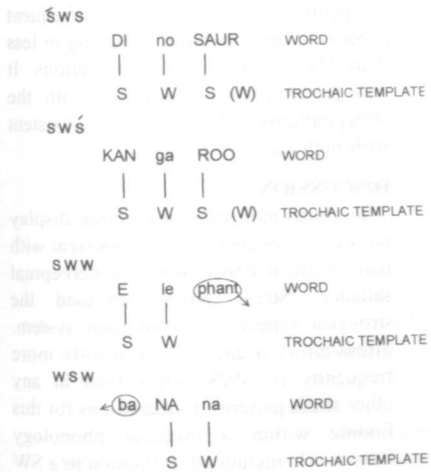


Figure 1. Pattern of truncations predicted by metrical constraints in three-syllable words.

patterns, ŚWS and SWŚ, because the addition of a syllable results in a canonical SW template.

These predictions were examined across different age ranges in order to determine if there were developmental trends in truncation patterns and across different segmental patterns, in order to determine if segmental effects influence truncation patterns. Pilot work indicated that words in which the unstressed syllable had a non-stop onset were more susceptible to deletion.

METHOD

The subjects included 18 children: 6 children aged 22-, 28-, and 34-months. The children participated in semi-structured elicitation tasks where they produced multiple tokens of both novel and familiar three-syllable words. Novel words were employed to control for familiarity and segmental effects on children's productions. The target words included four metrical patterns: ŚWS, SWŚ, SWW, and WSW words, and two segmental patterns: Words in which the unstressed syllable had a stop consonant onset versus a non-stop consonant onset. All productions were digitized and subject to both acoustic and perceptual analysis. A subset of the data was reanalyzed for inter- and intra-examiner reliability. All reliability measures exceeded 80%.

RESULTS

Results indicated significant stress and segmental pattern effects and significant age by stress pattern interactions. Results will be discussed separately for truncations, stress errors, and epenthesis.

There was a significant stress pattern effect for children's two-syllable truncations (i.e., productions in which two syllables are realized). In the novel word condition, children truncated ŚWS words significantly less than SWW and WSW words. However, there was no corresponding difference between the truncation rates of SWŚ, SWW, and

WSW words. In the familiar word condition, children truncated SWŚ words less frequently than all other stress patterns but there was no significant difference between the truncation rates of ŚWS and SWW words. These results were almost uninterpretable without further examination of segmental effects.

The study confirmed the significant segmental effect on truncations observed in the pilot study: Weak syllables were more frequently realized when they contained stop-consonant onsets than non-stop onsets. An item analysis revealed that this effect pertained specifically to internal weak syllable in ŚWS, SWŚ, and SWW words and was linked more closely to sonorant versus obstruent than to stop versus non-stop, as originally hypothesized. Words that contained intervocalic sonorants (e.g., /n/ or /l/), such as *TElePHONE* and *Animal* were more frequently truncated than words that contained intervocalic obstruents, such as *CROcoDILE* or *OCtopus*, regardless of metrical pattern. This pattern was interpreted as reflecting children's syllabification tendencies: Children are more likely to syllabify an intervocalic sonorant with the preceding stressed vowel than an intervocalic obstruent, leading to an onset-less medial syllable which is subsequently deleted. The intriguing stress pattern results described above were almost entirely interpretable in light of these segmental effects.

A comparison of stress pattern effects across age group revealed the following findings: The 22-month-old children truncated words from all stress patterns equally frequently; the 28-month-old children truncated WSW words more than all other stress patterns, consistent with the large body of evidence that weak syllables in word-initial position create difficulty for children; the 34-month-old children truncated words with intervocalic sonorants more frequently than other words.

An examination of error patterns in both novel and familiar SWW words indicated that the final weak syllable was most frequently preserved in children's truncations.

There was a significant stress pattern effect for children's stress errors. Stress errors refer to the perception of level or incorrect stress. Children displayed significantly greater numbers of stress errors in SWŚ words. They also made stress errors in other words, e.g., ŚWS words, consistent with quantity effects in a metrical framework. The analysis of stress errors across age and stress pattern revealed the following findings: 22- and 34-month-old children produced stress errors predominantly in SWŚ words whereas 28-month-old children produced stress errors in other stress patterns as well, thus suggesting that there is a period in development when stress patterns may be quite unstable. A closer examination of stress errors in the 28-month-old children showed a strong tendency for word-final stress.

Epenthesis was an extremely infrequent process in the data base, occurring in less than 1% of novel word productions. It was predominantly associated with the stress patterns, ŚWS and SWŚ, consistent with metrical predictions.

DISCUSSION

Findings indicate that children display prosodic strategies that are consistent with both metrical constraints and perceptual salience. Stress errors provided the strongest support for a metrical system. Stress errors occurred significantly more frequently in SWŚ words than in any other stress pattern. Explanations for this finding within a metrical phonology framework include: deformation to a SW metrical constraint, acquisition of an extrametricality rule, or acquisition of a main stress rule that assigns stress to the initial foot of a multisyllabic word. Epenthesis, although infrequent, was invariably associated with stressed syllables only (particularly primary stressed syllables), thus, preserving canonical feet or SW templates, as predicted in a metrical analysis.

The truncation results were the most difficult to interpret in metrical terms. Some aspects of the truncation findings showed that children were guided by metrical structure: For example, 28-month-old children truncated WSW words more frequently than other stress patterns. For the large part, however, results with two-syllable truncations did not show that children distinguished amongst the metrical patterns: ŚWS, SWŚ, and SWW. The truncation rates were most strongly influenced by a segmental effect which appeared to reflect children's syllabification tendencies.

It is hypothesized that children's truncation patterns are most consistent with a parsing strategy that scans from right to left and circumscribes at the position of stress [5]. The selection of syllables for production is guided not only by metrical templates but syllable structure constraints [2], and prominence factors, related to vowel quality and the acoustic salience of final position. The order of syllable mapping may be determined by a stored weighting system influenced by perceptual factors, or alternatively by a complex system of phonetic and phonological production constraints. The findings are less consistent with proposals that prosodic units are circumscribed [4], although it is true to say that children's one-syllable truncations (i.e., productions in which only one syllable is realized) almost always consist of the stressed syllable or foot closest to the end of the word. Nevertheless, the way children extend their productions suggest that segmental and syllable-level constraints play as important a role as metrical constraints.

REFERENCES

[1] Allen, G.D. & Hawkins, S. (1980), Phonological rhythm: Definition and development. In G. Yeni-Komshian, J. Kavanagh, & C. Ferguson (Eds.). *Child phonology: Volume 1. Production.* New York: Academic Press.

[2] Gerken, L. (1994), A metrical template account of children's weak syllable omissions from multisyllabic words. *Journal of child language*, vol. 21, pp. 565-584.

[3] Echols, C. (1993), A perceptually-based model of children's earliest productions. *Cognition*, vol. 46, pp. 245-296.

[4] Fikkert, P. (1994), *On the acquisition of prosodic structure*. Dordrecht: Holland Institute of Generative Linguistics.

[5] Archibald, J. (1995), The acquisition of stress. In J. Archibald (Ed.). *Phonological acquisition and phonological theory*. New Jersey: Lawrence Erlbaum Associates.

CONSONANT CLUSTERS IN DISORDERED L1 ACQUISITION: A LONGITUDINAL ACOUSTIC STUDY

James M. Scobbie, William J. Hardcastle, Paul Fletcher, Fiona Gibbon*
Queen Margaret College, Edinburgh, Scotland
**University of Reading, England*

ABSTRACT

We report a longitudinal instrumental study of the acquisition of English word-initial consonant clusters by children with phonological disorders. Interim results pertaining to /st/ are given for four children with phonological disorders and two normally-developing children. Duration measures show that before the successful acquisition of /st/, the stop closure is similar to that of the independent stops, but after, the stop closure duration for /st/ is shorter than for /t/ or /d/.

INTRODUCTION

This longitudinal instrumental acoustic study is motivated by the relative paucity of detailed quantitative information about the phonetic strategies used in the acquisition of consonant clusters by speech-disordered children. Most analyses of cluster acquisition in normal and disordered speech are transcription-based.

It is essential to explore *both* these aspects of the acquisition of phonological contrast. The developmental maturation of a speaker's ability to render adult-like phonetic encodings of a contrast, and the perceptual discontinuity at which the speech community (or phonetician) judges a contrast to be successfully acquired are separate processes.

Macken and Barton [1], for example, show that in the acquisition of the stop-voicing contrast, differences in VOT which are imperceptible to adult listeners are produced by some children. Such 'covert contrast' has also been demonstrated in [2,3,4]. The children then proceed to the production of an acceptable contrast, though initially it is not adult-like.

Applying these findings to consonant cluster acquisition, we might expect a stage of covert contrast in which the phonological contrast between cluster and singleton is expressed phonetically in a manner which can be detected instrumentally but is effectively imperceptible. Weismer [4] discusses a case in which, for example, [t] realising reduced /st/ has greater closure duration than [t] realising /t/. We would also expect to observe the gradual approximation of the child to the adult model.

The detection of covert contrast and the discovery of the ways in which children approach phonetically mature expressions of contrast would have important clinical implications [2].

The results presented here are taken from a larger study which will be able to explore the issues raised in greater depth.

METHODOLOGY

The four experimental subjects DB (4;1), SR (5;7), KG (4;1) and IB (4;3) were from the Edinburgh area, speaking varieties of Scottish English. The subjects were the age indicated at the start of the study. All exhibited functional cluster reduction of /st/ to [d] or [t], although subject KG had resolved already by the first session. All were undergoing courses of speech and language therapy having been diagnosed as phonologically disordered. In addition, we refer to two normally-developing children, RM (4;0) and JS (3;3).

The children uttered nine target words (forming minimal triples, Table 1) in a naturalistic manner as part of a series of ten picture-naming games, each game comprising three dissimilar targets from a larger dataset. No minimal pairs appeared in the same game. The carrier phrase was 'give me *x* (please)'.

Table 1. Coronal targets.

sty	tie	dye
steer	tear	deer
store	tore	door

Multiple tokens of each target were elicited within a game in random order, except for IB, who produced several tokens of one target together before moving on to the next. The recordings took place in a sound-treated booth, and used a Sony DAT DTC-690 tape deck, Alice microphone amplifier MIC-AMP-PAC-2 and a Sony ECM-77 lavalier microphone placed for optimum recording clarity.

Six tokens of each target were digitised at 40960Hz on a KAY CSL 4300. The waveforms were annotated to indicate the boundaries of significant acoustic events for durational analysis. Seven annotation points (t1-t7) were chosen (Table 2) and are illustrated in Figure 1. Often, annotation points were coincident.

Table 2. Annotation Point Definitions

t1	Start of breathy vowel offset at the end of carrier 'give me', if present.
t2	End of breathy vowel offset.
t3	Start of frication noise, if present.
t4	End of frication noise or beginning of stop closure, if present.
t5	Release of stop closure (burst).
t6	Onset of periodic phonation after burst.
t7	End of target word.

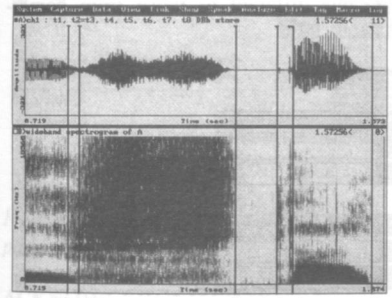


Figure 1. Annotated 'store' (DB).

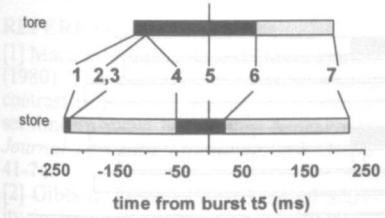


Figure 2. Key to graphs. 1-2 = carrier-final breathy vowel, 3-4 = [s], 4-5 = stop closure, 5-6 = VOT, 6-7 = target rime.

RESULTS

Mean durations of t1-t7 for each onset type are graphed in Figures 3-7 (n=18, 17 or 16 unless otherwise indicated). Figure 2, in conjunction with Figure 1, exemplifies the key to the graphs.

Figures 3a, 4a, 5a, 6a refer to each subject's first session, while Figures 3b, 4b, 5b, 6b refer to the second session, approximately four months later. In the text, DBa, for example, refers to DB's first session (Figure 5a).

DISCUSSION

In the cases with unreduced /st/ (Figures 5b, 6, 7), the mean stop closure duration (t4-t5) in [st] is less than in the independent stop phonemes /t/ and /d/. We might therefore have expected covert contrast to be signalled with a shortened closure in the stop realising reduced /st/. Alternatively, since unreduced [st] has greater duration as a whole than the closure phase of /t/ or /d/, then we might have expected, with Weismer [4], to find a *greater* stop closure in reduced /st/. In fact, in the sessions where subjects exhibit cluster reduction (Figures 3, 4, 5a), stop closure duration in reduced /st/ is similar to the independent stop phonemes.

The mean duration of [st] as a whole is greater than the singleton durations mainly because of the relatively long [s] (t3-t4). In DBb, where the cluster was only recently acquired, [s] was often extremely long, incorporating two or more separate energy peaks (Figure 1). KG's [s] sometimes exhibited near-complete cessation of frication medially.

Duration measures can hide the variability in the individual productions. In JS, the stop component of /st/ was typically spirantised, and effaced completely 50% of the time. (Nevertheless, most of these spirantised tokens contained bursts.) If the mean duration is calculated for just those closure durations which are not 0ms, her closure durations for /st/ are roughly commensurate with the independent stops. Such a revised mean would hide JS's typically weak stop closure, however, so all her tokens are used. This spirantisation could be due to her fast speaking rate or young age. A different type of variability was shown by SR, who typically reduced /st/ to [t]. In session 2, all six tokens of

'store' exhibited cluster reduction to [s:]. These strategies are graphed separately.

Voicing during the stop closure was open to great intersubject variability. Most closures were largely voiceless, but at the other extreme, IB's were often fully voiced during the stop closure, followed by voiceless aspiration (i.e. [d] for /d/ and /st/, [dʰ] for /t/).

The rime duration (t6-t7) is useful in relativising the duration of closure or VOT to that of the whole word. Rime duration is influenced heavily by speaking rate and phrasal position. The carrier phrase incorporated a final 'please', but this was only employed by JS, IB, DBb, and KGb. In addition, individual differences are a strong determinant of this measure.

Carrier-final breathy vowel offset (t1-t2) was generally greater for /t/ than /d/, and was a very consistent factor for KG. Given the immature nature of [s] in the children's /st/, a high value for this duration is to be expected for [st]. A similar increase might have been expected to appear in *reduced* /st/, as a covert contrast with /d/. Slowed closure or devoicing of [ɪ] might have boosted this duration, but no consistent increase was observed. The audible character of t1-t2 varied from [h] to [j] to [ç].

VOT (t5-t6) was similar for /d/ and reduced /st/. VOT was calculated from the release of closure to the detectable presence of periodic phonation. (F2 was detectable a cycle or two after t6.) An accurate measure of IB's VOT in /d/ and /st/ is made difficult by noise immediately after the burst obscuring any low amplitude phonation that might be present. (For IB's /t/, however, there is a clear long lag before voice onset.) In general, long-lag VOT for /t/ was often in excess of our Scottish English estimate of 55-75ms. The short lag values may also be rather high. This suggests a maturation model for the subjects with over-long VOT in the early stages, approximating later to adult values.

Given the excessive length of [s] in DBb, presumably the same pattern of hyper-duration followed by progressive approximation to the adult mean will be observed in his later maturation of /st/. As with all the measures, further sessions with these subjects will chart such development.

Finally, note that the use of different articulators, e.g. in [sp], might lead to results different to those from homorganic [st].

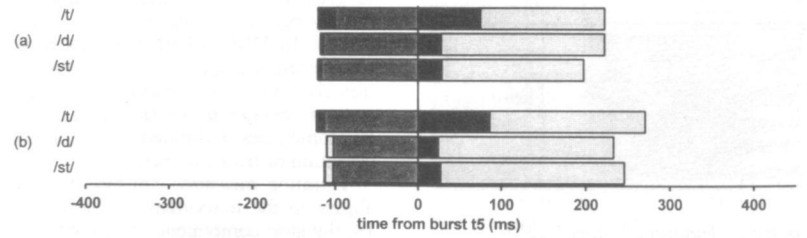


Figure 3. IB 4;3 (a) and 4;7 (b). Target phrase-medial, nonrandom order.

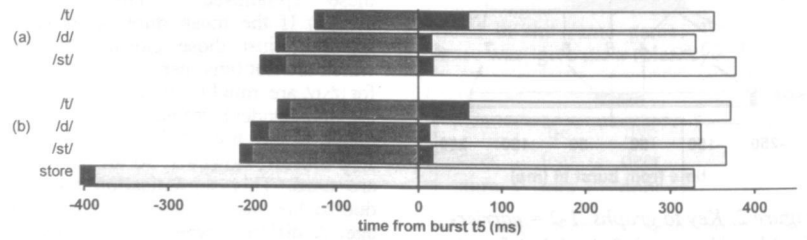


Figure 4. SR 5;7 (a) and 5;11 (n=12 for /st/) (b). Target phrase-final.

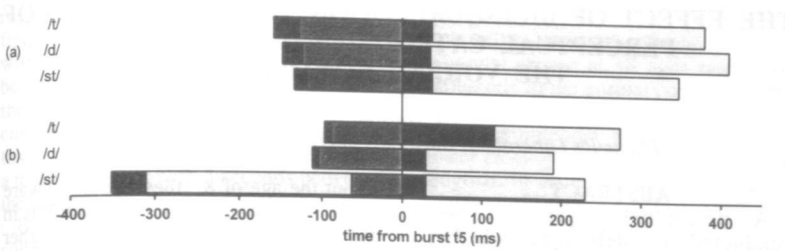


Figure 5. DB 4;1 phrase-final (a) and 4;5 phrase-medial (b).

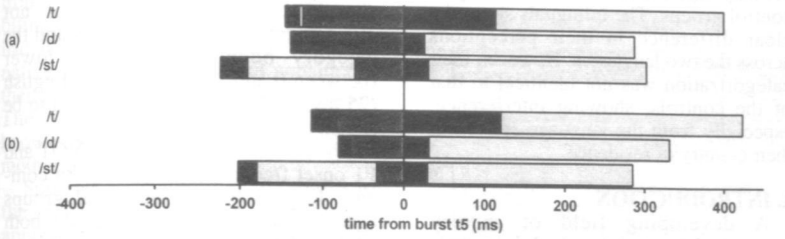


Figure 6. KG at 4;1 phrase-final (a) and 4;5 phrase-medial (b).

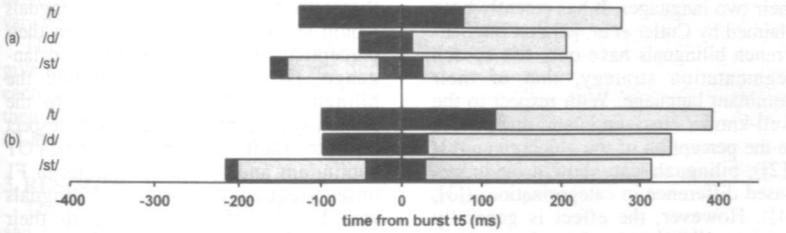


Figure 7. Control JS at 3;3 phrase-medial (n=8 for /t/) (a) and control RM at 4;0 phrase-final (b).

ACKNOWLEDGEMENTS

We gratefully acknowledge the support of the British Medical Research Council, grant #G 9117453 N.

REFERENCES

[1] Macken, Marlys A. and David Barton (1980) The acquisition of the voicing contrast in English: a study of voice onset time in word-initial stop consonants. *Journal of Child Language*, vol. 7, pp. 41-74.

[2] Gibbon, Fiona (1990) Lingual activity in two speech-disordered children's attempts to produce velar and alveolar stop consonants: evidence from electropalatographic (EPG) data. *British Journal of Disorders of Communication*, vol. 22, pp. 302-217.

[3] Forrest, Karen, Gary Weismer, Megan Hodge, Daniel A. Dinnsen and Mary Elbert (1990) Statistical analysis of word-initial /k/ and /t/ produced by normal and phonologically disordered children. *Clinical Linguistics and Phonetics*, vol. 4, pp. 327-340.

[4] Weismer, Gary (1984) Acoustic analysis strategies for the refinement of phonological analysis. In Mary Elbert, Daniel A. Dinnsen and Gary Weismer (eds.) *Phonological theory and the misarticulating child*, ASHA Monograph 22, pp. 30-52, Rockville MD: ASHA.

THE EFFECT OF BILINGUALISM ON THE ACQUISITION OF PERCEPTUAL CATEGORIES UNDERLYING THE VOICING CONTRAST

I. Watson
Phonetics Laboratory, University of Oxford, Great Britain

ABSTRACT

A perceptual experiment was conducted to determine whether bilingual children's development of the categories underlying the voicing contrast resembled that of monolingual control groups. The bilinguals showed a clear difference in their perceptions across the two languages. However, their categorization was not identical to that of the controls, showing interference, especially from the language spoken in their country of residence.

1. INTRODUCTION

A developing field of inquiry concerns the extent to which bilinguals are able to maintain strict separation between their processing strategies in their two languages. It has recently been claimed by Cutler *et al.* [1] that English-French bilinguals have only one speech segmentation strategy, that of their dominant language. With respect to the well-known cross-linguistic differences in the perception of the voicing contrast ([2]), bilinguals can show a language-based difference in categorization ([3], [4]). However, the effect is generally much smaller than has been found when comparing monolingual speakers of the relevant languages. Very little is known of the development of perception categories in bilinguals and the existing literature is largely concerned with secondary bilinguals (see, e.g. [5]). As shown by Simon and Fourcin [6], the ages between 4 and 10 involve a gradual development of adult-like perceptual responses, at least as shown by experiments based on synthetic speech-like continua. This development is cross-linguistically heterogeneous. In English, children acquire the ability to respond in a sharply categorical fashion to a VOT continuum at the age of 5. The average 50% crossover point in their labelling functions is higher than that of adults and slowly diminishes towards the adult norm over the next 5 to 7 years. At about the age of 8, they become aware of the perceptual salience of variations in the onset frequency of F1, with higher values causing a downward shift in the VOT category boundary, corresponding to fewer voiced percepts. In French, sharply categorical labelling is not attained until the age of about 8 and the category boundary is much lower (between 0 and 10 ms) than for English (25 ms). The F1 cue is not found to be salient in French.

The present experiment uses VOT and F1 onset frequency parameters to compare English-French bilingual groups aged 6, 8 and 10 to monolinguals, both adult and child. There were two groups of bilinguals at each age, one resident in England, the other in France. The hypotheses to be tested were: (i) bilinguals would show a clear difference in their perceptual responses according to language; (ii) at each age studied the bilinguals would be identical to the monolingual control groups with respect both to their response to the VOT continuum and to manipulations in F1 onset frequency; (iii) the bilinguals would not differ according to their country of residence, and (iv) they would show, like the monolinguals, a clear progression towards adult norms in both languages.

2. METHOD

A perceptual identification experiment was designed based around two synthetic VOT continua. The first contained tokens in which VOT varied from -30 to +50 ms in 5 ms steps. The tokens were heard as 'gash' or 'cash' in English and as 'gâche' or 'cache' in French. This continuum (henceforth the "normal"continuum) was used to investigate VOT categorization. The second continuum differed from the first in that the F1 onset frequency was held artificially high at a frequency of 685 Hz in short-lag tokens (with VOTs up to 40 ms). This was used in comparison to the normal continuum to assess responsiveness to a high F1 onset frequency. The continua were produced with each of two carrier phrases, one being the English "I say..." and the other, the French "Je dis...". For each continuum, each group of carrier plus token was recorded onto a tape recorder a total of 6 times with 3 seconds between the tokens.

Both monolingual and bilingual subjects were resident in Paris or London. The bilinguals had been exposed to both languages in the home from birth and at schools specifically for bilinguals. The adult monolinguals had had at least some schooling in their non-native language but did not considered themselves proficient in that language. The younger monolinguals had no knowledge of languages other that their mother tongue.

Subjects indicated their responses to the perceptual test by ticking the appropriate box under pictures which illustrated the target words. Bilinguals performed the experiment once in each language, the order of presentation being varied.

The resulting data were grouped to give overall identification curves for each subject group. These curves were then subject to logit analysis using the GLIM General Linear Models package.

3. RESULTS

The results are presented in Tables 1. and 2. Table 1. presents the mean values for the 50% - crossover points for each group for the normal continuum. Table 2. shows the effect of the F1 onset frequency manipulation in terms of the shift it produced in the group category boundary relative to the normal continuum. This style of presentation has been used due to limitations of space, rather then the traditional representations of identification curves. However, it must be borne in mind that it is the overall curves and not just the crossover points abstracted from them which formed the basis of the statistical processing. A difference of, for example, 3 ms difference in crossover point between the two continua may represent a significant shift in response curve in some cases, but not others.

3.1 Cross-linguistic comparison of bilinguals

It will be clear from Table 1. that all bilingual groups at all ages responded differently to the normal continua in the English and French conditions. The expected outcome was obtained of a lower cross-over point in French than in English. In all cases, the difference is highly significant ($p < 0.005$). There was also a strong tendency for the labelling functions to be less sharply categorical in French than in English .

Table 1. Mean VOT 50%-crossover values for normal continuum

a)

6 yr olds	English		French	
E. bilings		21.5	15.5	
Mono	25			8*
F. bilings		20	8.5	

b)

8 yr olds	English		French	
E. bilings		22	16.5	
Mono	22.5			-3
F. bilings		20	12	

c)

10 yr olds	English		French	
E. bilings		19	14.5	
Mono	21.5			10
F. bilings		17	11	

d)

Group	English		French	
Adult Mono	18.5			5*

N.B. * indicates non-monotonic function with more than one 50% - crossover point

3.2 Distinction between bilingual and monolingual groups

Contrary to the hypothesis in section 1, in most cases there are significant differences between bilinguals and the relevant control groups with respect to their categorization of the normal continuum. In all cases, these differences result from a greater similarity between the response curves of the bilinguals response curves being more similar than between those of the monolinguals.

Amongst the 6 year olds, the bilinguals resident in England differ to a statistically significant extent (here and in all further cases, $p < 0.01$) from both of their monolingual peer groups. The Paris-based bilinguals are dramatically different from the English control group but indistinguishable from their French

peers, (even though the French monolingual group has a non-monotonic labelling function with crossovers at both -0.5 and +8 ms.)

In contrast, the 8 year old Paris-based bilinguals responses differ significantly from those of both their French and English contemporaries. The London-based 8 year old bilinguals differ from the monolinguals in their French, but not their English labelling functions.

For the 10 year olds, the same pattern obtains as for the 8 year olds, that is, the bilingual results differ significantly from the monolinguals, except in the case of the London-based monolinguals' English responses.

There is thus a general pattern of difference between the bilinguals and monolingual control groups, although the older child subjects living in England produce results in line with their English monolingual counterparts.

The results for the F1 manipulation, are shown in Table 2., expressed as the shift (in ms) of the category boundary produced by the second (high F1) VOT continuum relative to the first.

Table 2. Shift for F1 manipulation

a)

6 yr olds	English		French	
E. bilings		1.5	**20**	
Mono	1.5!			3.5!
F. bilings		2.5	8.5	

b)

8 yr olds	English		French	
E. bilings		2	2	
Mono	**4**			0
F. bilings		**4**	0.5	

c)

10 yr olds	English		French	
E. bilings		**2**	1.5	
Mono	**3.5**			**7**
F bilings		0.5	1	

d)

Group	English		French
Adult Mono	**2**		**3.5!**

N.B. ! indicates a shift in the opposite direction to that predicted, i.e. an increase in /g/-reponses. Figures in bold indicate a shift which is statistically significant ($p < 0.01$).

The bilinguals depart from the monolinguals in a number of ways. However, these departures are no consistent, and the monolinguals themselves produce more heterogeneous results than those of Simon and Fourcin [6]. In general, there is a small movement of the category boundary in English. This is significant for the 8 and 10 year old monolingual groups, for the Parisian 8 year old bilingual group and for the 10 year old London-based bilingual group. None of the 6 year old groups show significant effects, nor do the Paris based 10 year old bilinguals.

In French, it will be recalled, no effect of the F1 manipulation is expected. In fact, responses vary dramatically from group to group. Most of the observed differences are insignificant and some bilingual groups which responded to this cue in English failed to do so in French. Nevertheless, there are some significant shifts of category boundary in French. The most dramatic is that of the 6 year old London-based bilinguals (20 ms) but the 6 year old Paris-based bilinguals and the 10 year old French monolinguals also show sizeable shifts of 8.5 and 7 ms respectively. The monolinguals' result argues against any straightforward account of these findings in terms of language contact. The bilingual subjects might be transferring perceptual skills from English to French. However, if subjects with no knowledge of English exhibit the same behaviour, such transfer cannot be the only available explanation.

No clear patterns emerge, therefore, from the F1 manipulation. In general, the F1 cue is used by the older children in English but not French, and bilinguals are capable of using it in one language, but not the other. However, the contrast between the two languages is less stark than earlier studies had suggested.

3.3 Differences between bilingual groups

Although the differences in category boundary between the bilingual groups are frequently small in terms of milliseconds, they are nevertheless all statistically significant. The effect is apparently related to the language of the country of residence dominating the other. Thus, London-based bilinguals always have higher category boundaries in both languages than Paris-based bilinguals. This difference does not appear to vary across the age groups.

3.4 Progression towards adult norms

As shown in Table 1., the 50% crossover for the English adult group is situated at 18.5 ms. The English monolingual groups do show a pattern of slow decrease in crossover values toward this figure, the age-based difference being significant ($p < 0.01$). There is a similar, and similarly significant, trend in both bilingual groups.

In French, there is no consistent pattern. The adult category boundary falls at +5 ms, although in this case, there is again a non-monotonic function with a crossover at -2.5 as well as +5 ms. There are statistically significant distinctionsbetween the different age groups, and between the children and tha adults, but no obvious developmental pattern emerges. Given that the adults do have a lower category boundary than most of the child groups, . it may be that adult-like values are attained at a later stage by French speakers than is encompassed in this experiment.

The majority of crossover points observed in French in this study lie in the region between +5 and +17 ms. This is a somewhat higher range than has been referred to in earlier literature ([6]) but it is consistent with data from a production study conducted with the same subjects. This shows that although the majority of voiced tokens in French are produced with pre-voicing, a substantial minority have VOT in the short-lag region.

4. DISCUSSION

Cutler et al ([1]) have shown that in some respects even strong bilinguals have a dominant language. The present study demonstrates that the same is not entirely true with respect to the voicing contrast.. All the bilingual groups showed a clear difference in identification functions dependant on the language they believed they were hearing. Furthermore, in several groups the F1 cue trading relation was in evidence in one language - English - but not the other, even thought the stimuli were identical in both cases. However, the disparities between the bilingual and monolingual groups, show that bilingualism does affect perceptual processing. Furthermore the different responses of the two bilingual groups indicate that even while they maintained different categories in their two languages, that spoken in their country of residence had the greater influence.

Despite these differences in detail, no clear differences in developmental pattern emerge between bilinguals and monolinguals. The youngest bilingual subjects dealt with here had already developed distinct perceptual categories in English and French. Those categories develop in a similar way to monolinguals in English (lowering of category boundary, development of F1 trading relation). In French, there is no such apparent development, but this is equally true for the monolinguals.

Bilingualism may modify the details of perceptual processing and its acquisition, but it is possible to be perceptually bilingual.

Acknowledgement

I should like to thank the staff of all the schools and nurseries which have allowed me access to their children, especially the Lycée Charles de Gaulle in London, and the Lycée International in Saint Germain en Laye.

REFERENCES

[1] Cutler, A., Mehler, J., Norris, D., & Segui, J. (1989) "Limits on bilingualism", *Nature*, vol. 340, pp. 229 - 230.

[2] Lisker, L. and Abramson, A.S.. (1964) "A cross-language study of voicing in initial stops: acoustical measurements",. *Word*, vol. 20, pp. 384-422.

[3] Elman, J., Diehl, R., & Buchwald, S. (1977) "Perceptual switching in bilinguals", *Journal of the Acoustical Society of America*, vol. 62, 971-974

[4] Hazan, V. & Boulakia, G. (1993) "Perception and production of a voicing contrast by French-English bilinguals", *Language and Speech*, vol. 36 pp. 17-38.

[5] Flege, J. & Eefting, W. (1987) "Production and perception of English stops by native Spanish speakers", *Journal of Phonetics*, vol. 15, pp. 67-83.

[6] Simon, C. & Fourcin, A.J. (1978) Cross-language study of speech pattern learning. *Journal of the Acoustical Society of America*, vol. 63, pp. 925-935.

THE PROSODIC STRUCTURE OF LEFT-DETACHED PHRASES IN FRENCH INTERROGATIVE UTTERANCES

F. Sabio, A. Di Cristo & D.J.Hirst,
Institut de Phonétique , URA CNRS 261 "Parole et langage"
Université de Provence, France

ABSTRACT

Previous studies have shown that statements and questions in French differ in the iterative pre-nuclear pitch patterns : statements being generally characterised by a sequence of (downdrifting) rising patterns and questions by a sequence of downstepped pitches. This study examines the patterns on left-detached phrases before statements and questions. Their tonal patterns appear intermediate between that of statements and questions irrespective of the modality of the following construction.

INTRODUCTION

In French as in many languages there are a number of different ways of asking questions. Yes-no questions, for example, can be syntactically marked by postposing a subject pronoun (but not a full noun phrase) after the verb : *"As-tu la clef?"* (Have you got the key), or by use of the expression "est-ce que" : *"Est-ce que tu as la clef?"*. Yes/No questions can also be syntactically unmarked : *"Tu as la clef?"*.

Prosodically, although questions can be produced with either a final rising or a falling pattern, a common characteristic of different question intonation patterns in French seems to be the use of a recurrent downstepping pattern which can be observed on both Yes-No questions and Wh-questions as well as on other interrogative patterns such as alternative questions (*"Tu as la clef ou tu l'as pas?"* (Have you got the key or haven't you?)), tag questions (*"Tu as la clef, non?"* (You've got the key, haven't you?) and elliptical questions (*"Et la clef?"* (What about the key?)) [1, 2].

This recurrent downstepping pattern found on questions in French contrasts with the recurrent rising pitch pattern (with downdrift) found on non-final accented syllables in declaratives.

In this presentation we examine the case of a more complex structure in French : that of left-detached phrases. This construction, which is extremely frequent in colloquial French, consists in extracting either the subject or the object of a verb and replacing it with a clitic pronoun as in : *"La clef, elle est sur le bureau."* (The key, it is on the desk.) *"La clef, tu l'as trouvée?"* (The key, have you found it?). The phenomenon of detachment can be interpreted pragmatically as a form of topicalisation. Since the detached sequences are liable to appear both in declaratives and questions, we wanted to see whether the opposition mentioned earlier between the recurrent downstepping pattern and the recurrent rising pitch pattern would also apply to detached sequences.

METHOD

Most of the examples of left-detached sequences in our corpus of spontaneous speech were too short (one or two accent groups) for us to examine the pre-nuclear iterative pattern which only appears on sequences containing at least three accent groups. We consequently decided to test this feature on a simulated dialogue.

Corpus

We selected 6 noun phrases from an written dialogue which was read aloud by 10 native French speakers. The reading of one speaker was rejected for lack of naturalness. The dialogue contained 6 noun phrases which were structurally identical : consisting of an embedded NP of the type

[N [Prep + N [Prep + N]]]

They also had the same lexical content, and hence the same segmental structure. The final word varied (Paris / Roissy / Madrid) but in each case the final vowel was [i].

The three phrases were :

les horaires des départs vers Paris
les horaires des départs vers Roissy
les horaires des départs vers Madrid

(the schedule for the flight to Paris/Roissy/Madrid)

The sequences differed in syntactic structure and modality :

Three of the phrases were part of interrogative utterances :

a) Final Question (QF) : independant interrogative noun-phrase.

les horaires des départs vers Paris ?

b) Detached Question (QD) : left-detached phrase before an interrogative construction.

mais les horaires des départs vers Paris, vous pouvez me les donner?

[but ..., can you give it to me ?]

c) Non final Question (QN) : This construction - frequent in spoken conversational French - is similar to the " detached question " (b), but in this case only the detached part is expressed, giving the construction an elliptical nature.

et alors pour les horaires des départs vers Madrid?...

[so what about...?]

The other three phrases were part of declarative utterances :

a) Final Declarative (DF) : declarative phrase ending a turn.

Je voudrais connaître les horaires des départs vers Madrid.

[I'd like to know...]

b) Detached Declarative (DD) : left-detached phrase before a declarative construction.

Les horaires des départs vers Madrid, on vous les donnera si (...).

[..., it will be given to you if...]

c) Non-Final Declarative (DN).

Je voudrais les horaires des départs vers Paris, mais (...)

[I'd like ... but (...)].

All of the sequences share the rhythmic property of constituting a single Intonation Unit comprising three Tonal Units (or accent groups) of three syllables each :

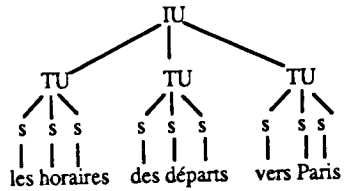


Figure 1 : prosodic structure for the noun phases used in the analysis.

The utterances were digitalised and Fundamental Frequency was modelled as a sequence of target-points defining a quadratic spline-function using an automatic f0 modelling program MOMEL [3].For the first two tonal units, two targets were sufficient ; for the last TU, three f0 targets were often necessary. Thus, each sequence was reduced to 7 f0 values for the study. The Hz values were then converted to the ERB scale [4], in order to allow comparisons across speakers.

RESULTS

The only factor which significantly distinguished all three types of questions from all three types of declaratives was the value of the initial f0 peak which was systematically higher for questions than for the corresponding statements ($p < 0.02$).

For the two turn-final patterns QF and DF, the f0 patterns produced by the 9 speakers were very consistent. The two patterns were distinguished systematically both by the value of the final target point and by the drop between the non final accented syllables and the following syllable which was much greater for the declarative pattern than for the interrogative as can be seen in Figure 2 illustrating the mean values for the two patterns over the 9 speakers.

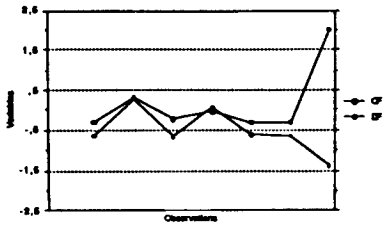


Figure 2 : mean values in ERBs for the target points for turn final question (QF) and statement (DF).

The distinction between a downdrifting pattern for declaratives and a downstepping pattern for interrogatives [5, 1] seems to be a fairly robust characteristic of questions as can be seen from the pattern of non-final utterances illustrated in Figure 3 where this characteristic is the only one which distinguishes the two patterns significantly.

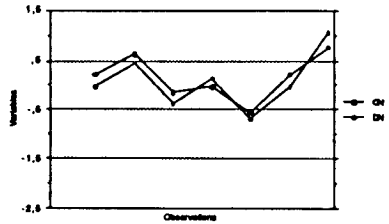


Figure 3 : mean values in ERBs for the target points for non-final question (QN) and statement (DN).

Left-detached phrases seem however to behave rather differently from other phrases in French. As can be seen from Figure 4 the mean values of the question and statement patterns do not seem to differ on this feature : on the contrary both seem to be somewhat intermediate between the patterns observed on final or non-final questions and statements.

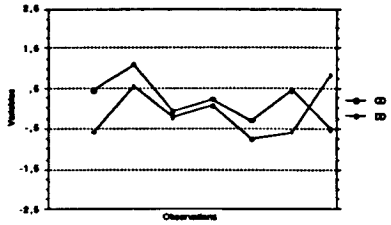


Figure 4 : mean values in ERBs for the target points for detached question (QD) and statement (DD).

DISCUSSION

First of all, our data confirms some previous observations about prosodic differences between declaratives and questions in French [2].

We can note a tendancy for final questions (QF) and elliptical questions (QN) to be produced with an iterative downstep, and for declarative utterances - both final and continuative - to have a downdrifting contour (where L and H tones alternate).

However, if we consider the mean values drawn from of our corpus, the detached sequences (QD, DD) seem to differ from both downdrifting and downstepping contours. Their intonation pattern - intermediate between those two patterns - would be better described as a "reduced" downdrift contour : the second TU does show a tendancy towards tonal alternation between H and L tones, but these are of significantly smaller amplitude than that observed on declaratives. In addition, the fact that the detached sequence was followed by a declarative or interrogative construction seemed to play no part in determining the nature of the reduced downdrift. Thus we can hypothesise that the phonological opposition between the downdrift and the downstep - semantically linked to the distinction between assertion and question - is neutralised in detached sequences. A semantic explanation for this neutralisation of modality would be beyond the scope of this study : we can however note that a major characteristic of topicalised detached phrases seems to be that they remain outside the scope of the modal marks carried by the verbs. For example, it is obvious that answering

"non" to a question like "Les clés, elles sont sur la table ?" [the keys, are they on the table ?] would not negate the existence of the keys, but only their being on the table [6].

As for the pattern of the last Tonal Unit, detached questions were characterised by greater variability than the other patterns. By contrast with the Detached Declarative sequences, which all show a globally rising contour on their last TU, the last segment of the Detached Question was produced as a flat or falling contour most of the time, but as a rising contour by two speakers. However, DD and QD sequences appear globally to clearly differ regarding their last TU. In particular the QD sequence finished with a tonal value which can be interpreted as downstepped.

It remains to be seen whether the tonal characteristics which we have observed on questions, statements and detached phrases are perceptually relevant cues for the identification of sentence modality.

References

[1] Di Cristo, A. (forthcoming) "intonation in French" in Hirst & Di Cristo (eds) *Intonation Systems : a Survey of Twenty Languages* . Cambridge: Cambridge University Press.

[2] Di Cristo, A. & Hirst, D.J., (1993), "Prosodic regularities in the surface structure of French questions", *Working Papers*, vol.41, pp. 268-271.

[3] Hirst, D.J. & Espesser, R. (1993), "Automatic modelling of fundamental frequency with a quadratic spline function", *Travaux de l'Institut de Phonétique d'Aix*, 15, 71-85.

[4] Hermes, D.J. & Van Gestel, J.C. (1991), "The frequency scale of speech intonation", *JASA*, 90, 97-102.

[5] Hirst, D.J. & Di Cristo, A. (1984), "French intonation : a parametric approach", *Die Neueren Sprachen*, 83, 5, 554-569.

[6] Sabio, F. (1995) "Micro-syntaxe et macro-syntaxe du français. L'exemple des compléments antéposés", *Recherches sur le français parlé*, 13, 111-157.

PROSODIC FEATURES OF FINALITY FOR INTONATION UNITS IN FRENCH DISCOURSE

Magali Vincent , Albert Di Cristo, Daniel Hirst
Institut de Phonétique, URA CNRS 261 "Parole & Langage"
Université de Provence, 13621 Aix en Provence, France
email : {phonetic; di cristo; hirst} @univ-aix.fr

ABSTRACT

This study aims to bring to light the prosodic features which contribute to the identification of finality in discourse. Perception tests applied to spontaneous speech extracted from radio interviews allowed us to derive two indices of finality. The main acoustic correlate was the F0 parameter. We found no evidence that silent pause or duration played an effective role in the perception of finality, however there was a significant drop in intensity on the final syllable.

INTRODUCTION

It is generally assumed that one of the main linguistic functions of prosody is a structural or organisational one, which consists in segmenting the flow of speech into coherent units of different sizes [1], [2].

At some high structural level, prosody serves to indicate a binary distinction between finished and unfinished utterances. At lower levels, prosody can also be used to signal boundaries of intermediate units such as prosodic phrases and prosodic words. In recent years, a large number of studies dealing with different languages have been devoted to investigating the correspondance between prosodic parameters and boundary type and/or magnitude. While the results of these studies are consistent with the fact that some parameters (mainly F0 and pauses), are used in similar ways across languages to signal different levels of boundaries, the situation is less clear in the case of segmental lengthening. Data on this subject is often controversial. The literature on this topic suggests that the role of duration as a boundary marker is probably dependent on both language, [3], [4], [5], and mode of discourse [6], [7].

Practically all research on this topic has been concerned with carefully controlled laboratory speech. More recently, however, a series of perceptual and acoustic studies of the role of prosodic cues of finality (essentially contour type, register and range) have been carried out in discourse for Dutch [8], [9].

For French there is very little data on this subject. Crompton [10] suggested that preboundary lengthening in French is correlated with the degree of the following boundary, being greater before final boundaries than before non-final ones. This claim was only partially confirmed by Fletcher [11] in a study of isolated sentences.

In this paper we present preliminary results of a study concerning the contribution of prosodic cues : silent pause, intensity, duration and pitch-patterns, to the identification of finality in discourse, as well as the relative importance of these different cues in context and in isolation. Although there are almost certainly cues which pre-signal boundaries quite early in an utterance [12], we concentrate in this study on acoustic parameters in the immediate vicinity of the boundary.

CORPUS

Our corpus is composed of two extracts of radio broadcast interviews. For each interview we extracted and analyzed the answer to a question asked by a journalist. The first extract is uttered by a female speaker and the second one by a male speaker. Both speech turn which last approximately 1.5 minutes each correspond to a complete utterance, characterized by syntactic and semantic coherence.

TEST 1

A preliminary listening test was carried out to see how far listeners would agree on the presence of boundaries in spontaneous utterances, and how far, they would agree on the nature of the boundaries.

Experimental procedure : Ten postgraduate phonetics students listened to both recordings and were asked to add punctuation (restricted to comma, semi-colon and full-stop) to an orthographic transcription of the recording which contained no punctuation marks.

TEST 2

The aim of the second listening test was to see whether subjects were able to identify utterances as utterance final out of context.

Experimental procedure : Ten subjects with no previous experience in prosody listened first to recording 1 and then to recording 2. The recordings were broken up into 26 extracts for recording 1 and 23 extracts for recording 2. Each extract corresponded to a (final or non-final) boundary which had been identified in Test 1 by at least 50% of the subjects. The extracts were presented in random order. Each subject heard each extract three times so that in total each subject listened to 147 extracts.

Subjects were asked to key their responses to each extract as being either utterance final or utterance non-final. The responses were recorded on a microcomputer.

The results of the two tests were used to calculate two indices : a *contextual finality index* calculated as the number of subjects who had marked a full-stop at the corresponding boundary and an *isolated finality index* calculated as the number of responses designating the boundary as utterance final, divided by 3 to give a score out of 10.

Although the two indices were highly correlated ($r = 0.75$, $p < 0.0001$) they were significantly different ($t = -4.651$, $p < 0.0001$). As can be seen in Figure 1 the contextual finality index was practically always lower than the isolated index.

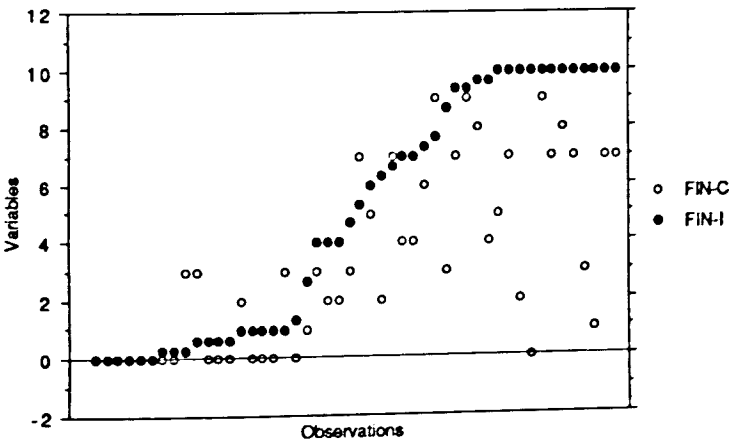


Figure 1 : isolated (FIN-I) and contextual (FIN-C) finality indices sorted by increasing value of the isolated index.

For statistical analyses the boundaries were split into three categories FINAL, NON-FINAL and INTERMEDIATE for each of the two indices, using 25% and 75% as threshold values. Thus scores between 0 and 2.5 were classified as NON-FINAL, scores between 2.5 and 7.5 as INTERMEDIATE and scores between 7.5 and 10 as FINAL. The contingency table for the nine intersecting categories thus defined is given in table 1.

	final	int.	non-final	Totals
final	5	0	0	5
int.	10	8	3	21
non-final	3	4	16	23
Totals	18	12	19	49

Table 1 : contingency table for boundaries classified as final, intermediate or non-final in context (rows) and isolated (columns).

It can be seen immediately that there is a major dissymmetry in that all 5 values coded as final by the contextual index are also coded final by the isolated index but that more than two thirds of the values coded final by the isolated index are coded as intermediate or non-final by the contextual index.

This is consistent with the idea [12] that the perception of a final boundary is carried out essentially on the basis of cues which precede the boundary. Cues which follow the boundary may weaken the perception of finality but they will not convert a non-final boundary into a final one.

Acoustic Analysis :

The corpus was digitised and analysed on a SUN-Sparc station using a MES, a signal editing environment developped by R. Espesser. Phoneme labels were placed manually. F0 was analysed and modelled as a sequence of target points defining a quadratic spline function using an automatic modelling algorithm with manual correction [13]. For each of the 49 boundaries used in the auditory tests, the following parameters were measured :

Silence : duration of silent pause following the boundary.

Intensity and duration : intensity was measured at 1/3 of the duration of the vowel. The duration of vocalic nucleus for the penultimate, final and following syllable for each boundary were measured.

Fundamental frequency : the penultimate final and following target-points were taken.

Normalisation.

The duration of the vocalic nuclei were normalised using a z-transform [14]. F0 targets were normalised with an ERB scale offset to the mean of the speaker's range [15].

Dynamic aspects

In order to capture some dynamic aspects of the prosodic features, for each triplet of values described above as penultimate, final and following we calculated successive difference values (i.e. penultimate minus final and final minus following). In all 26 acoustic features were calculated for each of the 49 boundaries analysed.

An analysis of variance was carried out for each of the 26 acoustic features on both the Isolated and Contextual classifications of the boundaries.

RESULTS

Silence : there was no significant difference in the duration of the silent pause following the boundary for either the isolated ($p = 0.114$) or the contextual classifications ($p = 0.278$).

Intensity : a significantly lower intensity value was found for the final syllable preceding terminal boundaries than for boundaries classified as either non-terminal or intermediate. This difference was significant for both contextual ($p < 0.001$) and non-contextual classifications ($p < 0.0001$).

Duration : There was no significant difference between the normalised duration of the vocalic nucleus of the final syllable preceding a final and a non-final or intermediate boundary. From the raw data the duration of this vowel appeared significantly shorter for final boundaries than for non-final or intermediate boundaries but this effect disappeared after normalisation and was probably due to the intrinsic duration of the vowels in question. There was however a difference on the penultimate vowel which was significantly shorter before a final boundary ($p < 0.05$, $p < 0.005$).

Fundamental frequency : The strongest effects ($p < 0.0001$) were found on both the absolute value of the final F0 target which was lower before a final boundary than before a non-final or intermediate boundary, and the difference between the penultimate and the final values. These differences were highly significant both before and after normalisation and for both types of classification.

DISCUSSION

Our results did not confirm the hypothesis that preboundary lengthening is greater before terminal boundaries than before non-terminal ones. Similar results were reported for English by Wightman et al. [16] who suggested that the distinction between different types of major boundaries "must be distinguished by other cues such as pausing and intonation" [p. 1716].

There was, however, no evidence from our data that the presence and/or duration of a silent pause played an effective role in the perception of terminal boundaries.

As expected, the strongest effect in our data was linked to the F0 parameter, principally that of register as expressed by the normalised pitch target preceding the boundary.

An additional finding was that a drop in intensity on the final syllable was highly correlated with the perception of finality. Despite the extemely relative nature of intensity this parameter was highly significant even though the speech analysed was not recorded under laboratory conditions.

REFERENCES

[1] Bruce, G. (1985), "Structure and functions of prosody", *Proceedings of the French-Swedish Seminar on Speech.* Guerin, B. & Carré, R. (eds.), Grenoble, France, pp. 549-559.

[2] Beckman, M. (1986), *Stress and Non-Stress Accent*, Foris : Dordrecht.

[3] Thorsen, N. (1980), "A study of the perception of sentence intonation evidence from Danish", *Journal of Acoustical Society of America*, vol. 67, pp. 1014-1030.

[4] Berkovits, R. (1984), "Duration and fundamental frequency in sentence-final intonation", *Journal of Phonetics*, vol.12, pp. 255-265.

[5] Farnetani, E. (1989), "Acoustic correlates of linguistic boundaries in Italian : a study on duration and fundamental frequency", *Eurospeech* (Paris), vol.2, pp. 332-335.

[6] Klatt, D. (1975), "Vowel lengthening is syntactically determined in a connected discourse", *Journal of Phonetics*, vol.3, pp. 332-335.

[7] Klatt, D. & Cooper, W.E. (1975), "Perception of segment duration in sentence context", *Structure and process in speech perception*, Cohen, A. & Nooteboom, S. (eds.), Springer-Verlag, pp. 69-89.

[8] Swerts, M. (1994), *Prosodic Features of Discouse Units*, Doctoral Dissertation, Eindhoven University of Technology.

[9] Swerts, M. & Bouwhuis, Don G. and Collier, R. (1994), "Melodic cues to the perceived "finality" of utterances", *Journal of Acoustical Society of America*, vol. 96, pp. 2064-2075.

[10] Crompton, A. (1980), "Timing Patterns in French", *Phonetica*, vol.37, pp. 205-234.

[11] Fletcher, J. (1991), "Rhythm and final lengthening in French", *Journal of Phonetics*, vol.19, pp. 193-212.

[12] Swerts, M. & Geluykens, R.C. (1994), "Prosody as a marker of intonation flow in spoken discourse", *Language & Speech*, vol. 37, pp. 21-43.

[13] Hirst, D.J. & Espesser, R. (1993), "Automatic modelling of fundamental frequency with a quadratic spline function." *Travaux de l'Institut de Phonétique d'Aix*, vol. 15, pp. 71-85.

[14] Campbell, W.N. (1992), *Multi-level Timing in Speech*, PhD thesis, University of Sussex.

[15] Hermes, D.J. & Van Gestel, J.C. (1991), "The frequency scale of speech intonation.", *Journal of Acoustical Society of America*, vol. 90, pp. 97-102.

[16] Wightman, C., Schattuck-Hufnagel, S., Ostendorf, M. and Price, P. (1992), "Segmental durations in the vicinity of prosodic phrase boundaries", *Journal of Acoustical Society of America*, vol. 91, pp. 1707-1717.

THE ACCENTUAL PHRASE AND THE PROSODIC STRUCTURE OF FRENCH

*Sun-Ah Jun * and Cécile Fougeron °**
** Dept. of Linguistics, UCLA, USA ° Inst de Phon., CNRS URA-1027, Paris III*

ABSTRACT

A model of French intonational phonology is proposed based on pitch tracks of three Parisian speakers' read utterances. The lowest prosodic level defined based on tone is an Accentual phrase and has a LHLH pattern. This model is compared with the models of previous proposals.

INTRODUCTION

French intonation has long been known to have a sequence of rising pitch movements. Most studies so far have been phonetic and acoustic descriptions of French intonation. Phonological descriptions of French intonation have recently been proposed, among others, by Hirst and Di Cristo (H&DC) [2] [3], Mertens [5], and Post [8]. They agree that a tone is associated with a stressed syllable and stress is rhythmic or postlexical. They also agree that an utterance is hierarchically organized with different prosodic levels, but they disagree in the levels below the intonational phrase (IP) and their tonal representation.

H&DC propose two levels below IP: a Rhythmic Unit "governing" a Tone Unit, while Mertens proposes an Intonation Group "governing" a Stress Group. These prosodic units are delimited by a prominence-lending syllable (or stressed syllable). On the other hand, based on the syntactic structure, Post proposes a Phonological Phrase as the domain of a pitch accent, which is a tonal property of a stressed syllable. The Tonal Unit, Stress Group, or a pitch accentable syllable are all based on stress including the secondary (or initial) stress, and the Rhythmic Group and Intonation Group are delimited by the primary (or final) stress.

In this paper, we examine the tonal variation related to secondary and primary stress, their domains, and syllable-tone association. We propose that the lowest prosodic level defined by tone in French is the Accentual Phrase (AP) [1]. The AP, which can encompass more than one word, is the domain of the primary and the secondary pitch accent, and has a tonal pattern of LHLH. We call this prosodic unit an AP because, as in Japanese [7] and Korean [4], the AP-final rise in French is a delimitative tone. By assuming the underlying four tones associated with certain syllables within an AP and phonetic implementation rules [1,7], we can capture the relation between the secondary and primary pitch accents, and explain various realizations of French phrasal intonation.

METHOD

Pitch tracks of utterances read by three Parisian French speakers (1 male, 2 female) were analyzed using XWAVES. 59 sentences with different syntactic structures were repeated four times in random order. In these sentences, there were three major syntactic categories, Subject NP, VP and PP. Within each category, the number of syllables and the number and position of words were varied. Examples (selected):
"*Marie* mangera des bananes"
(Mary-will eat-some-bananas)
"*Marie-Anna* mangera des bananes"
(Marianna-will eat-some-bananas)
"*le mauvais garçon* ment à sa mère"
(The-bad-boy-lie-to-his-mother)
"*le garçon malade* ment à sa mère"
(The-boy-sick-lie-to-his-mother)
"Marion mangera *au dîner* des bananes"
(Marion-will eat-for-diner-some-bananas)
"Marion mangera des bananes *au déjeuner*"

(Marion-will eat-some-bananas-for-lunch)

In addition, 46 test words were embedded in a carrier sentence, "X est un mot utilisé par les français" (X is a word used by French people). We put the word in sentence initial position, where the secondary (initial) stress is more likely to be realized [6]. The number of syllables of the words were varied from two to eight syllables: e.g.

médite, méditer, méditation,
méditérranée, méditérranéen,
méditérranéiser, méditérranéisation.
Each sentence was repeated 4 times by each speaker.

RESULTS

1. Justification of AP as one group /LHLH/, not two /LH/s

As shown in Fig.1(a), one long word is realized as two rising movements: the first rising around the secondary (initial) stressed syllable and the second rising at the primary (final) stressed syllable. In Fig. 1(b), two words show two similar rising movements: the first around the secondary stressed syllable of the first word and the second at the primary stressed syllable of the second word.

The sequence of rising patterns (LH) as shown in Fig. 1(a, b) has been analyzed as two groups, e.g. Tonal Units or Stress Group. But we found that these two patterns are closely related with each other. They look similar regarding the timing of the two rising movements: the initial and final rising each takes about 200-300 ms regardless of the number of syllables in each rising part. But when we look at the falling movement, as shown in Fig. 2, we see that the falling timing from the first peak to the beginning of the second rise increases as the number of syllables increases between the two points. Thus, we propose that the two tones, (initial) H and (the following) L, are target tones within one AP, and the interpolation between these two tones covers the pitch realization of the intermediate syllables.

More important, if we compare this falling time within one AP with the falling time between the AP-final-H and the following AP-initial-L (Fig. 3), we can see that the falling timing across APs does not depend on the number of syllable of the following AP, but is rather constant, about 100-200 ms.

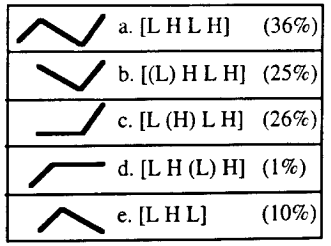
Thus, if we consider a sequence of LH as a separate tonal unit, we cannot explain why the timing between two tonal units, LHLH, is highly correlated with the number of syllables in the second tonal unit, and why it is not the same across all LH tonal units. We need to distinguish the different fallings, and group the connected tonal units into the same group: /LHLH/$_{AP}$ /LH...

2. Tone association with a syllable

The AP initial L and final H are realized in the first and the last syllable of AP, respectively. The L preceding the final H is associated with the penultimate syllable of an AP. For APs shorter than 3 syllables, this L tone is often realized on the final syllable. The initial H shows a variation in its realization on a syllable; the first to third syllable of the first lexical word. It may be "loosely" associated with the second syllable of the lexical word and realized earlier or later. Variation in the realization of the initial H, i.e. the initial stress, has been a big question in French, and, as shown by Pasdeloup [6], various factors such as phonotactic, rhythmical, contextual constraints may account for this variability.

3. Realization of AP

We observed five patterns of AP shown below. The frequency of each pattern out of 466 tokens (Subject NP position only) is indicated in parenthesis.



a. [LHLH]: all underlying tones are realized; the most common pattern, especially in phrases longer than 4 syllables.

b. [LH]: undershoot of medial HL; common in one or two syllable phrases.

c. [LLH]: initial peak is not realized by phonetic undershoot, or for intentional or pragmatic reasons.

d. [LHH]: undershoot of L between two Highs; when a short phrase has both initial and final accent.

e. [LHL]: (i) when the AP has an initial H (initial stress) and the following AP also has an initial H. (ii) when the AP is in the IP final position.

The four patterns (a-d) can be explained by phonetic rules like undershoot, while the realization of the last pattern (e) depends on the tonal

context and the prosodic hierarchy. To explain the last pattern, we propose a constraint to avoid three H tones. (*HHH : a sequence of three H tones are maximally avoided). A sequence of two H tones are also often avoided (*HH), explaining the low frequency of type 'd' AP.

In addition, when the AP is in the IP final position, we assume that the AP's H boundary tone is preempted by a higher level (IP)'s boundary tone. Thus, when the IP boundary tone is L%, AP is realized as [LHLL].

DISCUSSION

Our model differs from H&DC's model in that we assume our AP (two of their Tonal Units) as the lowest prosodic level defined by a tone, and we consider Word as a part of prosodic hierarchy. H&DC's Tonal Unit boundary does not necessarily match the word boundary, violating the Strict Layer Hypothesis [9]. Their phrasing is more tone-driven, thus their model is closer to the surface representation

Our model is similar to Mertens' model in that both models assume that the lowest prosodic unit based on tone, i.e. our AP and his Intonation Group (IG), includes both the primary and the secondary pitch accent. But our AP differs from his IG (=((NA)AI) (NA) AF (NA)) in that we assume four underlying tones while he assumes only one underlying tone, AF, and an optional tone after the final stressed syllable. In addition, he assumes each tone can be L or H with four levels of height. Thus, his representation is much closer to the phonetic representation.

Our AP also differs from Post's Phonological Phrase (PP) in the way the prosodic unit is defined. His PP is defined based on a syntactic structure. Our AP formation is constrained by, but not predictable from, syntactic structure. But, as well known, a tonal unit does not always match a syntactic structure. Our AP also differs from his PP in that we allow no APs without pitch accent and we allow only H toned pitch accents.

By assuming an Accentual Phrase with an LHLH underlying tonal pattern, tone interpolation, and phonetic undershoot, we can explain and predict various tonal contours of one AP, the slope from the initial peak (H) to the following L, and the same tonal pattern of a long one word AP and that of AP having more than one word.

Next, we can also compare French AP patterns and its realization with those of other languages such as Japanese and Korean. The Japanese AP has at most one pitch accent, which is linked with the underlying pitch-accented syllable, while French AP has two pitch accents, which are linked with the postlexically stressed syllables. On the other hand, Korean (Standard dialect) AP (LHLH) has no pitch accent, but only has phrasal tones which are linked to a certain position within the AP [4]. As in French, the falling timing (LHLH) in the Korean AP is correlated with the number of syllables within an AP.

CONCLUSION

We propose that, in the intonational phonology of French, the Accentual Phrase has the tonal pattern of LHLH and it is the lowest prosodic level that is tonally defined. The AP includes both the initial and the final pitch accented syllable, and thus has at most two peaks. When an AP includes more than one word, the initial pitch accent is realized at the initial stressed syllable of the first lexical word and the final pitch accent is realized at the final stressed syllable of the last word within the AP. But due to undershoot and the tendency to avoid adjacent H, we observe five possible tonal realizations.

So far, we proposed two prosodic levels, AP and IP, in French. Further research is needed to find out if French has a prosodic level intermediate between AP and IP, i.e. an intermediate phrase (ip) as in Japanese and English.

Since this model is based on read speech in the laboratory setting, we plan to apply our model to utterances with different speech styles and to spontaneous speech.

REFERENCES

[1] Beckman, M. & J. Pierrehumbert (1986) "Intonational structure in Japanese and English," *Phonology Yearbook* 3, 255-309.

[2] Di Cristo, A. & D. Hirst (1993) Rythme syllabique, rythme mélodique et representation hierarchique de la prosodie du français. *Travaux de l'Inst. de Phon. d'Aix, 15*, 9-24.

[3] Hirst, D. & A. Di Cristo (1984) French intonation: a parametric approach. *Die Neueren Sprachen* , 83.

[4] Jun, S.-A. (1993) *The Phonetics and Phonology of Korean Prosody*. Diss. Ohio State Univ.

[5] Mertens, P. (1993) Intonational grouping, boundaries and syntactic structure in French. *ESCA Workshop on Prosody, Lund WP.* 41: 155-159.

[6] Pasdeloup, V. (1990) *Modèle de règles rythmiques du français appliquées à la synthèse de la parole.* Doctorat Univ. Aix en Provence.

[7] Pierrehumbert, J. & M. Beckman (1988) *Japanese Tone Structure*, MIT Press, Cambridge, Mass.

[8] Post, B. (1993) *A Phonological Analysis of French Intonation*, M A thesis. The Netherlands.

[9] Selkirk, E. (1986) "On Derived Domains in Sentence Phonology," *Phonology Yearbook* 3:371-405.

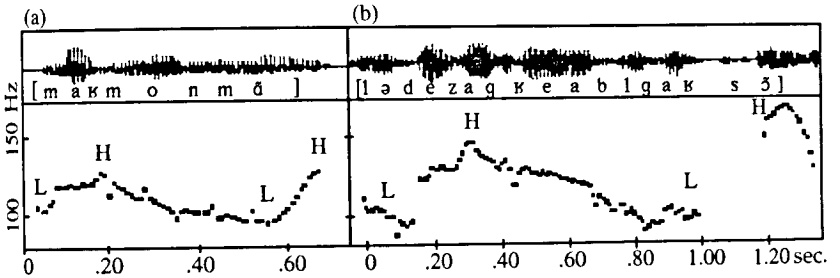


Figure 1. Pitch tracks of (a) One-word AP : "Marmonement est un mot utilisé par les français" and (b) Two-words AP : "Le désagréable garçon ment à sa mère", both produced by a Parisian-French male speaker (only the underlined part is shown).

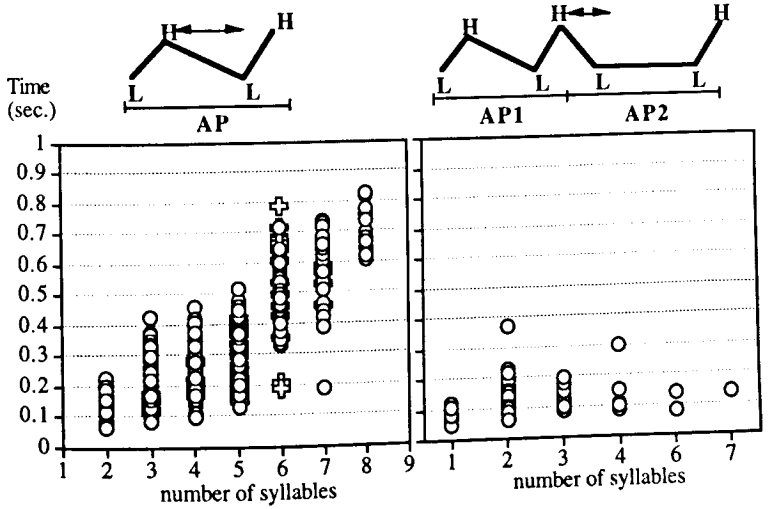


Figure 2. Falling timing within one AP, from the initial H to the following L, depending on the number of syllables in one-word APs (circles) and two-words APs (crosses).

Figure 3. Falling timing across APs, from AP-final H to the following AP-initial H, depending on the number of syllables in the second AP.

PREDICTABILITY OF DIFFERENT ATTITUDINAL INTONATION IN STANDARD CHINESE

Zong-ji Wu
Institute of Linguistics, CASS, Beijing, P.R.China

ABSTRACT

In the sentence intonation of Standard Chinese(SC), the range registers of phrasal contours(PC) are shifted to different extents according to various attitudes. Normalization by frequency transposition after converted the F0 into chromatic scale, the PC values are rather identical that can be predicted based on given patterns.

INTRODUCTION

The intonation contours of SC sentences are a compound of interrelative constituents of tonemic and poly-syllabic tone-sandhis patterns in PCs; as well as attitudinal modified inter-PC glidings, initial and ending drifts in global sentential contours. Most of them present the surface forms rather different from their original patterns. In our lab., fundamental studies on PC's tone-sandhis were carried out[1,2,3], for which poly-syllabic tone-sandhi rules were derived and had been tested in synthesis[6].

As for the SC intonation, two obstacles, say, the largely deviations of the absolute registers and the frequency ranges caused by the modification of sentence attitudes, had bewildered phoneticians and engineers for long time. It can be analyzed by F0/ST(semitone) conversion and frequency transposition. Thus, predictability of SC intonation may be realized in the near future.

MEASUREMENTS AND ANALYSIS

For the problems of absolute registers and frequency ranges, a number of spoken sentences with different moods collected from different sources were analyzed[4]. Measurements were done by segmentation of the PCs according to the margins of sense groups. It was found that the tone contours of some PCs with their registers modified by the sentential expressions will be higher or lower; however, their contour patterns are similar to the forms of normal moods. Also, the contour ranges of all the PCs were identical to each other in the scale of melodic intervals if the linear F0 scale had been converted into chromatic semitones(ST). Thus, any PC in different register can be normalized to a given pattern by frequency transposition or change key process. It means that all the melodic ranges among the PCs in the sentences of different moods spoken by the same speaker are rather consistent. Table I shows two sentences with different moods measured both by F0 scale and chromatic one. The absolute F0 and their ranges are different among all PCs, while the absolute key notes of the PCs are shifted to different extent, however, their ranges in ST are all identical, with some exceptions in ending drifts. Moreover, another experiment was arranged[5], in which hundreds sets of four SC syllables were spoken by two Pekinese speakers, male and female. The speakers were asked to pronounce isolately with tone 1 tuned to three

Table I: The upper-lower thresholds and the F0 range of different PCs in an question sentence and an exclamatory one

Sentence moods	Phrasal contour	F0 Range in Hz*			Melodic key in ST**		
		Upper	Lower	Width	Upper	Lower	Width
Inter-rogative	PC1	260	170	90	C4	F3	7
	PC2	350	230	120	F4	B3	7
	PC3	270	170	100	C4	F3	7
Exclama-tory	PC1	250	120	130	B3	B2	12
	PC2	350	180	170	F4	F3	12
	PC3	390	200	190	G4	G3	12
	PC4	270	140	130	C4	C3	12

Note: *Upper threshold, lower threshold and width of PC range
**ST=semitones(concert scale)

different keys. The result showed however the upper thresholds of the four tone ranges were different from each other, their lower thresholds were all shifted to an extent of the same range-width in ST scale.

DISCUSSION

In the PCs of sentence, even their surface forms are far different from their underlying patterns, it can be recognized by rules[3]. As for the analyzing of global contours, the PCs modified by intonation and the inter-PC glidings, can be processed by PC segmentation, ST conversion and register transposition. The initial and ending drifts modified by sentence intonation are appeared in different surface forms such as the raising or lowering contours, the neutralization of the first and/or the last syllables of the PCs. These are much more verified by different moods and individual speakers respectively, they were not easy to be normalized and further studies have to be going on.

There is another problem in the analysis of SC intonation, i.e., how do the prosodic features other than tones play the roles in intonation. It is worthwhile to mention that there were not a few papers presented to a symposium of prosody in Yokohama on account of the relation of the three prosodic features to intonation. Lehiste stated that perceptions of prosody are differed in language backgrounds, e.g.,

stress as prominent cue in English and duration in Estonian. Ohala gave the result that there is no significant difference between a "clear" speech and repeated one. It is interesting to compare these results with that in SC.

As in Chinese, It is true that the tone pitch is the chief prominent cue in intonation. A number of tentative experiments on speech synthesis had shown that in a synthesized exclamatory sentence the absolute register of a prominent PC raised to a certain extent is not significant in loudness unless a certain degree of amplitude and/or duration are added to an appropriate proportion.

Measurements of further more samples show that the proportions of the three prosodic features are differed in different cases. Fig.1(a,b,c) are three samples of SC sentences with their prominent cues in PCs contributed by different prosodic features: (a)lengthening duration in 1st PC; (b)increasing of all three features in last"repeated"PC; and (c)a cheer with three"long live"s louder and louder by increased durations. It shows that the prominence cues can be represented by any of the prosodic features, or both, or all, in SC intonation. Percentages of the three features taken part in different prominence cues will be a current topic in SC intonation studies.

REFERENCES

[1]Wu,Z.J.(1982), "Rules of intonation in Standard Chinese", *Reprints of Papers for the Working Group on Intonation,* 13th Inter. Cong. Ling. pp.96-108, Tokyo

[2]Wu,Z.J.(1985), "Rules of tri-syllabic tone-sandhi in Standard Chinese", *Bull. of Chinese Ling.*, vol.2, pp.70-92. (in Chinese)

[3]Wu,Z.J.(1990), "Can poly-syllabic tone-sandhi patterns be the invariant units of intonation in spoken Standard Chinese?", *Proc. 1990 Int. Conf. on Spoken Language Processing,* Kobe, section: 12.10.1.

[4]Wu,Z.J.(1993), "A new method of intonation analysis for Standard Chinese: Frequency transposition processing of phrasal contours in a sentence", *Rep. Phon. Res., Inst. of Linguistics, CASS,* 1992-1993, PP.1-18.

[5]Wu,Z.J.(1994), "Further experiments on spatial distribution of phrasal contours under different range registers in Chinese intonation", *Proc. Int. Symp. on Prosody,* Yokohama, pp.65-73.

[6]Yang,S.A.(1994), The Chinese speech synthesis technique oriented acoustic-phonetics, *Document Press of Social Sciences,* Beijing. pp.11-13.

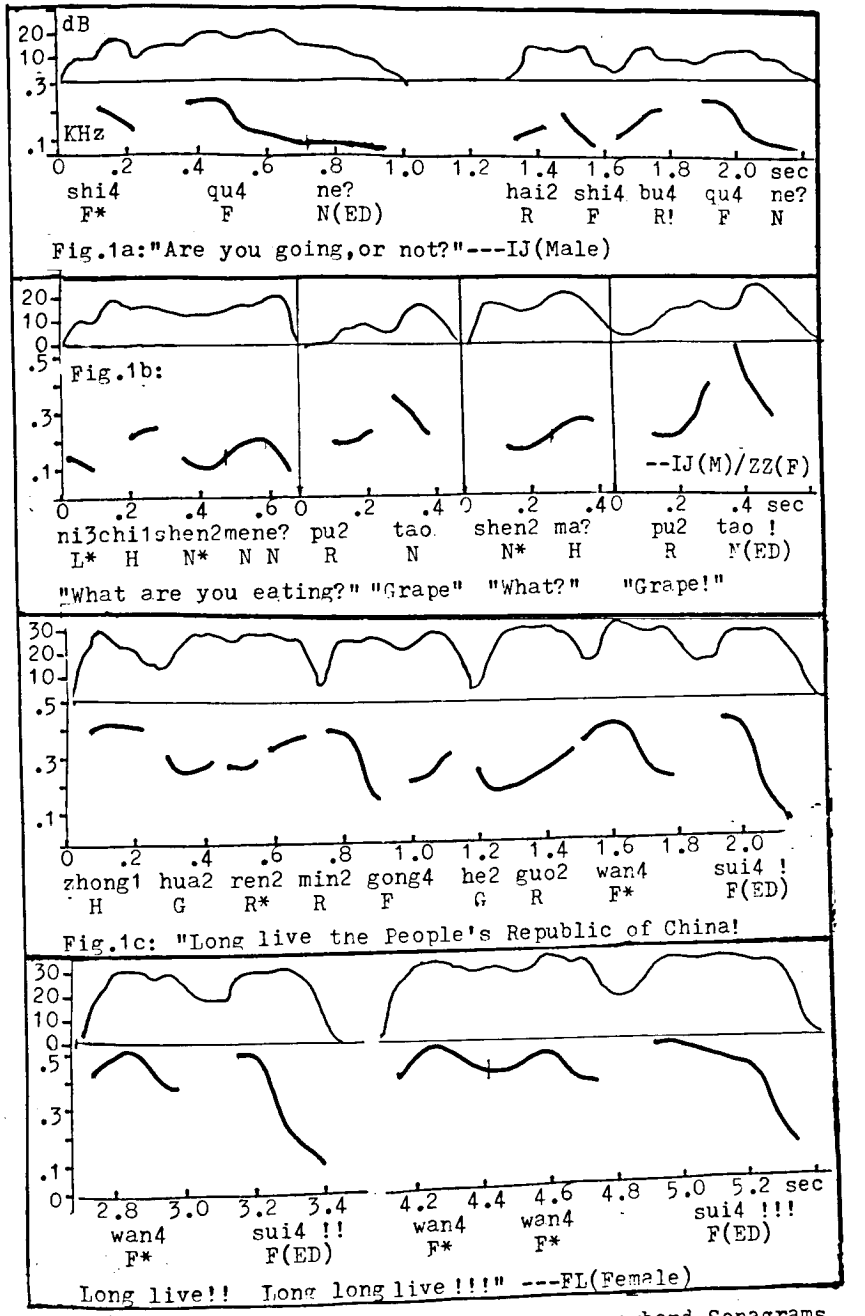


Fig.1: Intonation contours traced from narrowband Sonagrams. (a) A yes/no question sentence. (b) Question/answer between two persons. (c) A cheer in the National Day Celebration.

Effects of Phrasal Length and Time Distance between Peaks on Peak Height in Mexican Spanish

*Pilar Prieto, Holly Nibert *and Chilin Shih AT&T Bell Laboratories*

Abstract

A speaker of Mexican Spanish read a total of 809 declarative utterances of varying length (from 2 to 5 pitch accents) and of varying distance between pitch accents (from 1 to 3 intervening unstressed syllables). The resulting contour can be described as a series of simple peaks or H* pitch accents, following [1], where each peak is lower than the one preceding it. An analysis of the data indicates that neither phrasal length nor distance between adjacent pitch accents have an effect on peak height. Rather, the height of a given peak is determined by its position within the utterance and can be predicted quite successfully by reducing the previous peak's pitch value by a constant *downstep ratio*.

INTRODUCTION

The purpose of the present experiment is to examine two controversial questions regarding the behavior of pitch downtrend (i.e., the general lowering of pitch over the course of a phrase) in Mexican Spanish: a) what is the effect of time distance between peaks on peak height? b) what is the effect of phrasal length on peak height?

The first question concerns the nature of downtrend: is it governed by a gradual time-dependent *declination* effect, by a ratio- driven *downstep* effect, or by both? The former type of effect, which has been claimed to occur in many languages (e.g., [2] for Japanese, [3] for Danish, [4] for English, among many others), predicts that the *length of time interval* between two peaks in a fundamental frequency contour will affect the pitch level of the second peak: the greater the distance between the two peaks, the lower the F_0 value of the second peak. Yet, studies like [5] found that adjacent F_0 peaks in English descending contours displayed an *invariant* ratio of decay, regardless of the distance between them. In our experiment, we will examine whether the time-dependent declination effect is a necessary component in pitch downtrend modelling.

The second issue investigated by the present experiment is the effect of phrasal length (measured by the number of pitch accents in an utterance) on the scaling of initial peaks. Recent instrumental studies make contradictory claims about the relationship between the length of an utterance and the height of the phrase-initial peak. While authors like [6] report a significant increase in the height of the first peak in longer utterances, other authors find that peak values are more or less constant in a given position, regardless of utterance length, see [5], [7], [8], [9].

EXPERIMENTAL DESIGN

The database consisted of 120 declarative *listing* phrases, all comprised of a noun phrase modified by an increasing number of embedded noun or prepositional phrases. They were obtained by exhaustively combining phrases with 2 to 5 pitch accents with 1 to 3 intervening unstressed syllables. The three following utterances correspond to the 2-accent group (2 accents with 1, 2, and 3 unstressed syllables in between): 1) *ráyo lúna*; 2) *ráyo de lúna*; 3) *ráyo de la lúna*. To facilitate the comparison between peaks in the same position, the ordering of the content words has been kept constant in all of the utterances. Thus, the structural characteristics of the database make possible a strict analysis of both the distance between pitch accents and phrasal length on pitch height.

A male speaker of Mexican Spanish read the 120 utterances at least five times in random order, for a total of 809 sentences. The speaker was trained to read the sentences at a normal speech rate using a descending pitch pattern. Figure 1 illustrates a typical utterance with five pitch accents together with our labeling scheme. In accordance with [1] the exhibited contour could be characterized as a series of simple peaks, or H* accents. Each peak is downstepped and is lower than the one preceding it.

*University of Illinois at Urbana-Champaign

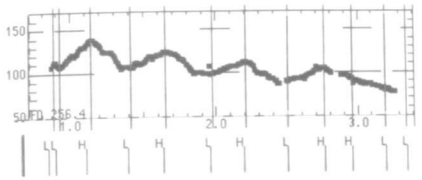


Figure 1: Five-accent utterance

F_0 peaks and valleys of each utterance were manually labelled using the Waves speech analysis package of Entropic Inc. The points labeled H in Figure 1 (the highest absolute F_0 values for each accent) constitute the target data in the present study. The utterance-final pitch accent was excluded from the analysis, since it did not display an F_0 peak in our data. The contour in Figure 1 is a 5 pitch-accent sentence with only four clear peaks.

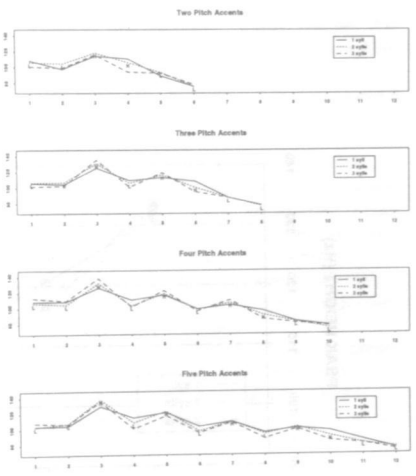


Figure 2: Schematized F_0 contours

RESULTS

Before the analysis, we checked that the utterances were produced over a more or less constant pitch range, since pitch range variation is an additional factor that could confound the effects of the factors under study. In our data, pitch range (measured as the distance in Hz from the lowest F_0 point to the following peak) was rather stable for a given phrasal position (which, of course, decreased steadily over the course of an utterance).

Effects of Phrasal Length

Figure 2 shows the schematized mean F_0 contours of utterances of different lengths (from 2 to 5 pitch accents). The height of the first peak is nearly constant in the 3 to 5-accent cases (around 138 Hz). The first peak of a 2-accent utterance, however, is much lower. For each utterance length, there are a set of three mean contours that show a steady increase in the number of unstressed syllables between accents (1 to 3). Solid lines represent a distance of 1 unstressed syllable between accents, dotted lines correspond to 2 unstressed syllables, and dashed lines show 3 unstressed syllables. The peak heights are not affected by the varying number of intervening unstressed syllables.

In general, our data do not exhibit phrasal length effects on peak height, leaving aside the first peak of 2-accent utterances (which is not included in further data analyses). Therefore, utterance length need not be of concern when we analyze the effects of the distance between pitch accents.

Effects of Distance between Peaks

Is a time-dependent declination effect present in the descending F_0 peaks produced by our speaker? If so, we would expect the second one of two adjacent peaks to be lower as the distance between the two increases. We measured this distance in two ways: in terms of number of unstressed syllables between accents (from 1 to 3), and in terms of real time.

Table 3 shows the mean distance in Hz between adjacent peaks, grouped into utterances of different lengths and number of intervening unstressed syllables. The results show that, for a given phrasal position, the distance in Hz between adjacent accents is nearly the same, except in one case: the data seem to show that the one-syllable condition triggers less peak decay. This pattern is also observed in Figure 2, which plots the mean absolute values of four successive F_0 peaks, using utterances with one, two, or three unstressed syllables before the target peak. T-tests comparing the peaks in the three conditions (1, 2, 3 preceding unstressed syllables) show that while peaks in the two and three-syllable conditions belong to the same group, peaks in the one-syllable condition are significantly different (at $p. < 0.01$).

We claim that the effect triggered by the

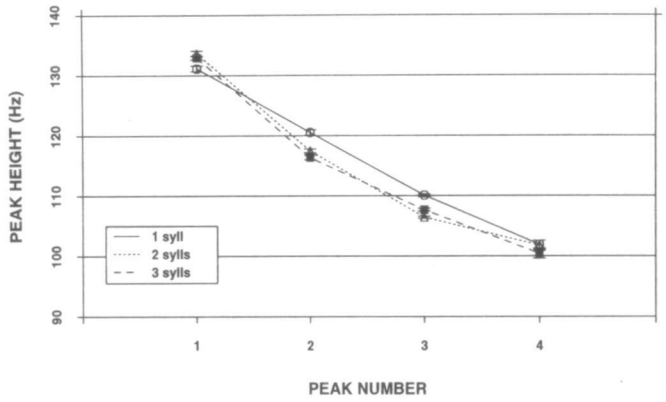


Figure 3: Mean adjacent peaks corresponding to an increase in the number (1 to 3) of unstressed syllables between pitch accents. Solid lines link the peaks separated by 1 unstressed syllable, dotted lines link those separated by 2 unstressed syllables, and dashed lines link those separated by 3 unstressed syllables.

Table 1: **Mean absolute peak didfference in Hz between adjacent peaks with differing number of intervening unstressed syllables (1 to 3) in the position indicated.**

Number of accents	Interv Sylls	Pos. 1-2	Pos. 2-3	Pos. 3-4
3	1	9.33		
	2	14.11		
	3	16.10		
4	1	7.62	10.03	
	2	14.55	14.80	
	3	14.93	15.60	
5	1	7.93	8.10	5.84
	2	15.46	13.25	8.82
	3	16.29	14.03	10.47

one-syllable condition is attributable to a slight pitch reset created by a syntactic boundary separating two simple NPs (the one-syllable condition is the only one with this syntactic configuration). This phenomenon has been reported for other languages (see [3] Danish), and we will leave this issue for further investigation.

The clearest evidence for the lack of a time-dependent declination effect on peak height can be seen by measuring the distance between the peaks in terms of real time. We performed a within-group correlation analysis between the time interval (in ms) between two peaks and their difference in Hz. Surprisingly, if anything, there was a small tendency to increase the height of the second peak as its distance to the previous peak increases (the negative correlation coefficients were highly significant (at $p < 0.01$)).

CONCLUSION

The analysis of the data indicates that neither phrasal length nor distance between adjacent pitch accents have an effect on peak height in the downstepped F_0 contours produced by our Mexican speaker. Rather, the height of a peak is determined by its position within the utterance: for example, the second H peak in a sequence will show a more or less constant pitch value, regardless of whether it appears in a three, four, or five pitch accent sentence or whether it is separated from the preceding accent by 1, 2, or 3 unstressed syllables. Further calculations show that peak height is predicted quite successfully by a constant reduction (also called *downstep ratio*) of the previous peak's pitch value. Thus, our results point to a lack of time-dependent declination effect and provide evidence against the view that global declination is an automatic and universally pitch mechanism [2], [6], [10], [11].

REFERENCES

[1] Pierrehumbert, J. (1980) *The Phonology and Phonetics of English Intonation.* Ph. D. Dissertation, MIT.

[2] Fujisaki, H. (1983) Dynamic Characteristics of Voice Fundamental Frequency in Speech and Singing. In *The Production of Speech* (P.F. MacNeilage, editor), pp. 39-55. New York and Berlin: Springer-Verlag.

[3] Thorsen, N. (1980) Intonation Contours and Stress Group Patterns in Declarative Sentences of Varying Lengt. n ASC Danish. *Annual Report of the Instit e of Phonetics*, University of Copenhagen 1 1-29.

[4] Liberman, M. (197 , The Intonational System of English. Ph. D. Dissertation, MIT. Cambridge, Massachussets.

[5] Liberman, M. & Pierrehumbert, J. (1984) Intonational invariance under changes in pitch range and length. In *Language Sound Structure* (M. Aronoff & R. Oehrle, editors), pp. 157-233. Cambridge: MIT Press.

[6] Cooper, W. E. & Sorensen, J.M. (1981) *Fundamental frequency in sentence production.* Heidelberg: Springer.

[7] Thorsen, N. (1981) Intonation Contours and Stress Group Patterns in Declarative Sentences of Varying Length in ASC Danish. — Supplementary Data. *Annual Report of the Institute of Phonetics*, University of Copenhagen 15, 13-47.

[8] Sternberg, S., Wright, C.E., Knoll, R.L. & Monsell, S. (1980) Motor programs in rapid speech: additional evidence. In *Perception and production in fluent speech* (R.A. Cole editor), pp 507-534. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

[9] van den Berg, R., Gussenhoven, C. and Rietveld, T. 1992. Downstep in Dutch: Implications for a model. In *Papers in Laboratory Phonology II: Segment, Gestures, Tone.* (G.J. Docherty and D.R. Ladd, editors), pp. 335-367. Cambridge: Cambridge University Press.

[10] Lieberman, P. (1967) *Intonation, perception, and language.* Cambridge: MIT Press.

[11] Fujisaki, H. (1988) A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour. In *Vocal Physiology: voice production, mechanisms and functions* (O. Fujimura, editor), pp. 347-355. New York: Raven.

A LONG-DISTANCE DEPENDENCY IN YORÙBÁ TONE REALIZATION

Yetunde O. Laniran, Northeastern University, Boston
G. N. Clements, CNRS, UA 1027, Paris

ABSTRACT

This paper reports an experimental investigation of the F_0 interpretation of tone sequences in Yorùbá, with special reference to the phenomena of downstep and H Raising. It examines a long-distance dependency between the F_0 realization of an initial H tone and the choice of a tone occurring several syllables away, and discusses its consequences for models of F_0 interpretation.

1. INTRODUCTION

Downstep is a pattern in which later tones are scaled lower than earlier tones of the same category within a given prosodic unit. Most current models of F_0 interpretation treat downstep either in terms of a rule which iterates from left to right across the phrase, computing the F_0 value of each new tonal target as a function of the immediately preceding F_0 target value [1], or in terms of a downward resetting of the overall register within which tones are realized [2,3]. We present evidence that these models are not able to account for the specific nature of downstep in Yorùbá, and suggest an alternative interpretation of F_0 downtrends in terms of regressive upstep (right-to-left register raising). The experimental corpus and methodology upon which this study is based are described in [4], to which the reader is referred for further details.

2. TONE IN YORUBA

Yorùbá, spoken in Nigeria, is a tone language with three lexically distinctive tone levels, H (high), M (mid), and L (low). We follow the usual practice of marking H tones with an acute accent and L tones with a grave accent, leaving M tones unmarked. Minimal pairs include ìlú (LH) 'city', ìlu (LM) 'perforator', and ìlù (LL) 'drum'. A regular rule of Tone Spread extends a L tone to the beginning of a following H tone syllable, causing e.g. 'city' to be realized with the high-rising pattern [ìlǔ]; similarly, underlying HL sequences are realized as high-falling [4,5].

Previous studies have elicited certain generalizations about the phonetic realization of surface tone sequences in Yorùbá [4,6], of which two are of special relevance here. *Downstep* occurs non-distinctively in Yorùbá, where it is triggered by HL sequences occurring early in the sentence; no other tone sequences are regular downstep triggers. *H Raising* causes H tones in HL sequences early in the sentence to be realized with higher values than H tones in other contexts.

Given the similarity of the contexts in which Downstep and H Raising occur, Connell and Ladd ([6]: 17-19) raise the question whether apparent Downstep effects on H tones in Yorùbá might be entirely reducible to H Raising; their data was unable to satisfactorily resolve this question. One of the specific concerns of this paper (as it was of the larger study from which it is drawn [4]) will be to determine whether H tones really do "step down" in HLHL sequences. A more general concern will be to sort out the relative contributions of Downstep and H Raising in tone realization, and to determine whether both principles are independently needed to account for phonetic generalizations.

3. EXPERIMENTAL DESIGN

In order to address these questions, we constructed a set of test sentences devised to allow us to examine the F0 values of H tones (henceforth called the "target H tones") in a variety of contexts. Each sentence contained a target H tone as its first and third tone. Two parameters were systematically varied: the tone following the first H tone and the tone following the second one. The test sequences can be schematized by the formula "HXHY", where X and Y are to be understood as variables taking H, M, and L tones as their values. These sequences were placed in a constant tonal frame to eliminate any final lowering effects. The sentences constructed in this way are given below:

Set A: HXHY, X=H
1. Déwálé rélá bọ́ fún Láyẹmí
2. Déwálé rolú bọ́ fún Láyẹmí
3. Déwálé ràwé bọ́ fún Láyẹmí

Set B: HXHY, X=M
4. Móyẹmí rélá bọ́ fún Láyẹmí
5. Móyẹmí rolú bọ́ fún Láyẹmí
6. Móyẹmí ràwé bọ́ fún Láyẹmí

Set C: HXHY, X=L
7. Máyọ̀mí rélá bọ́ fún Láyẹmí
8. Máyọ̀mí rolú bọ́ fún Láyẹmí
9. Máyọ̀mí ràwé bọ́ fún Láyẹmí

All sentences in sets A-C have identical syntactic structure, in which a proper-noun subject is followed by a serial verb predicate introduced by a contracted V+N sequence (e.g. *rélá* is the contraction of / ré / 'pick/ + / ilá / 'okra' created through a regular process of vowel elision). Glosses in each set are 'X (picked okra / bought mushrooms / bought books) for Layẹmi', respectively. The first five syllables contain only sonorants, to eliminate the perturbatory effects of obstruents [7]. The variable X of the HXHY schema is instantiated by a H tone in set A, a M tone in set B, and a L tone in set C, and the variable Y is instantiated by H, M, and L tones in that order in the three sentences of each set. Thus all possible instantiations of the HXHY schema are represented in the data.

These sentences were recorded as part of a larger experimental corpus by two native speakers of standard Yorùbá, TJ (a male in his thirties) and KG (a male in his twenties). Each subject read the full list of randomized, tonally-annotated sentences twelve times. Sentences judged to have been read with incorrect tones were discarded. The remaining sentences were then digitized at a sampling rate of 10 kHz using the Entropics Waves+ software on a SUN Workstation. After further elimination of sentences with incomplete F_0 tracks, ten tokens of each sentence from TJ and nine tokens by KG were retained for analysis.

These sentences are appropriate for testing the Downstep and H Raising rules described above. If speakers apply Downstep at HL junctures, we expect the second H tone (H_2) to be lower than the first (H_1) in the HLHY sentences of set C, but not in the others. If speakers apply H Raising, we expect H tones to be higher in HL sequences than they are in otherwise similar HH and HM sequences. (Furthermore, we expect H Raising to apply twice in the HLHL sequence of example 9, but at most once in other target sequences.)

A further expectation is that all the sentence-initial raised H tones of set C should have about the same F_0 values, regardless of the tonal composition of later portions of the sentence. In particular, the initial H in the HLHL sequence of example 9 should have the same value as the initial Hs in the HLHH or HLHM sequences of examples 7 and 8. This is because neither Downstep nor H Raising, as formulated in current models, access information from later, nonadjacent tones in the tonal string. In other words, we should find no anticipatory long-distance dependencies. Indeed, such non-local access would be theoretically prohibited by the principle of adjacency, which requires that rules can make reference only to elements that are adjacent in the representation [8], if we extend this principle to phonetics.

4. RESULTS

Figure 1 presents the results for TJ in column 1 and for KG in column 2. Graphs (a) and (d) present the data from Set A, graphs (b) and (e) from Set B, and graphs (c) and (f) from Set C. Each graph overlays the averaged F_0 tracks for the three sentences of the given set. Tracks labelled with circles show sentences in which Y=H, those labelled with lozenges show sentences in which Y=M, and those labelled with triangles show sentences in which Y=L. Two F_0 values are given for each syllable, one taken toward the beginning and one toward the end, as indicated on the x-axis.

In Figure 1, realizations of target H tones are assigned alphanumeric labels and boxed for ease of reference. Each H tone followed by a L tone is realized by three F_0 values, two coinciding with the syllable bearing the lexical H tone and one occurring at the beginning of the following L tone syllable, to which the H tone extends by Tone Spread (section 2). These H tone values are labelled a, b, and c. All other H tones are represented by only two values, those labelled a and b.

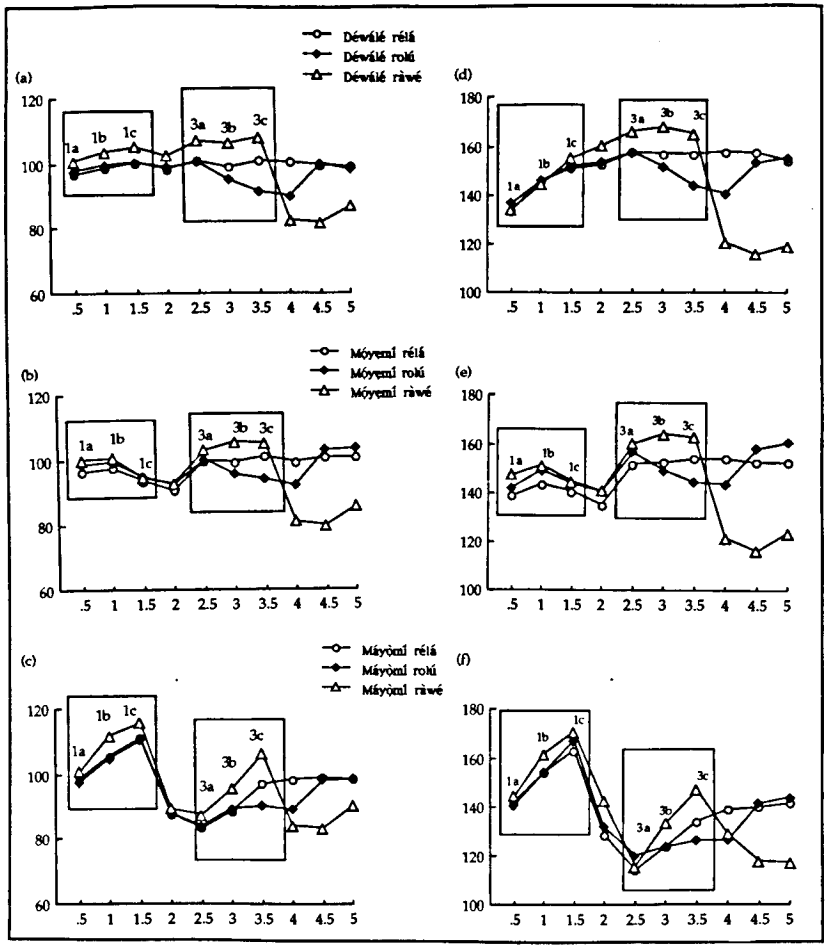


Figure 1. Overlay of F_0 contours of the sentences of sets A, B, and C.

A comparison of all graphs shows that our first two expectations are confirmed. First, we see that H_2 is lower than H_1 in the set C sentences (graphs (c, f)), but nowhere else; indeed, in the other sets H_2 is usually realized at values slightly higher than H_1. This result confirms that Downstep applies across HLH sequences, and cannot be reduced to H Raising.

A second result is that H Raising applies to H tones before L tones, also as expected. This is shown by a comparison of the H_1 values labelled 1a-c in the set C comparable H_1 values in the other graphs. It is also shown by a comparison of the H_2 tracks labelled with triangles in all graphs (representing sentences 3, 6, and 9, in which Y=L) with the other tracks. In all cases, these tracks reach higher F_0 values at points 3b and 3c than do the other tracks; these are, of course, the points that represent a H tone before a L tone.

In contrast, our third expectation is not confirmed, as is shown by an examination of the F_0 tracks in the bottom graphs, (c) and (f). Here the F_0 tracks labelled with

quence of sentence 9, display higher H_1 values than the others (points 1a-c). In other words, the selection of L as the value of Y in the sequence HLHY correlates with higher values of H_1.

As the F_0 differences here are somewhat smaller than the previous ones, we performed a two-way ANOVA to detect possible interactions between tones X and Y and the H_1 values observed at points 1a, 1b, and 1c. The results are shown in Table 1.

	TJ	KG
P-values: point 1a		
X	n.s	n.s
Y	.0140	.0191
X*Y	n.s	n.s
P-values: point 1b		
X	.0001	n.s
Y	.0005	n.s
X*Y	n.s	n.s
P-values: point 1c		
X	.0001	.0001
Y	.0065	n.s
X*Y	n.s	n.s

Table 1. ANOVA results.

P-values under .05 are taken as showing that the null hypothesis—that there was no effect of the identity of tones X and Y on F_0 values at points 1a,b,c—should be rejected. We see that the identity of tone X significantly affected the value of H_1 at point 1c for both speakers, and also at point 1b for TJ, confirming the effect of H Raising. Moreover, it shows that the identity of tone Y significantly affected the value of tone H_1 at point 1a for both speakers, and also at points 1b and 1c for TJ. In other words, the choice Y=L is significantly correlated with the raising of H_1. (There was no significant interaction between tones X and Y.)[1]

5. DISCUSSION

As mentioned earlier, current models of downstep predict that the F_0 value of a sentence-initial raised H tone in sentences like those of set C cannot be affected by the choice of nonadjacent tones to its right. In such models, the anticipatory long-distance dependency observed in sentence 9 can be accounted for only by making explicit reference to nonadjacent tones. Such an analysis would violate the principle of adjacency, and would be quite arbitrary.

This result suggests that "Downstep" and "H Raising" in Yorùbá may be artefacts of an inadequate F_0 interpretation model. In Laniran's 1992 dissertation [4: chs. 5, 6], an alternative view is proposed in which both phenomena are viewed as manifestations of *regressive upstep*. In this view, HL sequences do not trigger downstep (register lowering) to their right, but upstep (register raising) to their left. Since upstep is cumulative, the precessive H tones in a HLHL sequence will be assigned higher values as we scan from right to left. This model has the advantage of accounting for downtrends across HLHL sequences and H Raising before L in terms of a single mechanism, and directly explains the otherwise coincidental fact that both apply only in HL sequences.

ACKNOWLEDGEMENT

We thank John Kingston for helpful guidance on several aspects of this study.

REFERENCES

[1] Pierrehumbert, J. (1980) The Phonology and Phonetics of English Intonation, unpublished Ph.D. dissertation, MIT, Cambridge, Ma.

[2] Clements, G.N. (1983) "The Hierarchical Representation of Tone Features", in I.R. Dihoff, ed., *Current Approaches to African Linguistics* vol. I, Foris Publications, Dordrecht, 145-176.

[3] Inkelas, S. and W.R. Leben (1990) "Where Phonology and Phonetics Intersect: the Case of Hausa Intonation," in J. Kingston and M. Beckman, eds, *Papers in Laboratory Phonology 1*, CUP, Cambridge, 17-35.

[4] Laniran, Y.O. (1992) Intonation in Tone Languages: the Phonetic Implementation of Tones in Yorùbá. Ph.D. dissertation, Cornell University.

[5] Ward, I.C. (1952) *An Introduction to the Yorùbá Language*, CUP, Cambridge.

[6] Connell, B. and D.R. Ladd (1990) "Aspects of Pitch Realization in Yoruba," *Phonology Yearbook* 7, 1-30

[7] Hombert, J.-M (1976) "Consonant Types, Vowel Height, and Tone in Yorùbá," *UCLA Working Papers in Phonetics* 33, Los Angeles, 40-54.

[8] Odden, D. (1994) "Adjacency Parameters in Phonology," *Lg* 70, 289-330.

[1] A third speaker examined in the fuller study on which this paper is based [4] showed trends similar to those of TJ and KG, but which failed to reach statistical significance.

SPEAKER AND LISTENER SEX FOR SPEAKER HEIGHT AND WEIGHT IDENTIFICATION

Wim A. van Dommelen

Department of Linguistics, Trondheim, Norway

ABSTRACT

This study examines the ability of listeners to judge speaker height and weight from speech samples. The results show that mainly male listeners were able to estimate male speaker height and weight. Neither male nor female listeners could judge female speaker height or weight. The data suggest that the listeners correctly used speech rate information in judging males. But low F0 and formant frequency values were wrongly taken to indicate large body dimensions.

INTRODUCTION

This investigation takes a look at the impression of speaker height and weight conveyed by the voice of an unknown speaker. The first question is whether or not such impressions are correct. Though previous investigations do not seem to give definite answers to this question, existing data indicate that listeners are very consistent when estimating speaker height or weight [1, 2]. This gives rise to the question what kind of information contained in the speech signal is used by the listener and in what way. One potentially important factor could be speaking fundamental frequency (F0). F0 acoustic scale values varying from low to high presumably correspond to perceptual scale values varying from "tall speaker" to "small speaker" [3].

In order to shed light on the issues mentioned above, for the present study listener judgements on speaker height and weight were collected. To see whether these judgements were correct, estimated height/weight was compared with actual speaker height/weight. The listeners' use of information contained in the speech signal was investigated by relating the height/weight estimates to three different parameters: F0, formant frequency and speech rate.

EXPERIMENTAL PROCEDURE

Recordings

Speech recordings were made in the studio of the Linguistics Department of Trondheim University from a group of speakers consisting of 30 volunteers (15 males and 15 females), who were mainly recruted from the university's student population. Their ages ranged between 23 and 32 years for the males (with an average of $\bar{x}$= 26.0; s= 2.8) and between 20 and 36 years ($\bar{x}$= 25.5; s= 3.8) for the females, respectively. For the occasion of the recordings, the speakers' height and weight were measured with an accuracy of 1cm and 0.5kg, respectively, using a Lindeltronic 4000 from the Department of Sport Sciences.

The first part of the speech material that was recorded consisted of a list of 10 isolated Norwegian words. Secondly, the first two paragraphs of a Norwegian fairytale were recorded. The text was of an epic character without direct speech and was slightly modified to make its style more up-to-date. The speakers were instructed to read both words and text using their natural voices.

Analysis

Analyses of the speech recordings were performed with Signalyze [4] for each of the parts that were used in the listening tests: ten isolated words and each of the two text paragraphs. For each of the 30 speakers, the following signal parameters were investigated:

(a) average F0 of the voiced signal portions. Gross errors in the analysis were corrected by hand by deleting the actual frames. Given the three text conditions, this procedure resulted in three average F0 values per speaker.

(b) F2-frequency of the neutral vowel [ə], which occurred in final position in two isolated words and in four (paragraph 1) and seven (paragraph 2) words, respectively, from the fairy tale. Taking the relation v= f*λ (where v= velocity of sound in air, f= frequency and λ= wavelength) as a point of departure,vocal tract length l was estimated according to the formula l= 3/4*v/F2. Subsequently, average estimates were calculated based on n= 2, 4, and 7 measurements, respectively, for the three conditions. Henceforth, the estimates of vocal tract length will be abbreviated VT.

(c) speech rate, which was defined differently for the two text sorts. As a result of preliminary heuristic testing, for the isolated word condition the total duration of the ten words was taken as a measure, leaving out intermediate pauses. However, since pause durations in the texts obviously were highly speaker-specific, overall text durations were taken as a speech rate measure for paragraphs 1 and 2. In only a few cases, the total duration had to be corrected for speech errors. In the description, this measure will be referred to as Duration.

Listening tests

The speech material described above was used to prepare a set of 12 listening tapes, six for the groups of male and female speakers each (cf. Table 1). In preparing tapes 1-6 the original recordings were dubbed onto a second cassette; for tapes 7-12 the signals were band-pass filtered between 250Hz and 3.15kHz. All tapes were constructed according to the following pattern: each stimulus, i.e. the ten words, paragraph 1 or 2, was preceded by a voice announcing the next speaker number, and followed by a ca. 10 sec answering pause. Each tape contained each of the 15 male/female speakers once, each time with a different randomization. This means that tapes 7-12 contained the same speech material as tapes 1-6, the difference lying in randomization order and filtering condition. The main goal of presenting the same material also band-pass filtered was to collect more data, in order to be able to investigate the listeners' consistency.

Table 1. Presentation order of the 12 listening tapes. Each tape contained randomized speech samples from 15 speakers.

	unfiltered	filtered
words	(1) males	(7) males
	(2) females	(8) females
paragr. 1	(3) males	(9) males
	(4) females	(10) females
paragr. 2	(5) males	(11) males
	(6) females	(12) females

A total of 20 listeners (10 males, 10 females) were asked to estimate the speakers' height and weight using prepared answering sheets with height/weight scales for each speaker. The tapes were presented over headphones in two separate sessions (tapes 1-6 and 7-12).

RESULTS

Voice and body characteristics

First of all, it was investigated whether the acoustic cues analyzed here vary systematically with actual speaker height and weight. The results of a multiple regression analysis for the male speakers showed that neither average F0 nor vocal tract length estimate VT are reliable predictors of speaker height and weight. A remarkable exception in this connection is that heavier men had a tendency to speak at slower speech rates: all three correlations between male body weight and speech rate under the different text conditions Words, Paragraph 1 and 2 were statistically significant (r= 0.592, p= 0.024; r= 0.762, p= 0.001 and r= 0.712, p= 0.004, resp.). Male body height, however, does not correlate with speech rate, in spite of a general positive correlation between height and weight. From this, it can be concluded that it is the surplus of weight going beyond the normal weight increase due to greater height that is responsible for the lower speech rate.

As far as the female speakers are concerned, only one out of 18 correlations between height/weight and acoustic cues reached statistical significance (VT for height in Paragraph 2; r= 0.516, p= 0.048). The lack of a significant correlation between female weight and speech rate reveals an interesting sex-specific difference in speech production.

Estimation of height and weight

This paragraph deals with the question whether the listeners were able to estimate the speakers' height and weight from the speech samples. The data concerning the correlation between estimated and actual height and weight are presented in Table 2. For the male speakers, a considerable number (14) of positive correlations between estimated and actual height/weight were found. Eleven out of these 14 significant correlations go back on the male listeners, indicating that it was mainly this listener group that had accurate ideas concerning male height and weight. This in contrast with the far less clear results for the group of female

Table 2
Correlation coefficients and probabilities for correlations between estimated speaker height/weight and actual height/weight for two filtering conditions (unfiltered and band-pass filtered) and text conditions words, paragraph 1 and paragraph 2. 10 male (a) and 10 female (b) listeners. Only values statistically significant at a 5% level of probability are given. None of the correlations for the tests involving female speakers reached statistical significance.

(a) Male listeners - male speakers

		Words		Paragraph 1		Paragraph 2	
unfiltered	height	r= 0.535	p= 0.040	r= 0.526	p= 0.044	r= 0.578	p= 0.024
	weight	r= 0.591	p= 0.020	r= 0.724	p= 0.002	r= 0.840	p= 0.000
filtered	height	r= 0.530	p= 0.042	———	———	r= 0.609	p= 0.016
	weight	r= 0.865	p= 0.000	r= 0.722	p= 0.002	r= 0.790	p= 0.000

(b) Female listeners - male speakers

		Words		Paragraph 1		Paragraph 2	
unfiltered	height	———	———	r= 0.570	p= 0.027	———	———
	weight	———	———	r= 0.627	p= 0.012	———	———
filtered	height	———	———	———	———	———	———
	weight	r= 0.687	p= 0.005	———	———	———	———

listeners (compare Table 2a with 2b).

In addition to this sex-specific behaviour on the part of the listeners, a remarkable sex-related speaker factor is found: None of the correlations between estimated and actual female speaker height or weight turned out to be significant - neither for male nor for female listeners. This finding indicates that speech production differs significantly between the two groups of speakers and, in all probability, is sex-specific.

Use of acoustic cues

To shed light on the question of what kind of information the listeners used to arrive at their judgements, multiple regression analyses were performed involving F0, VT and Duration as predictor variables and estimated height/weight as dependent variables. Table 3 presents the results for male speaker height/weight as estimated by males and females. As can be seen from the table, mean F0 correlated negatively with both estimated height and weight (17 out of 24 cases significant). So, a low F0 was taken as an indication of a tall, heavy speaker, whereas high F0 values pointed to small bodily dimensions. Similarly, the data suggest that low formant frequency values (high VT values) were associated with large body dimensions. A long text duration, i.e. a slow speech rate, gave rise to the impression of a tall and (especially) heavy speaker. All the tendencies noted here are stronger for the male than for the female listeners (compare Table 3a with 3b).

Table 3
Number of times the pos(itive) or neg(ative) correlations between estimated male speaker height/weight (EMH/EMW) and F0, VT, and Duration reached statistical significance (p< 0,05%). max= 6.

a) male listeners

	F0	VT	Duration
EMH	6 neg	2 pos	1 pos
EMW	5 neg	5 pos	4 pos

b) female listeners

	F0	VT	Duration
EMH	3 neg	0	1 pos
EMW	3 neg	3 pos	2 pos

Comparison of the data from Table 3 with the corresponding ones for female speaker height/weight (Table 4) reveals an effect of speaker sex also with respect to the use of acoustic cues. Only male

listeners show a weak tendency to associate low F0 values with greater height/weight (3 out of 12 cases significant). VT, however, correlates more often with perceived height/weight (9 out of 24 cases significant). In strong contrast with the results for male speakers, varying speech rate appears to have no influence on the height/weight ratings, neither for men nor for women.

Table 4
Number of times the pos(itive) or neg(ative) correlations between estimated female speaker height/weight (EFH/EFW) and F0, VT, and Duration reached statistical significance (p< 0,05%). max= 6.

a) male listeners

	F0	VT	Duration
EFH	1 neg	4 pos	0
EFW	2 neg	2 pos	0

b) female listeners

	F0	VT	Duration
EFH	0	2 pos	0
EFW	0	1 pos	0

DISCUSSION

In line with previous investigations [5, 6], the present correlations between mean speaking F0 and speaker height/weight were statistically nonsignificant. Estimated vocal tract length, too, did not vary systematically with body dimensions. Since mean F0 and formant frequencies are largely dependent on the dimensions of the laryngeal and supralaryngeal parts of the speech mechanism, this outcome suggests that these dimensions vary independently of the rest of the human body.

An unexpected result was the observed tendency for male speakers to have reduced speech rate with increasing body weight. This raises the question whether this phenomenon is due to biological or to socio-cultural factors. At present, it is also an open question why the women in this study behaved differently in this respect.

Interestingly, the sex-specific behaviour in speech production is also reflected in the identification of speaker height/weight in that the listeners' ability to estimate these dimensions was confined to the group of male speakers. It was mainly the group of male listeners who were able to do so. This suggests that this group exploited the typically male speech rate behaviour.

The results have shown that speech rate is not the only source of information used by the listener. The final impression of the speakers' body characteristics on the listener turned out to be multi-facetted. Obviously, also F0 and spectral information are important factors. Differently from the correct use of speech rate for estimating male speaker weight, F0 and spectral parameters were exploited inappropriately. Low F0 as well as low low formant frequency values were taken to indicate large body dimensions. High values were interpreted as originating from a small, light speaker. The incorrect use of these parameters, however, had no crucial impact on the (correct) use of speech rate information. In all probability, the reason for this must be sought in the fact that there was no correlation between speech rate on the one hand and mean F0 or formant frequencies on the other. This means that possible modifications of speech rate based judgements were stochastical, rather than systematic.

References

[1] Lass, N.J., Phillips, J.K. & Bruchey, C.A. (1980), "The effect of filtered speech on speaker height and weight identification", *Journal of Phonetics*, vol. 8, pp. 91-100

[2] Dommelen, W.A. van (1993), "Speaker height and weight identification: a re-evaluation of some old data", *Journal of Phonetics*, vol. 21, pp. 337-341

[3] Ohala, J.J. (1984), "An ethological perspective on common cross-language utilization of F0 of voice", *Phonetica*, vol. 41, pp. 1-16.

[4] Keller, E. (1993), *Signalyze™. Signal Analysis for Speech and Sound.* Network Technology Corporation, Charlestown.

[5] Hollien, H., Green, R., and Massey, K. (1994), "Longitudinal research on adolescent voice change in males", *Journal of the Acoustical Society of America*, vol. 96, pp. 2646-2654.

[6] Künzel, H.J. (1989), "How well does average fundamental frequency correlate with speaker height and weight?", *Phonetica*, vol. 46, pp. 117-125.

SPEAKER CHARACTERISTICS IN THE COARTICULATION OF THREE DUTCH VOWELS /a,i,u/

H. van den Heuvel, B. Cranen & T. Rietveld
Dept. of Language and Speech, University of Nijmegen, The Netherlands
E-mail: heuvel@let.kun.nl

ABSTRACT

We investigated the coarticulation in the first three formants of the Dutch vowels /a,i,u/, and the speaker variability in this coarticulation. We found the largest amount of coarticulation in the vowel /u/, somewhat less in /a/, and hardly any in /i/. The amount of coarticulation as a function of context turned out to be speaker-dependent for /a/ and /u/. However, as for our data coarticulation proved not to be an important parameter for speaker identification by computer.

INTRODUCTION

Speaker variability in articulation patterns has not been investigated on a large scale, because the invariant aspects of speech production have been considered more interesting from a linguistic point of view. Nonetheless a few studies were published dealing with this topic [1,2]. In the light of these studies it can be hypothesised that coarticulation may exhibit substantial between-speaker variability.

For some phonemes a confirmation for this hypothesis has been found in the field of (automatic) speaker identification. Su, Li & Fu (1974) [3] noted that the amount of coarticulation in nasals (and especially in /m/) varies highly among speakers and can, as a result, effectively be used in automatic speaker identification. Comparable experiments are reported for /l/ and /r/ by Nolan (1983) [4]. In his study the coarticulation in /l/ appeared more speaker-specific than that in /r/. Similar experiments for vowels have not been carried out so far, which is surprising since the coarticulation in vowels as such has been studied extensively.

The main aim of this paper is to ascertain firstly whether coarticulation in vowels is speaker-specific, and secondly to investigate if it can be beneficially used in a speaker identification task. Before we examine this, we will first have a look at the speech data used and at the coarticulation that was observed in /a,i,u/.

SPEECH DATA

In order to keep the experiment within practical limits and to have control over the number of factors that may effect the vowel formants, we opted for a rather restricted dataset. The data set used consisted of 24 /CVCə/ (mainly) pseudo-words spoken in isolation. The three nucleus vowels used were /a,i,u/ and the eight consonants, which appeared once as C_1 and once as C_2 for each vowel /a,i,u/, were /p,t,k,d,s,m,n,r/. See table 1.

The 24 words were printed in a random order on ten 30-word word lists, in a, for Dutch, plausible spelling. All ten word lists were read out by each speaker in one recording session. The initial three words served as fillers, as did the final three. In this way (24 words · 10 repetitions =) 240 words were obtained for every speaker. Since fifteen speakers participated in the experiment a total of 3600 word tokens were collected. The subjects were all male native speakers of Dutch and aged between 20 and 30 years.

Table 1: The /CVCə/ pseudo-words used in the experiment, in phonemic representation.

V =	/a/	/i/	/u/
C_1 = /p/	/pasə/	/pinə/	/pudə/
C_1 = /t/	/tanə/	/tirə/	/tumə/
C_1 = /k/	/kadə/	/kimə/	/kunə/
C_1 = /d/	/dakə/	/disə/	/dupə/
C_1 = /s/	/sapə/	/sidə/	/surə/
C_1 = /m/	/marə/	/mikə/	/mutə/
C_1 = /n/	/natə/	/nipə/	/nukə/
C_1 = /r/	/ramə/	/ritə/	/rusə/

The speech data were digitised with a 12-bit AD-converter at a sampling frequency of 16 kHz. Each word was segmented into phoneme-sized units; the nucleus vowel (i.e. /a,i,u/) was additionally split up into a steady-state part flanked by two transitions.

The formants were calculated by means of an LPC-analysis (pitch-asynchronous autocorrelation method; window length 25 ms; frame shift 5 ms; order 20). Root solving was used to obtain the formant values (in Hz). The target formants were extracted from the middle frame of the steady-state of each vowel token (/a,i,u/). Only the first three formants F_{1-3} were used. The formant positions of F_{1-3} in all selected vowel middle frames were checked by hand. They were converted to Barks to prevent that the variations in the higher formants F_2 and F_3 would obtain a dominating weight.

MEASURING THE COARTICULATION IN /a,i,u/

In our score model the coarticulation (COART) in a formant i in a specific context c in replication r as realised by speaker s is given by

$$COART(s,c,r,i) = (f_{scr}(i) - f_{s(ref).}(i))^2,$$

where $f_{scr}(i)$ refers to a raw formant value (obtained from the middle frame of a vowel token) and $f_{s(ref).}(i)$ to the (speaker-dependent) reference value of the vowel formant.

The best reference for coarticulation is the vowel spoken in isolation or in a /hVt/-context [5]. However, our speakers did not produce /a,i,u/ in isolation nor in a /hVt/-context. In order to obtain good estimates of these formant values for our experiment, we used the vowel formant values for /a,i,u/ as published in [5:1094] as initial references. Next we used each context of a vowel as reference and looked which context resulted in COART-values with the closest match to the COART-values resulting from the values given in [5]. These contexts were selected as the ultimate references. It were /pasə/ for /a/; /ritə/ for /i/; /surə/ for /u/.

An ANOVA on the (3 vowels · 8 contexts =) 24 COART-values of the three vowels, followed by a Tukey HSD post-hoc comparison ($\alpha = .05$), showed that the COART-values for /u/ were significantly higher than for /a,i/. Other ANOVAs made clear that significant between-context differences in coarticulation were present only in /a/ and /u/, but most in /u/. The smallest COART-values and between-context differences were observed for /i/.

SPEAKER VARIABILITY IN COART

In the previous section we encountered significant differences in coarticulation between vowel contexts of /a/ and, especially, /u/. We may ask whether the same pattern of between-context differences is observed for all speakers, or whether the pattern is speaker-dependent. If the latter is the

case, we may conclude that coarticulation is speaker-specific.

A set of ANOVAs was carried out to answer this question. The ANOVAs were performed on the COART-values of each vowel /a,i,u/, separately. COART was computed for individual speakers, contexts and replications, averaging over formants:

$$COART(s,c,r) = \frac{1}{3}\sum_{i=1}^{3} COART(s,c,r,i).$$

Factors Speaker (fifteen levels) and Context (eight levels) were crossed in the ANOVAs. The interaction CxS proved to be significant ($p < .001$) for /a/ ($F_{98,1080} = 3.21$) and /u/ ($F_{98,1080} = 8.28$), but not for /i/ ($F_{98,1080} = 1.64$). This demonstrates that, indeed, the pattern of coarticulation over the contexts is speaker-dependent for /a,u/, but most for /u/. The speaker variability in the coarticulation of /u/ is shown in figure 1.

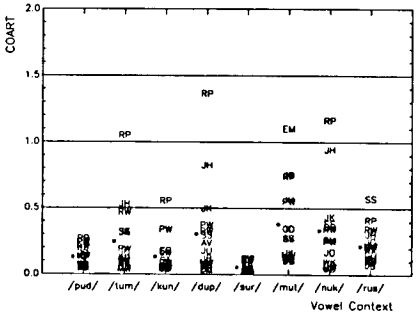


Figure 1: The speaker distribution of COART for each context of /u/. The speaker means are denoted by the speakers' initials; the context's mean is denoted by a black circle, slightly shifted to the left.

The figure illustrates that, indeed, speakers do not coarticulate uniformly. The most salient observation is that the mean COART-values in contexts with alveolars in C_1-position (/nuk/, /dup/ and /tum/) are pushed up due to the behaviour of two speakers: JH and, in particular, RP. Nonetheless, the interaction CxS remains significant if the data of these two speakers are removed from the ANOVA for /u/ ($F_{84,936} = 8.34$, $p < .001$).

SPEAKER IDENTIFICATION USING COART

Our results have indicated that coarticulation in the vowels /a/ and /u/ of our data set (as expressed by COART) is speaker-specific. This suggests that the amount of coarticulation may be used for speaker identification. The question that we tested was: do speaker identification scores improve if COART is used as an additional parameter to F_{1-3}? This is, of course, only a sensible question if COART is not highly correlated to (one of) the formants. The highest correlation observed between COART and a formant is the one between COART and F_2 of /u/; it was $r = 0.55$ ($n = 1200$), which is rather low. This makes it interesting to evaluate the question.

Speaker identification percentages were acquired by utilising the classification option of Linear Discriminant Analysis (LDA) and by introducing the 15 speakers as the groups to discriminate. For the present purpose, LDAs were carried out (a) for separate contexts of each vowel, and (b) for the pooled contexts of each vowel. In condition (a) there were (15 speakers · 10 replications =) 150 cases for each LDA; in condition (b) this number was multiplied by 8 (contexts), yielding 1200 cases. The identification percentages were based on three functions both for the LDAs on F_{1-3} and for the LDAs on F_{1-3} combined with COART. In this manner the analysis results were kept compatible. The results are presented in table 2.

Table 2: Percentages for correct speaker identification of the three vowels /a,i,u/. See text.

Condition	Vowel	F_{1-3}	COART + F_{1-3}
Pooled contexts	/a/	48.50	49.42
	/i/	43.83	42.33
	/u/	32.33	40.17
Split contexts	/a/	59.33	59.92
	/i/	62.67	66.58
	/u/	59.84	60.33

The differences between the two analysis settings (F_{1-3} vs COART $+F_{1-3}$) are clearly not substantial. The only exception is /u/ in the pooled contexts condition, where the improvement is about 8%. Thus, we find that the COART-index was not found a useful cue in speaker identification for most analysis settings.

GENERAL DISCUSSION

In this study we examined the speaker variability in the coarticulation of /a,i,u/. The amount of coarticulation in a vowel context was quantified using a score-model based measure COART.

We concentrated first on the coarticulation effects in the vowel contexts as such. It was observed that the effect of coarticulation upon /u/ was much stronger than upon /a/ and /i/.

In the next section we looked at the speaker variability in the observed coarticulation phenomena. It was found that the coarticulation in precisely the vowels that showed significant differences between contexts, proved to be speaker-specific as well (i.e. the vowels /a/ and especially /u/).

Guided by the finding that COART in /a,u/ had turned out to be speaker-specific, we tested if COART is a useful additional parameter for automatic speaker identification. Our results quite convincingly indicate that the COART-index is not a valuable coefficient for this task. Similar findings have been reported for /l/ and /r/ by Nolan (1983) [4]. As yet high speaker identification scores for a coarticulation measure have been presented only by Su, Li & Fu (1974) [3] for /m/. But also in their paper there is no proof that the coarticulation measure performs better than or just as good as simple spectral coefficients of /m/.

It is evident that the experimental setting described in this paper deviates considerably from the conditions normally encountered in (automatic) speaker recognition. There, the setting will be less formal and the recording background and transmission channels more noisy. Moreover, automatic speaker recognition nowadays operates increasingly more on sentence material and less and less on isolated words. Hence, stronger coarticulation and probably more between-speaker variability in coarticulation can be expected in such more complex speech data.

REFERENCES

[1] D.P. Kuehn and K.L. Moll (1976), "A cineradiographic study of VC and CV articulatory velocities", *J. of Phonetics*, Vol. 4, pp. 303-320.

[2] K. Johnson, P. Ladefoged and M. Lindau (1993), "Individual differences in vowel production", *JASA*, Vol. 94, pp. 701-714.

[3] L.-S. Su, K.P. Li and K.S. Fu (1974), "Identification of speakers by use of nasal coarticulation", *JASA*, Vol. 56, pp. 1867-1882.

[4] F.J. Nolan (1983), *The phonetic bases of speaker recognition* Cambridge: Cambridge University Press.

[5] L.C.W. Pols, H.R.C. Tromp and R. Plomp (1973), "Frequency analysis of Dutch vowels from 50 male speakers", *JASA*, Vol. 53, pp. 1093-1101.

RELATIONSHIP BETWEEN GESTURES AND VOICE IN VERBAL INTERACTION: PROSODIC AND KINESIC ASPECTS OF BACK-CHANNEL SIGNALS

Roxane Bertrand, Jacques Boyer*, Christian Cavé*, Isabelle Guaïtella* and Serge Santi***

**Laboratoire Parole et Langage, U.R.A. 261 CNRS*
Université de Provence, Aix-en-Provence, France
*** Laboratoire de Phonétique, GRI-DESYCOLE,*
Université de Franche-Comté, Besançon, France

ABSTRACT

This paper points out the theoretical merits of (1) investigating the relationship between vocal and gestural activity in speech, and (2) using this perspective to study back-channel signals (which can be verbal, vocal and/or gestural). The main results of two preliminary studies on back-channel signals are reported.

1. THEORETICAL BACKGROUND

1.1 Three-modality communication and the relationship between gestures and voice

In a trimodal model of communication [1, 20, 11, 14], interpersonal exchanges are based on three communication modalities: verbal, vocal, and gestural. It was hypothesized here that among these three modalities, vocal and gestural activities are tightly linked [14, 15]. Some phoneticians contend that the gestures which co-occur with speech are linked to intonation by their temporal features, and above all, by their semiotic characteristics [17, 7, 14, 15]. The study of the relationship between eyebrow movements and variations in fundamental frequency has shown that, although the interaction between the two depends on the context and the speaker, these two kinds of movement are clearly synchronized [9, 16, 2]. Phoneticians working on phonatory gestures and their interaction with expressive phenomena [5, 6, 12, 13, 21, 22] have contributed to our understanding of the link between vocal and gestural expression in speech. If we agree that the expressivity of the voice is related to a "glottal movement", i.e. that the voice is the audible trace of a physiological activity, and that such "internal gestures" are comparable to "external gestures" such as facial expressions, then we can assume that there is a link between the visual and auditory channels. The lack of a link would be surprising in that it would reflect the disconnection of internal and external gestures. Some researchers in non-verbal communication or human ethology [4, 10, 11] have addressed the dual question of the micro-analysis of gestures (in Condon's terminology) and their link to the vocal component. The problem of bimodal perception can also be considered relevant to this line of research.

1.2 Conversational Feedback

During conversation, the listener contributes to the interaction by exhibiting an active listening attitude, or by showing his/her desire to speak through the production of specific signals [18, 23, 19]. It is well known that turn-taking is controlled by listener-produced feedback called back-channel signals. Back-channel signals can be vocal, verbal, gestural, or any combination of the three. The acoustic characteristics of vocal back-channel signals have not been sufficiently described (see however [24]). Likewise, little research has been conducted on the forms and functions of gestural signals [8]. Such studies could help us gain insight into the relationships between the forms and functions of back-channel signals, and allow us to suggest a precise and objective typology.

2. EXPERIMENTAL STUDIES

Two studies were conducted to describe the formal and functional characteristics of vocal and gestural back-channels signals. In the first study, the prosodic characteristics of vocal signals were analyzed in relation to their functions. In the second study, the same types of analyses were performed, but gestural activity was also taken into account.

2.1 Experiment 1: Prosodic and Functional Analysis

In this preliminary study [3], the prosodic and pragmatic aspects of back-channel signals in turn-taking were investigated. Most of the observed signals (10 out of 15) had a flat or slightly falling prosodic contour. The listener appears to use these signals to show that he/she is listening but does not wish to interrupt. Among the other five cases were two repetitions of the speaker's utterance. These signals had rising contours and can be interpreted as questions. There were also two isolated signals whose function appears to be to prompt the speaker to continue. For the remaining case, we do not have a functional interpretation to propose.

As a whole, the prosodic contours of the listener's signals were inverted with respect to the contour of the preceding speaking turn. In other words, back-channel signals with rising prosodic contours follow utterances with an overall falling intonation, while those produced with flat or falling contours follow utterances whose overall pattern is rising. Thus, two types of vocal back-channel signals can be distinguished: (1) those with a rising prosodic contour, which appear to have a continuation function, and (2) those with a flat or falling contour, which manifest an active listening attitude.

2.2 Experiment 2: Prosodic and Gestural Analysis

In this study, the prosodic analysis of vocal back-channel signals was extended by an analysis of the gestural feedback produced by the listener (Boyer, doctoral dissertation, in progress).

2.2.1 Corpus. The corpus consisted of a 3-person discussion recorded in a soundproof room. The speakers were seated in a triangular arrangement and were filmed by two synchronized video cameras. The topic of the discussion was their current work, but some informal exchanges also took place. The total duration of the recording was about 20 minutes.

2.2.2 Data analysis. The films were coded by visual inspection using a U-matic videotape recorder. The gestures noted were head movements, hand movements, and direction of gaze. The vocal parameters considered were fundamental frequency, sound intensity, and duration. Back-channel signals produced by both listeners at the same time (double back-channel) and by only one of the listeners (single back-channel) were analyzed.

2.2.3 Results and comments. The number of back-channel signals produced varied across subjects. The results are presented in two parts: (1) a prosodic description and (2) a vocal and gestural typology.

The prosodic analysis dealt with the fundamental frequency (in Hz), the intensity (en dB), and the duration of the segment analyzed (in ms). For complex patterns (mixed rising and falling intonation contours), the largest variation was considered. Flat intensity curves were rare and were included in the falling patterns.

For the most part, the vocal back-channel signals exhibited a drop in intensity and frequency. The frequency variation was greater for falling patterns than for rising ones. The durations were more stable for listeners J and R than for listener I. This may be related to individual differences or to the particular roles played by the interlocutors in this conversation. In fact, J and R tended to speak to I, who appears to have been the favored partner for the other two. When double back-channel signals occurred, they had the same characteristics for both speakers.

The parameters used in defining a vocal and gestural typology were the variations in intensity and fundamental frequency for the vocal channel, and the direction of gaze and head and hand

movements for the gestural channel. For the hand, movements involving only one hand were distinguished from coordinated movements of both hands. In fact, there were no gestures of the left hand alone.

We can see that only a few of the potential combinations actually occurred. The most frequent patterns were those where a drop in intensity and frequency was associated with a change in direction of gaze, or with changes in head orientation and direction of gaze.

3. CONCLUSION

In this study, we showed how gestures can interact with vocal parameters in the production of conversational feedback. These preliminary results will be extended by further studies aimed at determining the precise temporal organization of the relationship between vocal and kinesic back-channel signals. One of these studies, based on the data obtained from a movement analyzer, is currently in progress.

REFERENCES

[1] BALLY C., 1925, *Le langage et la vie*, Librairie Droz, Librairie Giard.

[2] BERTRAND R., 1993, *Mise en relation de l'activité de la face avec les paramètres prosodiques et la chaîne segmentale dans un corpus d'interview télévisé*, Mémoire de DEA de phonétique expérimentale, fonctionnelle et appliquée, Université de Provence.

[3] BERTRAND R., 1994, *Approche pragmatique et prosodique de l'interaction conversationnelle*, Mémoire de D.E.A. de linguistique générale, Université de Provence.

[4] BIRDWHISTELL R., 1970, *Kinesics and context*, University of Pennsylvania Press.

[5] BOLINGER D., 1946, "Thoughts on Yep and Nope", *American Speech*, 21, 90-5.

[6] BOLINGER D., 1972, "Accent is predictable (if you're a mind-reader)", *Language*, 48, 633-44.

[7] BOLINGER D., 1985, *Intonation and its parts*, Edward Arnold.

[8] BOYER J., 1993, *Mise en relation des mouvements des bras et de la tête avec les paramètres prosodiques dans un corpus d'interview-reportage télévisé*, Mémoire de DEA, Université de Provence.

[9] CAVÉ C., GUAÏTELLA I. & SANTI, 1993, "Fréquence fondamentale et mouvements rapides des sourcils: une étude pilote", *Travaux de l'Institut de Phonétique d'Aix*, 15.

[10] CONDON W.S., 1976, "An analysis of behavioral organisation", *Sign Language Studies*, 13, 285-318.

[11] COSNIER J., 1988, "Grands tours et petits tours", in: Cosnier, Gelas & Kerbrat-Orecchioni (eds), *Echanges sur la conversation*, Editions du CNRS, Lyon, 175-84.

[12] FONAGY I., 1962, "Mimik auf glottaler Ebene" *Phonetica*, 8, 209-19.

[13] FONAGY I., 1967, "Hörbare Mimik", *Phonetica*, 16, 25-35.

[14] GUAÏTELLA I., 1991, *Rythme et parole: comparaison critique du rythme de la lecture oralisée et de la parole spontanée*, Thèse de Doctorat, Université de Provence.

[15] GUAÏTELLA I., 1995, "Mélodie du geste, mimique vocale", *Semiotica*, 103, 3/4.

[16] GUAÏTELLA I., CAVÉ C. & SANTI S., 1993, "Relations entre geste et voix: le cas des sourcils et de la fréquence fondamentale", *Actes du colloque Images et Langages, Multimodalité et modélisation cognitive*, Paris, 261-8.

[17] HEESE G., 1957, "Akzente und Betleitgeärden", Sprachforum, 2, 274-85.

[18] KERBRAT-ORECCHIONI C., 1991, *Les interactions verbales*, tomes 1 et 2, Colin.

[19] LAFOREST M., 1993, *Le back-channel en situation d'entrevue*, Recherches sociolinguistiques, 2, Université Laval, Québec.

[20] MEHRABIAN A., 1972, *Silent messages*, Wadsworth, Belmont.

[21] OHALA J.J., 1980, "The acoustic origin of the smile", *J. Acoust. Soc. Am.*, 68, 33.

[22] OHALA J.J., 1984, "An ethological perspective on common cross-language utilization of fo of voice", *Phonetica*, 41, 1, 1-16.

[23] VION R., 1992, *La communication verbale, Analyse des interactions*, Hachette.

[24] WERNER S., 1991, "Understanding 'hm', 'mhm', 'mmh'", *Proceedings of the XIIth International Congress of Phonetic Sciences*, Aix-en-Provence, vol.4, 446-8.

Table 1. Mean (M), maximum, and minimum variation in fundamental frequency, intensity, and duration for the different prosodic patterns. Single and double back-channel signals are shown for each speaker.

		↓F0. variation in Hz			↑F0. variation in Hz			↓Int. variation in db			↑ Int. variation in db			Duration in ms		
speaker	Case	M	Min.	Max.	M	Min.	Max.	M	Min.	Max.	M	Min.	Max.	M	Min	Max
single R	14	41	5	110	30	2	59	6 5	3	16	2	1	3	268 2	161	395
double R	12	26 9	6	63	41 5	24	59	5 5	1	11	6 3	4	8	252 2	167	472
single I	13	26 8	9	46	29 8	3	63	6 7	2	17	5 5	4	7	350	145	877
double I	1	33	33	33	/	/	/	/	/	/	/	/	/	157	157	157
single J	12	31 5	15	62	8 6	1	21	8 5	1	15	/	/	/	234	134	314
double J	11	12	2	33	15	1	38	6 7	2	14	5	5	5	264 5	173	444

Table 2. Number of cases of each intensity and frequency pattern.

Speaker	F0 ↘	F0 ↗	Complex F0	Int. ↘	Int. ↗	Complex Int.
Single R	9	3	2	9	2	3
Double R	9	1	2	8	1	3
Single I	5	4	4	9	1	3
Double I	1	0	0	1	0	0
Single J	9	3	0	10	0	2
Double J	8	3	0	9	1	1

CHARACTERIZATION OF THE NON-LINGUISTIC INFORMATION OF VOWELS BY MATCHING VOWEL SYSTEMS

Jean-Sylvain Liénard, LIMSI-CNRS, Orsay, France
Maria-Gabriella Di Benedetto, INFOCOM, Univ. La Sapienza, Rome, Italy

ABSTRACT

Vowel system normalization does not succeed, in general, in totally cancelling the scattering areas of vowels. Considerable variations remain, which are due to the speaker and context peculiarities. In the present study we consider all types of information as equally important in the analysis of speech structures. Using the Peterson and Barney data we show that, by matching vowel systems, linguistic information can be associated with an average system considered as reference, while non-linguistic information lies partly in the parameters of the transform which gets an individual system close to the reference, and partly in the deviation of the transformed system with respect to the reference.

GENERAL PRESENTATION

Previous work on Vowel Systems (VS) normalization from three cardinal vowels resulted in some reduction of the non-cardinal vowels scattering areas, but failed in eliminating all classification errors [1]. The very notion of normalization implies some limitations : imposing too close a match between two VSs yields to neglect some relevant discrepancies among languages, dialects or individuals [2]. We propose here to compare VSs to each other as wholes, and to look into the parameters of the matching transformation for what corresponds to linguistic information (i.e. the explicit phonetic code of the language) as well as for what corresponds to non-linguistic information (i.e. the identity and vocal gender of the talker, the type of voice, etc.). This approach is an illustration of the Speech Pattern Processing paradigm [3], according to which all aspects of the perceptive information of the signal must be taken into account simultaneously.

From Peterson and Barney's vowel formant measurements [4] we determine a Reference Vowel System (RVS) ; then we define several transforms aiming at a "best match" of the RVS with the individual VSs. For each transform we compute the error-rate obtained in the classification of the vowels, speakers, and vocal gender (male, female, child). Finally we discuss the ability of each transform to provide an adequate representation of the relevant information.

CORPUS AND TRANSFORMS

Peterson and Barney's data comprise 10 American vowels uttered twice by 76 speakers (33 males, 28 females and 15 children). We only use the F1 and F2 measurements. The RVS is arbitrarily obtained by averaging the male VSs, which are the most represented in the database, and the less subject to formant frequency measurement errors.

We look for a transform which achieves an optimal matching of two sets of homologous points RVS and VS in the (F1,F2) coordinates (fig 1). In order to get some generality this transform must be simple : it is made of scalings and translation. Thus both systems cannot be exactly mapped onto each other : the direct transform changes RVS into a new system which best approximates VS ; the inverse transform changes VS into a new system, called Inverse Vowel System (IVS), which best approximates RVS. The quality of the approximation refers to the mean quadratic distance between homologous points of both systems,

measured in the (F1,F2) plane, with a weight of 2 in favor of the F1 dimension, to compensate for the interval (in Hertz) between extreme values, which is about half for F1 than it is for F2. The translation which gets a system as close as possible to the other is defined by the vector joining both centers of mass.

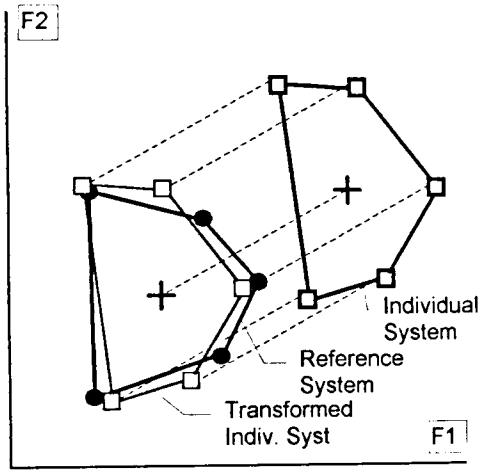


fig 1 : Vowel systems and transform

Such a translation, done with F1 and F2 replaced by their logarithms, corresponds to the multiplication of all formant frequencies by the same factor. It looks quite natural, in view of the perceptual theories based on the constancy of the formant frequencies ratio for a given vowel (cf, among other authors, Nearey's log-mean normalization reported in [2] and the formant-ratio theory evoked by Miller [5]). This transform, termed "Simple Log", is defined for each VS by the value of a single parameter (log of the best scaling factor in F1 and F2).

The "Double Log" transform does not presuppose this constancy : it applies a translation according to the F1 dimension, and another one in the F2 dimension. Thus it is defined by two parameters (logs of the best scaling factors in F1 and in F2).

Two other transforms, "Simple Bark" and "Double Bark", are similarly defined by replacing the log by the Bark function : close to the linear scale in the low frequency range and close to the log scale in the high frequency range.

We also used the Gerstman transform, which consists of linearly mapping the interval between extreme values in the F1 dimension of both systems ; the same process holds independently in the other dimension [6]. This transform requires 4 parameters.

The case for which the original data is used, i.e. no transform is applied, is called the Null transform.

For a given transform we make the following calculations. Each VS of the database transformed into an IVS as close as possible to the RVS. The parameters are kept, as well as the set of deviations (in both coordinates) that remain between homologous points of RVS and IVS. Thus there is no loss of information : each VS can be exactly reconstructed knowing the RVS, the parameters and the set of deviations. Then several error-rates are computed, concerning the vowel identity, vocal gender and talker identity (table 1).

table 1: error-rate in the classification according to several kinds of information and several transforms

transform	vowel err. from distances	gender.err. from params	talker err. from params	talker err. from dev. sets
random	*90.0*	*63.7*	*99.3*	*99.3*
Null	33.8	-	-	17.8
Simple Log	13.5	8.6	86.8	28.9
Double Log	11.5	9.2	65.8	36.8
Simple Bark	14.6	7.9	80.3	27.6
Double Bark	11.2	7.9	69.7	37.5
Gerstman	16.6	9.9	73.0	50.7

RESULTS

Errors on vowel categories

Each token of the IVS is compared to the RVS tokens (computation is based on the distances after transformation). The closest one is selected. If the vowel labels do not match an error is counted. Computation is extended to the whole database.

All the transforms improve the initial situation, which is normal. Double Log and Double Bark, with two parameters each, perform best. Simple Log and Simple Bark yield slightly degraded results but remain surprisingly efficient given the fact that they use one single parameter. Finally Gerstman seems less efficient despite its 4 parameters. This can be attributed to the fact that only 3 or 4 vowels, out of 10, contribute to the determination of the transform, making it non optimal for the whole VS. The same remark holds true in the further experiments.

Errors on the vocal gender computed from the transform parameters

First three average values of the parameters are computed, each one relating to a vocal gender, on the whole database. Then for each VS a quadratic distance is computed (on the parameters) to the three gender representatives, and the closest one is selected. An error is counted when the gender labels do not match. All transforms practically give the same result (8 to 10%), which seems good as compared to random gender allocation (63.7%). The most interesting observation is that one parameter is sufficient to characterize this aspect ; adding other parameters does not change anything.

Errors on talker identification from the parameters

For each VS the parameter values are compared using a quadratic distance to the parameter values of the 151 other systems of the database. The closest one is selected. If the talker labels do not match an error is counted.

Globally the error-rates are high, which shows that the parameters do not capture much talker-specific information. Double Log and Double Bark perform slightly better than their single-parameter conterparts.

Errors on talker identification from the deviation sets

For each VS the deviation set (10 components in F1, 10 in F2) is compared using a quadratic distance to the deviation sets of the 151 other systems of the database. The closest one is selected. If the talker labels do not match an error is counted.

Contrary to the previous experiment it appears that the deviation set captures most of the talker-specific information. Compared to the others, the one-parameter transforms leave more room to expressing the individual talker peculiarities in the proper shape of the transformed systems and consequently in the deviation sets.

On the relative merits of the transforms

Leaving aside the Gerstman transform for the above-mentioned reason, our experiments show that

- most of the gender information lies in a single parameter, namely the log-mean or bark-mean ratio,
- using two parameters instead of one results in assigning more talker-specific information to the parameters, more vowel-specific information and less talker-specific information to the transformed systems,
- the log and bark scales yield comparable results.

CONCLUSION

By considering the matching of vowel systems as wholes we showed, using the Peterson and Barney data, that linguistic and diagnostic kinds of information could be - at least partially - decorrelated.

Linguistic information, common to a group of talkers, lies mostly in the relative disposition of the vowels after some talker- or gender-specific transformation. This disposition may be materialized in the Reference Vowel System by vocalic categories defined by their prototypes. When a particular Vowel System is transformed so that it best approximates the Reference Vowel System the vowel error-rate drops below 15%.

Diagnostic information, related to the talker or voice specification, can be found partly in the transform parameters, partly in the set of deviations of the transformed system with respect to the Reference Vowel System. The proportions vary according to the number of parameters, that is the ability of the transform to match both systems more closely. It is remarquable that a single parameter, namely the scaling factor or its logarithm, conveys most of the information about the vocal gender.

REFERENCES

[1] Di Benedetto, M.G. and Liénard, J.S., "Extrinsic normalization of vowel formant values based on cardinal vowels mapping", ICSLP, Banff, Oct 12-16, 1992.

[2] Ferrari-Disner, S., "Evaluation of vowel normalization procedures", J.Acoust.Soc.Am. 67(1), 253-261, 1980.

[3] Liénard, J.S., "Speech Pattern Processing : integrating the linguistic and non-linguistic aspects of voice and speech", ICPhS, Stockholm, Aug. 13-19, 1995.

[4] Peterson, G. and Barney, H., "Control methods used in a study of the vowels", J.Acoust.Soc.Am. 24, 175-184, 1952.

[5] Miller, J.D., "Auditory-perceptual interpretation of the vowel", J.Acoust. Soc.Am. 85(5), 2114-2134, May 1989.

[6] Gerstman, L.J., "Classification of self-normalized vowels", IEEE Trans. on Audio and Electroac., AU-16, 1, 78-80, 1968.

Research supported by a cooperation agreement between the French Centre National de la Recherche Scientifique and the Italian Consiglio Nazionale delle Ricerche.

SPEECH PATTERN PROCESSING : INTEGRATING THE LINGUISTIC AND NON-LINGUISTIC ASPECTS OF VOICE AND SPEECH

Jean-Sylvain Liénard, LIMSI-CNRS, Orsay, France

ABSTRACT

In this paper we try to show that speech variability, instead of being a problem to be solved by desperately looking for acoustical invariants, actually reveals the fact that the problem is ill-posed. The central thesis is that at any level of abstraction the signal content should be fully specified, in a way integrating non-linguistic information, mainly related to voice, as well as linguistic information, mainly related to speech. By refering to a vowel classification experiment described in another paper we aim at evidencing in what the Speech Pattern Processing paradigm modifies the traditional approaches to speech automatic processing.

GENERAL PRESENTATION

In this paper the variability problem is related to the non-linguistic information of speech. We claim that the usual speech analysis processes (in the field of phonetics) as well as automatic recognition (in the field of automatic processing), by emphasizing the importance of the linguistic aspects of speech, implicitly admit that the non-linguistic information must be eliminated or controlled. A way to eliminate it is to look for invariants in the signal which would exclusively code the linguistic information. This approach generally does not yield the expected results because of the excess of variability

We propose another view [1], according to which speech processing cannot be reduced to extracting linguistically relevant items from the signal. Speech processing is viewed as a cascade of representation changes. At each abstraction level we associate a "complete representation", that is a set of linguistic and non-linguistic symbols or variables, from which it is possible to reconstruct a signal perceptively similar to the one which is under analysis. The set of differences between signals having the same complete representation constitutes the "residual variability". According to this view what is usually called "variability" is mainly the acoustic consequence of a lack of non-linguistic information in the higher level representation of speech. This general view applies here to the study [2] presented at ICPhS-95, concerning the matching of vowel systems.

SPEECH VARIABILITY

Variability is often presented as the particular ability of speech units to appear under different forms within a given linguistic category, at any abstraction level. Thus it is recognized as the main problem in the process of decoding the speech signal in artificial systems. Implicitly the decoding process is viewed as an information reduction process across the abstraction hierarchical levels, starting from the signal to end up with a set of semantic units. In other words, speech perception and recognition are conceived as reductive processes, guided at any step by the search for invariants.

As this view does not allow, in its general form, to build successful recognition systems, a set of constraints are imposed upon the process : vocabulary, syntax, speakers, recording conditions, task domain are limited or restricted. Eventually, the natural constraints that the system should take into account become constraints that are artificially imposed to the speaker. Thus as soon as a new speaker does not obey these constraints the system performances drop tremendously.

In the past few years the idea that variability should not be denied but recognized as the manifestation of some useful information has gained weight. Psychologists, particularly after Pisoni [3], have shown that human subjects can use non-linguistic (or "episodic") information contained in the signal in order to improve the perception and recall of linguistic (or "semantic") information. This idea is not yet used in artificial systems, which keep looking for absolute acoustical invariants or which, by default, aim at capturing all the possible variants for a given linguistic content.

SPEECH PATTERN PROCESSING

Another trend in psychology deals with categorization. According to Rosch [4] the signals emitted by the external world are categorized by the human cognitive system according to two principles, the first one taking into account the world structures as well as the properties of our perceptive system, the second one expressing the search for efficiency (the maximum information for the least cognitive effort). It would probably be wise to add a third principle, stating that at least some conceptual categories should be common to a given group of individuals.

Categorization theories, as well as most artificial systems, actually refer to what can be called the Pattern Recognition paradigm : at some point the problem always comes to putting several signals in the same category. The set of differences between the constituents of the same category constitutes the variability. By essence, the Pattern Recognition process is bottom-up and non reversible : specifying the category does not permit to specify a particular constituent. Usually one of them is chosen to represent all the others ; it is called "prototype", and each other is reputed to be more or less "typical" according to its distance to the prototype. In the traditional view this categorization or recognition process reproduces itself from level to level until it has (ideally) eliminated all of the "useless" or "redundent" information, i.e. the non-linguistic information.

However non-linguistic information is always present. It actually allows us to differentiate several signals pertaining to the same category. For instance the Peterson and Barney measurements show partly overlapping areas for adjacent vowels, which is a cause of classification errors. But the points in a given vowel area cannot be considered as equivalent ; when a point lies in a zone common to two categories, any additional knowledge about the talker may help in the desambiguation. If we know that the talker is a child the point lies preferably toward the external end of the area, while it probably pertains to the internal end if it comes from a male voice.

It should be pointed out that a definition "in extension" of the categories (i.e. a category is not defined by its prototype but is simply the collection of all its constituents or "exemplars") does not help in solving the problem : if two areas overlap there is still a misclassification problem in the common zone.

Instead of trying to compress the information content by using hypothetical invariants, prototypes or exhaustive inventories, we suppose now that the main task of the perceptive system is to gradually change the representation of the data. At each level of abstraction one looks for a "complete" representation preserving all of the perceptual content of the signal, while contributing to separate its different aspects a little bit more than at the immediately inferior level. For operational reasons it is sometimes possible to isolate a step of this process, joining two adjacent levels : for instance in the Peterson and Barney case the low level consists of the formant measurements, while the high level consists of the set of linguistic and non-linguistic descriptors (vowel category, vocal gender and talker identity).

In order to make sure that we have defined enough descriptors to form a complete description we should ask the question of the signal reconstruction from this description. If the resynthesized signal can be considered by the ear as perceptually equivalent to the original, then it means that the high-level description contains the relevant information. It does not mean that any variability has been eliminated : some variability may remain but it can be considered as "residual", i.e. non-relevant with respect to the perceptual process.

It is worth mentioning that the Speech Pattern Process is basically reversible (bottom-up and top-down) : a given high-level description gives birth to an unambigous signal, or to a set of signal which are perceptually equivalent by definition.

At any level the description is structured and constitutes a "pattern". Thus perception is viewed as a hierarchy of structured representations, ending at the upper level with a set of abstract descriptors well decorrelated from each other.

At this level the cognitive system selects the descriptors which convey the information required either to guide the system behaviour or to facilitate the signal decoding.

Pattern Processing does not conflict with Pattern Recognition, but completes it. If at any level the descriptors representing the non-linguistic information are suppressed, then Pattern Proccessing becomes Pattern Recognition. The top-down process is lost and variability becomes the major problem.

AN EXAMPLE

Let us now illustrate the above notions by an example taken in the field of vowel perception. In [2], from Peterson and Barney's classical measurements in the F1/F2 plane, we show that the apparent scattering of the measurements can be reduced by using speaker-specific transforms. A Reference Vowel System (RVS) has been computed by averaging the F1 and F2 values for each vocalic category, yielding 10 vowel prototypes.

Let us rephrase the problem from two different viewpoints :

- **Pattern Recognition or Categorization viewpoint** : *given a new, unknown vowel token defined by its F1 and F2 frequencies, pronounced by an unknown talker, determine its phonetic category.*

One way to solve this problem is to measure the distance of the unknown token to all of the RVS prototypes and to give the token the category of the nearest neighbour. This yields some 34% error-rate.

- **Pattern Processing or Multi-Categorization viewpoint** : *given a new token of which we know something, for instance the talker's vocal gender, complete its high-level description.*

This formulation implies that we also know something on the way in which talkers of different vocal genders deviate from the RVS. The study shows that, when the vocal gender effect is compensated (i.e. after the proper Simple Log or Simple Bark inverse translation), then the error-rate on the vowel category drops below 15%.

This is typically the kind of improvement that would come out from what is classically called "speaker adaptation" in the Speech Recognition domain. But the Speech Pattern Processing view is much wider. For instance let us formulate the problem in the following manner : *given a new token of which we have a high-level description (vowel category and speaker or pseudo-speaker combination) , complete its low-level description, i.e. compute the F1 and F2 values.* This sounds more like speech synthesis than speech recognition, although the same view is applied.

More generally, the Pattern Processing paradigm aims at relating two descriptions of the same signal content at different levels. Even if both descriptions are incomplete, it may happen that they complement each other. For instance : *given a new token of which we know F1 (one low-level descriptor out of two) and vowel category and speaker vocal gender (two high-level descriptors out of three), complete the low- and high-level descriptions, i.e. compute F2 and determine the plausible identity of the speaker.* In this case the Pattern Processing module simultaneously uses top-down and bottom-up processing to achieve its task. Thus such a module can be seen as part of an active perception process and cannot be reduced to a mere speaker adaptation mechanism.

VOICE AND SPEECH

Traditionally voice and speech are different, uncorrelated notions. Speech has something to do with the linguistic aspect of the signal, while voice mainly reflects some speaker properties. This idea may be related to the source-filter decomposition which prevails in the field of speech production. However there are many aspects of the signal content in which both notions cannot be clearly distinguished, for instance Fo evolution. A part of it seems to be governed by linguistic considerations of different levels (phonetic, lexical, syntaxic), another part is related to the intaction between interlocutors in a given situation, a third part can be attributed to extra-linguistic factors such as talker identity, mood, physical state, intentions, affectivity, awareness of the acoustical conditions, social relationship with the interlocutor, etc. Such considerations can also be formulated for the other aspects of prosody, with the supplementary remark that prosodic factors like duration or stress cannot be defined without some knowledge of the linguistic items to which they apply (phonemes, syllables, words).

At the highest level of perception we perceive the numerous aspects of the non-linguistic information as separated from each other, as well as from the linguistic information, despite the fact that they are intimately mixed in the signal. Besides we can at will focus our attention on whatever kind of information of interest.

If non-linguistic information is present at the cognitive level, it must also be present at the intermediary levels, in a most intricate form at the levels close to the signal and in a more separate or decorrelated form at the higher levels. Prosody is precisely one of those intermediary notions, in which the many aspects of information still strongly depend on each other. Thus at any abstraction level the study of voice and speech structures must be done jointly.

The biggest problem lies in the definition of the descriptors of the non-linguistic information. We know some relevant attributes of voice, mostly at the low level (average pitch, jitter, intensity, etc.), but the basis of a realistic and efficient description of voice at the highest level, for instance a number of prototypical voices on which an agreement could be obtained among different social groups, has not been firmly established yet. Let us point out the fact that the study of the so-called intra-talker variability, including the effect of the vocal effort, has been widely neglected until now.

We observe that this approach agrees with the problems presently met by speech synthesis from the text, which lacks some naturalness. The non-linguistic descriptors are not specified in the written text, and the characteristics assigned to the pseudo-speaker's voice are implicitly determined in an arbitrary and rudimentary manner. In other words, in order to provide more naturalness to the synthetic voice it is necessary to specify the necessary non-linguistic information, which is impossible if the adequate descriptors are not known.

CONCLUSION

Speech Pattern Processing generalises Speech Pattern Recognition which presently prevails in the study of artificial systems as well as of human perception. This new view yields the integration at any level of all kinds of relevant (perceptual) information, be it linguistic or not. It also takes into account the active aspect of perception, made of two flows of informations bottom-up and top-down. Thus it appears that speech recognition, speaker identification, recognition of the elocution and recording conditions, and even speech synthesis, actually form a single and the same problem, to be treated in a unified framework.

REFERENCES

[1] Liénard J.S. : "From speech variability to Pattern Processing : a non-reductive view of speech processing", in "Levels in Speech Communication : Relations and Interactions", C.Sorin et al. eds, Elsevier Science Publishers, 1995.

[2] Liénard J.S. and Di Benedetto, M.G. : "Characterization of the non-linguistic information of vowels by matching vowel systems", ICPhS, Stockholm, 1995.

[3] Pisoni, D.B. : "Some comments on invariance, variability and perceptual invariance in speech perception", 2nd ICSLP, Banff, 587-590, 1992.

[4] Rosch, E. : "Principles of categorization", in "Cognition and Categorization", eds E.Rosch and B.B.Lloyd, Lawrence Erlbaum Associates,, 1978.

INTER-JUDGE VARIABILITY IN PERCEPTION OF VOICE QUALITIES

F. D. Minifie, J. Green, J. Smith, and D. Z. Huang
Speech and Hearing Sciences, University of Washington, Seattle, WA 98195

ABSTRACT

Perceptual ratings by 10 speech-language pathologists of breathiness, harshness and hoarseness in disordered, and synthetic vowels were compared. The synthetic vowels were varied in jitter, shimmer and glottal noise. Rather wide variations in inter-judge ratings were observed.

INTRODUCTION

Perceptual judgements of voice quality are one of the key indicators in the clinical diagnosis of voice disorders, yet little is known about the sources of inter-judge variability in classifications of breathiness, harshness and hoarseness. The purpose of the present study was to compare individual differences in the ratings of breathiness, harshness and hoarseness by using 10 trained speech-language pathologists as judges. By using a repeated measures experimental design, it was assumed that rater variability within and among judges can be specified for the corpus of disordered vowel stimuli, thereby providing a general indication of the stability of measurement within and across judges. However, it was further assumed that the sources of perceptual variability could be studied in greater detail via the use of synthetic stimuli wherein a single acoustic variable at a time can be manipulated.

METHODS AND PROCEDURES

Digitally recorded tokens of sustained vowel sounds produced by talkers with verified laryngeal pathologies and computer synthesized sustained vowels generated to simulate varying levels of acoustic perturbations assumed to result from vocal pathology were used as stimuli in this perceptual (speech-language pathologists) were used as judges and asked to perceptually rate the sustained vowel tokens.

Disordered Vowel Stimuli

The disordered voice samples used in the present investigation were selected from among those collected by Hirano at Kurume University in 1987. Two hundred and eight recordings of the vowel /ae/ produced by talkers presenting disordered voice samples were used as the original corpus of disordered vowel tokens for this study. It should be mentioned that not all of the vowel tokens were perceived as the vowel /ae/ by listeners raised in the United States. The perceptual variations in vowel quality are probably related to differences in vowel space for Japanese versus American English listeners. Nevertheless, these vowel stimuli were acoustically analyzed to obtain measures of jitter, shimmer and normalized glottal noise energy (NNE), acoustic perturbations of the vocal sound source assumed to be related to perceptual judgements of breathiness, harshness and hoarseness. The durations of the sustained vowels produced by the voice disordered Japanese talkers varied widely - from 200 ms to 3000 ms due largely to the extent of laryngeal pathology and consequent difficulty in phonation. Some of the disordered vowel tokens were unanalyzable, due to unacceptable levels of noise from nonvocalic sources (environmental noises), or from the presence of silence gaps within the sustained vowel. The presence of the silence gaps in the disordered voice samples served to reduce the number of measurable contiguous pitch periods to an unacceptably small number. Hence, unreliable perturbation data on these voice samples rendered them unusuable for the present study. The remaining 158 sustained vowel tokens from the Kurume recordings served as the disordered vowel stimuli for the present study.

Synthetic Vowel Stimuli

A series of 462 synthetic vowel stimuli was created in which only one acoustic perturbation dimension at a time was varied. These stimuli were used to determine the specific perceptual consequences of jitter, shimmer and NNE on judgements of breathiness, harshness and hoarseness.

Three-formant synthesized vowels were generated using the *Dr. Speech Science for Windows* software developed by Tiger Electronics. The synthesized vowels generated were /i, I, E, ae, a, U, u/ . To account for gender differences, synthesized vowels were created using both male and female fundamental frequencies and formant frequency data [1]. The vowels averaged 400 msec in duration. Half of the vowels were produced using a flat intonation and half were produced using a rise-fall contour.

Each vowel was synthesized with a specified amount of jitter, shimmer or NNE. The amount of the acoustic variable imposed on a vowel was predetermined to fall within one of the five acceptable levels (ranges) for that acoustic dimension. Table 1 shows the values of jitter, shimmer and NNE at each of the five levels of variation used in this experiment. Level one represented minimal amounts of that acoustic dimension (a normal vowel production level), whereas level five represented a maximum amount of that acoustic variable (comparable to the level of variation observed in severely disordered vowel productions).

In order to isolate the perceptual effects of each acoustic variable, only one acoustic change was imposed on an interation of each vowel. Given that jitter, shimmer and NNE are not mutually exclusive, phenomena, the manipulation of one variable inadvertently has some effect on other variables. For instance, when synthesizing a vowel with extensive jitter, trace levels of NNE or Shimmer would also be detected. Although the contaminations were small, they will be addressed in subsequent papers.

Table 1: Acceptable Ranges of Jitter, Shimmer and NNE

	% Jitter
Level	Acceptable Range
1	(0.00-0.00)
2	(0.68-0.83)
3	(1.35-1.65)
4	(2.03-2.50)
5	(2.70-3.30)
	% Shimmer
Level	Acceptable Range
1	(0.00-0.00)
2	(3.60-4.40)
3	(7.20-8.80)
4	(10.80-13.20)
5	(14.40-17.60)
	% NNE
Level	Acceptable Range
1	(-18)-(-20)
2	(-14)-(-16)
3	(-10)-(-12)
4	(-6)-(-8)
5	(-2)-(-4)

Twenty percent of the vowel tokens rated in this experiment were duplicated and randomly inserted into the sequence of tokens to be rated. Repeated judgements on these tokens provided the data for evaluations of the intra-judge and inter-judge reliability in the perception of vowel quality.

Perceptual Judgements

Perceptual judgements of were made by 10 graduate students in speech-language pathology. They rated

breathiness, harshness and hoarseness on each sustained vowel token produced by the patients with laryngeal pathology, the disordered vowel samples, and for each synthetic vowel stimulus. The listening task took approximately 5 hours (3 listening sessions of about 100 minutes each). Subjects were required to complete the listening task in a sound proofed booth (IIC 1200 series). All stimuli were presented through loudspeakers at t comfortable loudness level.

Judges were required to judge only one voice quality at a time. For example, they listened to 208 disordered voice samples and 462 samples of synthetic vowels in approximately 100 minutes during which they rated only one of the voice qualities (breathiness, harshness or hoarseness). During that time the judges were provided a 1-minute break after every 52 samples, a 5 minute break after hearing the disordered voice samples, and a 5-minute break after 231 tokens of the synthetic vowels. These rest breaks were inserted to resist fatigue during the listening task.

Listener Training

All listeners in the perceptual judging task received training at the beginning of the first listening session. In addition, judges were provided with a definition of the voice quality to be rated at the beginning of each listening session. The definitions were adapted from Bassich and Ludlow [2]. Perceptual training included the presentation and rating of representative vowel tokens from the real voice samples and from the corpus of synthetic vowels. The tokens used during the training session were selected by an experienced voice clinician to be representative of the range of severity for each of the voice qualities (breathiness, harshness and hoarseness) to be rated. During training the students were familiarized with the rating form.

Perceptual Ratings

Vertically arranged continuous rating scales of 10 cm in length were used to obtain ratings of breathiness, harshness and hoarseness. The polar positions on the scales were identified with dimensional adjectives (e.g. "no breathiness" to "extremely breathy"). After the brief training period during which judges practiced rating 8 sample vowel tokens for the voice quality to be rated in that session, the blocks of 52 experimental vowel tokens were presented. Each token was presented two times in succession, separated by a 500 msec interstimulus interval, followed by a three second judging interval. Judges were instructed to indicate the severity of the vowel quality being rated by marking the rating sheet immediately following the second presentation of each vowel token. The rating sheets contained a separate vertical line for each vowel token. The top of the vertical scale was "most severe," and the bottom end of the scale was "normal" vowel quality. When judging tokens the subjects were instructed to make a horizontal mark across the vertical line at a point most representative of the severity of a specified voice quality.

RESULTS AND DISCUSSION

Three major findings of this study were obtained.

1. Listeners perceived greater changes in breathiness, harshness and hoarseness when rating the disordered voice samples collected from patients with pathological vocal mechanisms, than when rating the computer generated synthetic voice samples. This trend was particularly evident for the breathiness and hoarseness conditions, and less so for the harshness ratings. This trend may result from the relatively different levels of complexity of the stimulus from live vowels to synthetic vowel tokens.

2. Inter-judge reliability varied as a function of vowel and the voice quality being rated. As shown in Table 2, judges were more reliable in repeated estimates of breathiness than they were fore hoarseness and harshness. The perception of harshness appeared to be the most difficult for judges, based on the low reliability values reported.

Table 2
Reliability - Mean Perceptual Ratings

Token	N	Breathy	Harsh	Hoarse
/ae/	50	0.82	0.46	0.81
/a/	40	0.78	0.64	0.61
/e/	10	0.71	0.48	-0.24
/i/	50	0.70	0.67	0.62
/I/	10	0.36	0.31	0.00
/U/	10	0.92	0.38	0.70
/u/	40	0.72	0.69	0.59
All vowels		0.79	0.59	0.69

3. When the perceptual ratings on repeated syllables were compared within judges, there was considerable variation in the patterns of data obtained. Clearly, some of the judges were reliable when rating all of the voice qualities identified in this experiment. Other judges appeared to be considerably more reliable when rating a particular voice quality and not so reliable on another. (See Table 3).

Table 3
Intra-judge Reliability in the Perception of Voice Quality

Judge	Breathy	Harsh	Hoarse	All
1	0.33	0.91	-0.10	0.77
2	0.83			0.83
3	0.73	0.83	0.44	0.66
4	0.81	0.61	0.60	0.72
5	0.91	0.72	0.71	0.76
6	0.80	0.48	0.79	0.74
7	0.70	0.48		0.57
8	0.93	-0.05	0.69	0.55
9	0.87	0.74	0.77	0.79
10	0.34	0.50	0.76	0.51

Comparison of these data with transformed data from obtained three normalization procedures will be compared and the implications for reporting perceptual ratings of voice qualities will be discussed.

CONCLUSIONS

Based on the foregoing analysis of data obtained from perceptual ratings of voice quality, it would seem prudent to proceed cautiously when reporting group data or data from an individual judge when reporting perceptual ratings of vowel qualities. Individual judges appear to be more reliable when rating some voice qualities than others. These data are particularly troubling when attempting to rationalize inconsistencies in clinical ratings of voice qualities. Perhaps objective measurements of jitter, shimmer and glottal noise combined with perceptual judgements will provide increased stability of measurement.

REFERENCES

[1] Peterson, G. and Barney, H.L. (1952), "Control Methods used in the study of vowels," *J. Speech Hear. Res.*, vol. 9, 68-99.

[2] Bassich, J. and Ludlow, C. (1986), "The use of perceptual methods by new clinicians for assessing voice quality," *J. Speech Hear. Dis.*, vol. 51, 133.