MAKING SENSE OF DYNAMIC, NON-SEGMENTAL PHONETICS

J.K. Local
Department of Language and Linguistic Science, University of York, UK

ABSTRACT

This paper considers some aspects of the interpretation of dynamic approaches to phonetic representation. I argue that the most pressing challenge is that of relating a motivated dynamic, non-segmental phonetics to a phonetics-free phonological analysis.

INTRODUCTION

It may seem somewhat odd that towards the end of the twentieth century phoneticians should, at their international conference, devote a special symposium to dynamic non-segmental phonetics (DNSP). By doing so we might be seen treating the topic as contentious. How could this be? Are we suggesting that it might be possible to talk of a non-dynamic, segmental phonetics? Surely it is the case that from the earliest writings we find references to the continuous, co-ordinated nature of, for instance, articulatory activity in speech production and a concern with how best to represent this complex activity. Moreover, is there not a substantial literature that reports instrumental data and analysis of speech that shows that it is clearly dynamic and non-segmental?

It is certainly the case that linguists and phoneticians have for a long period recognised the inherent multidimensional nature of speech production and that instrumental phonetic investigations in various physical domains have attempted to provide precise details of the dynamics and inter-relateness of components in this system. (See eg Ohala's vignettes from the history of the phonetic sciences [1].) There is also a general, if tacit, assumption that attention to such details is crucial if we are to gain a full and accurate understanding of the organisation of speech. The hope is that that they will account for things that currently present themselves as problematic [2, 3].

Notwithstanding this, I do not think it odd at all to be having a such symposium. It seems to me that there are substantive issues to discuss here. There is a good deal of lip-service is paid to the importance and benefits which accrue from a phonetics which might be described as dynamic and non-segmental. However, at least part of the contemporary phonetics world behaves as if the best way to talk about speech is in terms of static segmented chunks of some kind that get glued together in production. Much of the rest of the phonetics world behaves as if it wished that speech were segmentally organised and that the dynamic aspect is at best an uncomfortable inconvenience.

Consider, for instance, the following quotations from a recent important book on the principles of phonetics [4]: '[Speech] is the most highly skilled muscular activity that human beings can ever achieve, requiring the precise and rapid co-ordination of more than eighty different muscles' and 'The stream of speech within a single utterance is a continuum.... the view that ... segmentation is mostly an imposed analysis, and not the outcome of discovering natural time-boundaries in the speech-continuum, is a view that deserves the strongest insistence.' (1, 101)

Now compare them with the following from the same book: 'Chapter 12 looks at how the articulation of adjacent segments is co-ordinated...' and 'Utterances will be treated as being made up of a linear sequence of segments, which will be phonetic events of normally very short duration, manifesting the phonological units of consonants and vowels.' (6, 113)

This illustrates what is a long-standing, common approach to phonetic analysis. The researcher recognises, even stridently asserts, the inherently continuous, non-segmental nature of production. That done the categories of interpretation employed to make sense of this data immediately seek to impose what is acknowledged as an arbitrary, convenient serial segmentation. The motivation for this segmentation is rarely, if ever, dealt with explicitly or at any length (though in [4] it gets considerably more argued attention than is usual).

It is this approach to the organisation of speech that continues to provide the structure for the interpretation of data in many phonetic investigations. The word segment may be replaced by another locution but the underlying linear, segmental assumptions are pervasive.

DOES DNSP EXIST?

Up to this point I have been talking as if a distinct and readily identifiable thing called DNSP actually existed. However, a noviciate earnestly seeking a definition would not find any general recognition of the concept 'DNSP' in the received literature. They would find reference to 'non-segmental features' such as word-accent or intonation and their 'physical correlates'. They would also locate a serious and ever-growing body of work that investigates the fine time-varying detail of speech production. To this extent they might conclude that, at least, an embryonic DNSP exists. However, the 'dynamic, non-segmental' part here seems to focus on the manner of data-collection and the nature of its representation, rather than on any distinctive theoretical premises. (Indeed the noviciate might well conclude that while such studies may be elaborating a dynamic phonetics there is not much in the way - over and above the unanalysed measurement data - which could be considered non-segmental.)

Perhaps this is all there is to DNSP. Perhaps the dichotomy between dynamic and non-segmental on one hand and non-dynamic and segmental on the other reduces to nothing more than a distinction between modes of practising phonetic science. By this I mean that it could be thought that experimental, instrumental phoneticians are necessarily doing DNSP simply by virtue of the kinds of techniques they employ in capturing physiological or acoustic data. Conversely, phoneticians who employ 'traditional' transcription techniques might seem to be necessarily doing linear, non-dynamic, segmental phonetics.

Both of these views are, of course, suspect. The most common kind of analysis of instrumental results, including those that focus on the spatio-temporal dynamics of articulation, is one in which the continuous complexes are related to a linearised segmental-type phonology. It is not uncommon to find comments such as: 'a model must account for observed differences in the relative magnitude and timing of the articulatory gestures for successive segments' [5] or '...unstressed [b] has a lower frequency [displacement of lower lip] before [a] than before [ɪ]' [6]. Some of this might simply be terminological imprecision of the kind Repp [7] considers. However, it is often not clear at what level the interpretative discourse is meant to be located. All too often one suspects that the underlying interpretative framework is one of cross-parametric, linear segmented phonetics derived from a linear segmental phonology. The patterns and inter-relationships in the continuous data representations are acknowledged but then they are made sense of by reference to sequences of segmental phonological objects.

To take a simple example, electro-palatographic (EPG) data would seem to cry out for a dynamic modelling. They do, after all, provide a (partial) representation of tongue dynamics. There have been attempts to explore some aspects of the dynamics of EPG data [8]. However, to the best of my knowledge, the interpretative categories which are employed in making sense of such data are typically the traditional cross parametric segmental phonetic ones found in the IPA. (See for instance the recent tutorial on EPG in *JIPA* [9]; cf. however, [10] for a non-segmental interpretation of such data.)

On the other hand there is no necessary reason why the use of a segmental notation system should inexorably commit a phonetician to a segmental view of speech. The work of Firthian Prosodic analysts [11, 12] is instructive in this respect. Although they employ traditional segmental representations their interpretation of them is in terms of non-

segmental phonological categories and structures. This is reflected, to some extent, in the phonetic terminology they use. Typically it is parametric rather than cross parametric. Thus a Prosodic Analyst is more likely to talk of a phonological category being exponed by 'labiablity, voice and plosion' than by a 'voiced bilabial stop', which suggests [b], which in turn suggests one segmental unit.

Recently Kelly and Local [13], following the tradition of Firthian analysis, have demonstrated ways of enhancing a segmental-type notation system to facilitate a thorough-going non-segmental view of speech. They show that even with symbol segmentation of the IPA kind it is possible to 'get a sense of the ways in which concurrent articulatory components of utterances are synchronised'. Their point is not that one can or should attempt a precise reconstruction of the dynamic components from segmental-type records. Rather K&L show that is the nature of the interpretation undertaken which matters not the form of the notation per se. (In taking up this issue in what follows my emphasis will be on 'non-segmental', rather than 'dynamic' aspects in the expectation that other participants in this symposium will have more insightful things to say on this topic than I have.)

THE INTERPRETATIVE CHALLENGE

K&L's claim concerning interpretation can be generalised to include all kinds of phonetic data representation. From this perspective 'dynamic' and 'non-segmental' are seen not simply as properties of the data representation itself but primarily as issues of interpretation. So instead of talking of DNSP as a kind of phonetic investigation it is more appropriate to talk about DNSP interpretations. It is a reasonably tractable task to provide a dynamic parametric description in some phonetic domain; it is more difficult to make linguistic sense of such descriptions. It is here that the real challenge for a DNSP lies: not in developing more sophisticated techniques for obtaining data or more sophisticated dynamic models of observed behaviour but rather in finding ways of relating the observed material to linguistically meaningful (phonological) organisation so that the detail in the dynamic phonetic description maximally preserved. The problem to be resolved is not what form the observed data representations take but what interpretative sense is made of them (see also [14]).

Scully provides a clear version of the standard formulation of 'the problem': 'Links are needed to bridge the gap between the analysis if speech as a set of discrete, ordered but durationless linguistic units and analyses of the continuously changing acoustic signals, defined along a time axis.'[15] Saltzman and Munhall [16] offer an equivalent formulation in the articulatory domain in which they 'attempt 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.' I will attempt to show that a solution to this problem can be developed by formulating a structured non-segmental phonology and elaborating a compositional phonetic interpretation function.

PHONETICS-PHONOLOGY RELATIONSHIP

I have, in a rather unsubtle manner, reformulated the challenge for a robust DNSP as being concerned with the problem of the relationship between phonetics and phonology. There would appear to be three main 'solutions' to this problem: (a) maintain a segmental analysis and propose intermediate levels of representation with sophisticated mapping functions [14]; (b) eliminate the distinction between phonetics and phonology and employ the same categories in both [17, cf also 18] (c) develop a non-segmental phonology with an interpretative dynamic non-segmental phonetics [19, 20, 21].

Other participants in this symposium will be addressing issues arising from (a) and (b). I will restrict myself to a consideration of (c). In doing so I will suggest that 'the problem' identified in [15] and [16] above which seems to arise when DNSP confronts a 'discrete phonology' is spurious. It arises from a view which, in espousing an intrinsic phonetic interpretation (IPI) hypothesis, misconstrues the timeless, relational nature of phonological representation.

Non-segmental phonology and the IPI hypothesis

Building on the work of Firthian prosodic analysts, colleagues and myself at York have been developing a radical non-segmental model of phonological structure. This model is implemented in the natural-sounding YorkTalk speech generation system [19, 21]. The architecture of this approach is derived from that of Firthian Prosodic Analysis [11, 12]. Phonological representations are treated as entirely relational. They encode no information about temporal or parametric events. In the York approach the phonological representations are constructed as complex attribute-value structures. The constituents of these structures are unordered, there is no distinguished type of phonological constituent and phonological information is distributed over the entire structure and not concentrated at the terminal nodes. These non-segmental representations make it possible to express phonological contrastivity over any appropriate domain in the structure - at phrase domain, word domain, at syllable domain, at constituent of syllable (onset, rime etc) for instance. The abstract phonological categories and structures of this model are given temporal and parametric interpretation in terms of a dynamic, non-segmental phonetics. A central aspect of this approach is the rejection of the IPI hypothesis which is propounded in a number of con-temporary 'non-segmental' approaches where features in the phonology are deemed to embody a transparent phonetic interpretation - typically cued by the featural name [17, 22].

The position I am outlining does not mean that I see no interesting or 'explanatory' links between phonetic phenomena and phonological structures. Rather my claim is that if we wish to develop a sophisticated understanding of the relationships between the meaning systems of a language and make sense of their dynamic exponents in speech, then being forced to provide an explicit statement of the detailed parametric phonetic exponents of phonological structure is an essential prerequisite. The feature labels for phonological units we employ may be given mnemonic labels but their relation to the phonic substance need not be simple. Because they are distributed over different parts of the syllabic structure, their interpretation is essentially polysystemic [11]. For example, the interpretation of the contrast given the feature label [+ *nasal*], say, at a syllable onset need not necessarily be the same as the interpretation of the contrast given the feature label [+ *nasal*] at a rime (see also [23] on the phonetic interpretation of 'alveolarity and plosion' in codas of English words). Moreover, the occurrence of the phonologically contrastive feature [+ *nasal*] at some point in the phonological structure may generalise over many more phonetic parameters than those having to do simply with lowering of the soft palate. (cf [24])

The consequence of this argument is that nothing at all hangs on the name of a phonological feature provided that the canonical naive view of the relationship between phonological categories and phonetic ones is eschewed. All that the 'naming of parts' achieves is some kind of mnemonic short hand. This means that provided the semantics of the phonological categories is explicitly and formally stated then it really doesn't matter what they are called. There are two aspects to specifying the semantics: (1) it is necessary to know how the phonological category(ies) in question relate to other phonological categories - that is provide a semantic statement of their place within the phonological systems and structures and (2) it is necessary to provide an explicit statement of the phonetic interpretation of the phonological categories because, in Firthian terms, it 'renews the connection' with the dynamic parametric phonetic data [11]. I will develop this position in the following section and show that we can construct a simple phonetic interpretation function which will relate non-segmental structures to a DNSP.

PHONETIC INTERPRETATION OF [ATR] IN KALENJIN

I will now use some data concerning the phonetic characteristics of the *[ATR]* harmony system in the Kalenjin to motivate an abstract non-segmental phonology and to show how such a phonology can be phonetically interpreted. The broad IPA transcriptions below give an impression of some of the phonetic exponents of *[+/-ATR]* in Kalenjin. *[+ATR]* words are given first for each pair:

1 [k̟ʰɛ̟ːβɪt̟ʰ] {to sprinkle}

[kʰɛ̟ːβɪt̟ʲʰ] {to grow}

2 [k̟ʰɛ̟ᵊgu̟ t̟ʰ] {to scrape up}

[kʰɛːɣ̟ u̟ʲʰ] {to blow}

3 [k̟ʰɛːβɑl̟] {to dig up}

[kʰɛ̟ᵊβ̟ a̟ l] {to dig}

4 [pʰe̟. ɲ] {meat}

[pʰɛɲ] {hardship}

5 [lo̟] {far}

[lʲɔ̟] {six}

There are a number of phonetic differences between words in the two categories. These occur not only in vocalic portions but also in the consonantal portions of such words. They include phonatory quality, vocalic and consonantal quality and articulation and durational differences.

Phonatory differences

The two sets of words exhibit different kinds of phonatory activity. Words of the *[-ATR]* set have audible breathy phonation as compared with words in the *[+ATR]* set. Measurements of the open quotient of the glottal cycle made from electrolaryngographic recordings and inverse filtering reveal (statistically significant) differences that can be taken to confirm breathiness of phonation (larger OQ values are found for *[-ATR]* words). Spectral characteristics of vocalic portions of the two classes also reveal differences commensurate with breathy versus non-breathy phonation. Examination of voice source measurements also suggest different kinds of laryngeal behaviour in moving from voice to voicelessness in the two sets of words. In *[+ATR]* voicing dies away slowly and continues at low level. In *[-ATR]* words, by contrast, voicing drops off rapidly.

Vocalic differences

There are striking auditory differences in vocalic quality between words in the two sets. Vocalic portions in *[-ATR]* words are noticeably more central (and frequently more open) than those in *[+ATR]* words. (Note that the open *[+ATR]* vocoid has a back [ɑ] quality in the region of CV5; the open *[-ATR]* vocoid has a *front* quality in the region of CV4 [a]. These vocoids harmonize with appropriate tokens from the *[ATR]* sets: thus [sɑmʲisʲ] ~ [sa̟ mʲɪsʲ], [tʰɑŋgusʲ] ~ [tʰaŋ̟gu̟sʲ].) Examination of plots of F1/F2 for tokens each of the *[±ATR]* vocoids in the data confirms the results of impressionistic listening (for example, *[+ATR]* vocoids show lower F1 values than their congeners *[-ATR]*).

Consonantal differences

Words of the two categories exhibit different types of stricture and ranges of variation in the consonantal portions. In *[+ATR]* words we final labial, apical and velar closure with burst release, or with close approximation. In comparable words which are *[-ATR]* closure with burst release is not found. In such words lax fricative portions occur but so do portions with open approximation. There are also noticeable variations in terms of place of articulation. 'Coronals' in *[+ATR]* words are exponed with apico-alveolar strictures whereas they may be exponed with either apico-alveolar or dental strictures in *[-ATR]* words.

Durational differences

Consonantal and vocalic portions are durationally different in *[±ATR]* words. Typically consonantal portions are shorter in *[+ATR]* words than they are in *[-ATR]* words. This is particularly noticeable in the closure and release phases of initial and final plosion. Averages of vocalic duration reveal a tendency for *[-ATR]* vocoids to be shorter than *[+ATR]* vocoids. However, *[+ATR]* words are routinely longer (measured from beginning to end of voicing) than are *[-ATR]* words.

COMPOSITIONAL PHONETIC INTERPRETATION

[ATR] harmony is canonically the kind of phonological organisation which has been given non-segmental status. Even a hard-core segmentalist would be likely to acknowledge that *[ATR]* in Kalenjin operates in terms of whole syllable structures. However, it is not immediately clear that extant phonological approaches (including eg autosegmental phonology and gestural phonology [16]) could deal in any coherent way with the phonetic interpretation of *[ATR]* here given the range of different phonetic exponents implicated. It would require a certain amount of ingenuity to postulate a non-segmental *[ATR]* feature with intrinsic phonetic content and find what there is in common between devoicing of coda approximants, breathy phonation, front or back secondary articulation, consonantal length, particular ranges of consonantal variability and any putative advanced position of the tongue root. Even greater problems might arise in making sense of the 'counter-intuitive' phonetic interpretation of the open *[+ATR]* vowel in the region of [ɑ] and the open *[-ATR]* vowel in the region of [a].

I suggest that a DNSP interpretation of the abstract phonological relationship designated *[±ATR]* is more appropriately accomplished with explicit statements of temporal and parametric phonetic exponency for various parts of word and syllable structure. This can be achieved by a compositional phonetic interpretation (CPI) function for partial phonological descriptions [19, 20, 21]. I sketch only the broad outlines of a CPI here.

In the CPI function phonological structures and features are associated with phonetic exponents. The phonetics is the semantics of the phonology [13, 19, 20] (cf [25]). As I indicated earlier, the phonological descriptions being interpreted are here taken to be unordered acyclical graph structures with complex attribute-value node labels. The statement of phonetic exponents in CPI has two formally distinct parts: temporal interpretation and parametric phonetic interpretation. Temporal interpretation establishes timing relationships which hold across constituents of a phonological graph while parametric interpretation instantiates interpreted dynamic 'parameter strips' for any given piece of structure (any feature or bundle of features at any particular node in the phonological graph). The resulting 'parameter strips' can be considered as sequences of ordered pairs where any pair denotes the value of a particular parameter at a particular (linguistically relevant) time. Thus in the general case:

{(node:partial_phonological_description,
(Time_start, Time_2, ... Time_end),
parameter section}

where the node represents any phonologically relevant contrast domain. The time values may be absolute or relative, fixed or proportional. The precise physical domain of the parameter strips (eg articulatory, acoustic, aerodynamic) is not of immediate relevance here.

The 'compositional' part of the interpretation function signifies that the 'meaning' of a complex expression is a function of the form and meaning of its parts and the rules whereby the parts are combined [26]. The phonological 'meaning' of a syllable equals the 'meaning' of its constituents. The compositional principle is instantiated by requiring any given feature or bundle of features at a given place in the phonological structure to only have one possible phonetic interpretation. So for instance, in the present case the words

(i) [k̟ʰʷo̟ːl̟], 'good planters' and

(ii) [kʰʷːo̟ːlː] 'plant!'

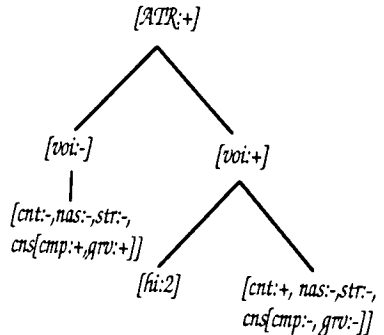
can be given the following Firthian like, partial representations:

(i) $^{[ATR+]}(\kappa o\lambda)$

(ii) $^{[ATR-]}(\kappa o\lambda)$

Here the syllable-domain *[ATR]* unit as well as being semantically distinctive serves to integrate the other syllabic material (paradigmatically contrastive units) with consequences phonetic exponency as illustrated above). Given this, then the interpretation of (i) is of the form: *CPI*([ATR:+] (κολ)) = {phonetic exponents of 'kol'}. A more fully specified representation of (i) might be given as: $^{[ATR+]}(^{f_1}(\kappa), ^{\neg f_1}(o,\lambda))$. Here the units within the syllable are treated as separate entities or sequences of entities.

The superscript symbols *ɦ* / *¬ɦ* placed before the units (κ) and (oλ) serve to indicate onset/rime domain contrasts (*ɦ* 'voicelessness'; *¬ɦ* 'voice'). Such a representation can be reconstructed as a graph with attribute-value node labels, thus:



A partial compositional interpretation of this schematic representation can be determined in the following quasi-articulatory fashion:

1. *CPI([cnt:-, nas:-, str:-, cns[cmp:+, grv:+]])* = {contact of tongue back with soft palate, closure of soft palate ...}
2. *CPI([hi:2])*={relatively mid tongue-height...}
3. *CPI([cnt:+, nas:-, str:-, cns[cmp:-, grv:-]])* = {contact of tongue apex with alveolar ridge...}
4. *CPI([voi:+]/[hi:2], [cnt:+, nas:-, str:-, cns[cmp:-, grv:-]])* = {partial overlap and succession of *CPI([cnt:+, nas:-, str:-, cns[cmp:-, grv:-]])* to *CPI([hi:2])*, relative length of *CPI([hi:2])*, relative slow decay of voicing of *CPI ([hi:2])*...}
5. *CPI([voi:-]/[cnt:-, nas:-, str:-, cns[cmp:+,grv:+]])* = {voicelessness, aspiration of *CPI([cnt:-, nas:-, str:-, cns[cmp:+,grv:+]])*...}
6. *CPI([atr:+]/[voi:-]/[cnt:-, nas:-, str:-, cns[cmp:+, grv:+]]), [voi:+]/[hi:2], [cnt:+, nas:-, str:-, cns[cmp:-, grv:-]]))* = {succession and partial overlap of *CPI([voi:-]/[cnt:-, nas:-, str:-, cns[cmp:+, grv:+]])* to *CPI([voi:+] ([hi:2], [cnt:+, nas:-, str:-, cns[cmp:-, grv:-]]))*, non-maximal backness of *CPI([voi:-] ([cnt:-, nas:-, str:-, cns[cmp:+, grv:+]])* and *CPI([voi:+] ([hi:2], [cnt:+, nas:-, str:-, cns[cmp:-, grv:-]]))*, relative palatality of *CPI([cnt:+, nas:-, str:-, cns[cmp:-, grv:-]])*, relative shortness of closure and release of *CPI([voi:-]/[cnt:-, nas:-, str:-, cns[cmp:+,grv:+]])*, tense phonatory quality and slow decay of voicing of *CPI([voi:+]/[hi:2], [cnt:+, nas:-, str:-, cns[cmp:-, grv:-]])*, ...} ...

We have formally tested and verified a CPI for Kalenjin within the YorkTalk declarative speech generation system employing acoustic parameters. Discussion and illustration of this work and quantitative details of the phonetic exponents of [*ATR*] in Kalenjin are given in Local and Lodge [27].

CONCLUSION

Recent phonetic work in laboratories across the world has provided a rich diet of DNSP data. Rather than reviewing this work I have chosen here to concentrate on isses surrounding the interpretation of DNSP data. I have done this because it seems to me that whilst we have seen considerable advances in data collection techniques (eg in the articulatory domain [28]) there has not been a commensurate advance in the linguistic interpretation of that data. By examining a small amount of material from Kalenjin I have tried to motivate the need for a consideration of non-segmental phonological categories in the interpretation of phonetic data. I have suggested that a small step in this direction can be achieved we if adopt a non-segmental phonology of the Firthian kind and reject analysis in terms of intrinsic phonetic interpretation. Such a step obliges us to devise an explicit phonetic interpretation function and to explore ways in which DNSP data might relate to abstract non-segmental categories. I think it also moves us towards the 'integrative phonology' for which Ohala argued so persuasively at ICPh91 [1].

REFERENCES

[1] Ohala, J. The Integration of phonetics and phonology. . *Proc. ICPhS91*. Aix-en-Provence. Vol. 1. 2-16.

[2] Nolan, F. (1991), Phonetics in the next ten years. *Proc. ICPhS91*. Aix-en-Provence. Vol. 1. 125-129.
[3] Perkell, J.S. (1991), Models, theory and data in speech production. *Proc. ICPhS91*. Aix-en-Provence. Vol. 1. 182-191.
[4] Laver, J. (1994), *Principles of phonetics*. Cambridge: Cambridge University Press.
[5] Bell-Berti, F. (1991), Comments on "Some observations on the organisation and rhythm of speech". *Proc. ICPhS91*. Aix-en-Provence. Vol. 1. 238-242.
[6] Browman, C.P. & Goldstein L.M. (1985), Dynamic modelling of Phonetic Structure. In V. Fromkin (ed). *Phonetic Linguistics*. New York: Academic Press. 35-53.
[7] Repp, B.H. (1981), On levels of description in speech research. *JASA*, 69(5): 1462-1464.
[8] Barry, M.C. (1991), Temporal modelling of gestures in articulatory assimilation. . *Proc. ICPhS91*. Aix-en-Provence. Vol. 4. 14-17.
[9] Byrd, D. (1993), Palatogram Reading as a Phonetic Skill: a short tutorial. *JIPA*, 23, 2: 59-72.
[10] Kelly, J. (1987), On the phonological relevance of some 'non-phonological' elements. *Magyar fonetikai füzetek* 21: 56-59.
[11] Firth, J.R. Sounds and Prosodies. *Transactions of the Philological Society*, 129-152.
[12] Sprigg, R.K. (1966), Phonological formulae for the verb in Limbu as a contribution to Tibeto-Burman comparison. In C.E. Bazell et al (eds). *In Memory of J.R. Firth*. London: Longmans. 431-453.
[13] Kelly, J. & Local, J.K. (1989), *Doing Phonology*. Manchester: Manchester University Press.
[14] Keating, P.A. (1990), Phonetic representations in a generative grammar. *Journal of Phonetics* 18, 3: 321-334.
[15] Scully, C. (1987), Linguistic units and units of speech production. *Speech Communication*, 6: 77-142.
[16] Saltzman, E.L. & K.G. Munhall, (1988), A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, 1: 333-382.
[17] Browman, C.P. & Goldstein, L.M. (1989), Articulatory gestures as phonological units. *Phonology*. 6.2: 201-251.
[18] Fujimura, O. (1994), The syllable: its internal structure and role in prosodic organisation. In *Proc. LP'94: Item order in (natural) languages)*. B. Palek (ed). Institute of Linguistic and Finno-Ugric Studies, Charles University: Prague
[19] Local, J.K. (1992). Modelling assimilation in non-segmental rule-free synthesis. In Docherty, G. & Ladd, R. (eds.) *Papers in laboratory phonology II*. Cambridge: CUP. 190-223.
[20] Coleman, J. & Local, J.K. (1992), Monostratal phonology and speech synthesis. In P. Tench (ed) *Studies in systemic phonology*. Pinter Publishers: London. 183-193.
[21] Ogden, R. (1992). Parametric interpretation in YorkTalk. *York Papers in Linguistics* 16, 81-99.
[22] Clements, G.N. (1985), The geometry of phonological features. *Phonology Yearbook*. 2. 225-252
[23] Manuel, S.Y., Shattuck-Hufnagel, S., Huffman, M., Stevens, K.N., Carlson, R & Hunnicut, S. (1992), Studies of vowel and consonant reduction. *Proceedings of the International Conference on Speech and Language Processing*. Volume 2, 943-946.
[24] Ladefoged, P. (1977), The abyss between phonetics and phonology. In *Proceedings of the 13th meeting of the Chicago Linguistic Society*. 225-235.
[25] Beckman, M. (1990),Metrical structure versus autosegmental content in phonetic interpretation. *Proc. ICPhS91*. Aix-en-Provence. Vol.1. 374-378.
[26] Partee, B.H. (1984), Compositionality. In F. Landman & F. Veltman (eds). *Varieties of Formal Semantics*. Dordrecht: Foris. 281-312.
[27] Local, J.K. & Lodge, K. (to appear) On the phonetic interpretation of [ATR] in Kalenjin. *York Papers in Linguistics*.
[28] Perkell, J.S. Cohen, M.H., Svirsky, M.A., Matties, I.G. & Jackson, M.T.T. Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. *JASA*, 92(6): 3078-3096.

PROSODIC ORGANIZATION OF SPEECH BASED ON SYLLABLES: THE C/D MODEL

Osamu Fujimura
Department of Speech & Hearing Science, The Ohio State University
Columbus, OH 43210-1002, USA

ABSTRACT

A syllable-based theory of phonetic implementation called the C/D model is reviewed, with remarks on its phonetic implications regarding prosodic control. The phonological feature specification assumed at the input is discussed, in connection with the underspecification scheme. A recent revision of the timing computation scheme accounts for some prosodic effects on temporal behavior of articulatory gestures for English /l/.

C/D MODEL

This paper discusses a new view of speech organization: the Converter-Distributor (C/D) model of phonetic implementation [Fujimura *et al.* 1991; Fujimura 1992, 1994a,b, in press]. It considers the prosodic organization of an utterance as the basic framework for describing the speech production process. A prosodic structure, represented by a metrical tree (see Liberman & Prince [1977]), is assumed with a phonetic augmentation for specifying utterance conditions. The prosodic structure is interpreted as a linear string of syllables and boundaries with varied magnitude values.

The flow of vocalic gestures characterizing the sequence of syllable nuclei forms the *base function* of the articulatory events that fit in the prosodic structure of the utterance. On this base function, consonantal gestures are superimposed, basically in the way Öhman [1966] depicted in his consonantal perturbation model. The base function is inherently multi-dimensional in the sense that different articulatory variables such as jaw opening, tongue body advancing or retraction, lip rounding and protrusion, and pulmonary and laryngeal conditions, behave more or less independently from each other. Prosodic effects are implemented mainly by mandibular, laryngeal, and pulmonary variables. Vocalic gestures are implemented in tongue body position and lip rounding dimensions, which physically interact with mandibular position, and represent a continuous flow of inherent articulatory gestures for unreduced syllable nuclei, constituting one aspect of the base function. The implementation process of vocalic and intonational aspects of the base function may be somewhat similar to existing acoustic models of F0 contours such as Pierrehumbert's [1980] or Fujisaki's [1988]).

To the extent that speech organization is described in terms of articulatory gestures, the C/D model is similar to the articulatory phonology proposed by Browman and Goldstein [1992]. There are many phonetic observations, particularly allophonic variations of phonemes in the traditional segmental description, usually expressed as context-sensitive rewrite rules in generative phonology, that are naturally explained by either theory as the consequence of using an assembly of autosegmental articulatory gestures in variable timing relations. Such variation is typically sensitive to the style of utterance, among other factors.

These two theories, however, basically differ from each other. While articulatory phonology assumes gestures to be the basic units in the lexical phonological representations, integrating everything together from lexical phonology to phonetic signal generation, the C/D model strictly respects the traditional distinction between phonology and phonetics. The phonetics, however, is strongly sensitive to the particular language or dialect, and it also handles abstract features until gestures are concretized at the output of the actuators. What was called the base of articulation in the traditional British literature, for example, is incorporated into the system

parameters that prescribe the signal generator design. The phonological structure as the input to the model reflects the lexical specifications and the syntagmatic organization of phonological phrases. At the same time, the numerical specifications attached to any node of the metrical tree produces prominence of the pertinent part of the tree structure, and additional numerical specifications of utterance characteristics including the speaker's habit, determine system parameters for the entire utterance, according to the situation of speaking.

The C/D model describes the phonetic implementation process, apart from the signal generator, in three sequentially ordered system components: converter, distributor, and a parallel set of actuators. The process is inherently multidimensional and superpositionally linear until the set of control time functions are derived. The signal generator, which takes these control functions as its input, is a complex, highly nonlinear and inherently three-dimensional dynamic system [Wilhelms-Tricarico, in press].

The C/D model uses syllables as the basic units of segmental materials that are concatenated into a temporal linear string, intervened by phonetic phrase boundaries. The latter can be empirically observed in articulatory movement patterns, as discussed in Fujimura [1990]. The prosodic structure of an utterance is represented completely, at one level of the phonetic representation, by a series of magnitude-specified pulses. The timing pattern of the series of abstract events for syllables is then derived from the magnitudes (abstract phonetic strengths) of syllables and boundaries.

It is emphasized that the signal generator component, as the last and physical stage of the model, determines critical characteristics of directly observable physical phenomena such as articulatory movement patterns, and based thereon, acoustic or spectrographic patterns, including durations of acoustically defined speech segments. The input to the signal generator may be interpreted to represent basically motor control time functions given to the physiological apparatus for speech production.

The prescription of articulatory gestures in the form of control functions is generated by the set of actuators, the third component of the model. These control variables as time functions can be significantly different from directly observable physical signals, whether articulatory or acoustic. Nevertheless, we claim that the model's general validity can be tested and its parameter values can be inferred, by evaluating physical signal characteristics, if powerful computational techniques are used to handle a large mass of data for inference of the underlying variables.

The basic assumption is that, however complex (with feedback loops, *etc.*) the signal generating system may be, it has a fixed physical design, containing only parameters that are sensitive to the speaker's conditions. In contrast, the process up to the output from the actuators including the table of impulse response functions for consonantal gestures, are parametrically sensitive to the language or dialect spoken.

The classical theory of generative phonology (see Chomsky & Halle [1968]), assuming a level of systematic phonetic representation, ascribes the switching from discrete specifications in phonology to continuous and numerical variable specifications in phonetics to additional suprasegmental variables like segmental duration and tonal inflection, while assuming a large number of phonetic segments (allophones) resulting from detailed but discrete alterations of articulatory states. This can not account for the intricate interaction between articulatory or acoustic gestures of individual consonants or vowels and prosodically conditioned suprasegmental parameters, including variable strengths of phonetic boundaries (see Fujimura [1970]). The continuous nature of phonetic phenomena stems not from the superimposed properties of individual segments, but from the inherently multidimensional nature of the articulatory organization interacting with prosodic conditions. Therefore, the "segmental" characteristics themselves continuously vary.

The C/D model seems to have the potential to account for much of the observed allophonic variation, whether coarticulatory or not, within the phonetic implementation process, according to a general phonetic principle combined with language-specific system parameters. The feature specifications are passed by the converter to the distributor for phonetic gesture specifications. Many apparently supplemental specifications of redundant information are automatically provided by the speech production process itself. For example, unspecified vocalic gestures for reduced syllables in English, can be left unspecified throughout the phonetic process, and computed by the signal generator as continuous time functions, according to the base function control.

Likewise, the place specification for the nasal segment in English coda when combined with a tense obstruent (*e.g.* in 'tent', 'tense', 'camp', 'honk') is not phonologically copied from the stop segment specification, but is implemented as a single articulatory oral closure gesture spanning over the nasal (lowered velum) and oral (raised velum) portions of the coda. In contrast, when an obstruent is voiced and follows a nasal consonant, as in 'lens', 'tend', 'sums', 'songs', *etc.*, the syllable-final voiced obstruent is always apical (alveolar or dental), and the place is specified for the nasal consonant. The final obstruent in such a situation (along with the final voiceless apical obstruent in an obstruent sequence such as 'act' and 'opt') are separated out from the syllable core as a syllable suffix (s-fix), based on the general rule of English syllables that a syllable-final apical obstruent that agree in voicing with the tautosyllabic obstruent in the coda is separated as a s-fix (Fujimura [1979]).

As the first component of the model, the converter's role is to evaluate the prosodic pattern as specified in the augmented metrical tree, to compute the phonetic strength of each syllable, and accordingly, to assign a magnitude value to each impulse that represents the syllable. The converter also creates a boundary pulse by evaluating the tree configuration, and assigns the magnitude value to each boundary pulse. Based on the series of magnitude-specified syllable-boundary pulses, the converter computes time intervals between contiguous pulses by an algorithm which is called a shadow computation (see below).

At the input level for the converter, utterance conditions such as speed of utterance, formality of utterance, and speaker idiosyncrasy (in multidimensional measures) are numerically specified. These affect the pulse train *via* adjustment of shadow slopes. This pulse train functions as the total prosodic control of the utterance (to the extent that the current approximation is effective) and determines the non-uniform temporal overlapping of gestures in each articulator. It should be noted, however, that the syllable type (heavy *vs.* light syllables, *etc.*), as a phonological property of the syllable represented by the feature specifications, controls the shadow coefficients, which affect the time intervals between contiguous syllable-boundary pulses (see Fujimura [1994a]). The numerical augmentation of a tree node for prominence, as an utterance specification, does not affect the shadow slopes.

The distributor interprets the feature specifications to distribute corresponding elemental gestures to pertinent articulatory dimensions to be implemented by specific articulatory organs, generating elemental gesture specifications for the next component, a parallel set of actuators. The parallel set of actuators generate time functions by exciting pertinent IRFs by the syllable pulse, which determines timing and amplitude of each IRF. Different IRFs are then superimposed in each dimension to form the time function of the articulatory control for phrasal units.

FEATURE SPECIFICATION

The syllable structure analysis in the C/D model adopts the principle of demisyllabic analysis [Fujimura 1976, 1979; Fujimura and Lovins 1978], that consonant clusters (in English) do not require any ordering specification within the syllable core, after separating out syllable affixes. This principle

recognizes CVC as the canonical syllable structure of English syllables, where the initial C can be zero, but the final C is mandatory unless the syllable is reduced as a supplement to the head (with stress) of a foot (there may be more than one such subordinate syllables). Tense vowels and diphthongs in English are treated as a combination of a vowel (V) and a glide (C). The syllable affix to the left of the core is called a p-fix (not applicable to English) and that to the right is called a s-fix.

The C in onset and coda, optimally an obstruent, marks the edges of the syllable core, to which a s-fix (or order-specified string of s-fixes) can be attached, when certain strong constraints are met for each consonant to qualify for the status of a s-fix. The p-fixes are similar in a mirror-image situation. The basic assumption is that within the core, in either onset or coda, no sequential ordering of features is given. Therefore, feature specifications, including sonorant features, for either onset or coda, are given as a set (not sequence) of several privative feature specifications, which may be divided into concomitant feature types (such as place and manner). The temporal organization of tautosyllabic articulatory gestures automatically emerges as the inherent properties of the evoked IRFs.

For this principle to work in English, it is critical to assume an abstract feature called {spirantized}, representing the combination of apical frication and oral closure in the phonemic consonantal sequences /sp/, /st/, and /sk/ in both initial and final position. This feature is an obstruent feature, as a member of the manner feature paradigm opposing it to {stop}, {fricative}, {interdental} and {nasal}. The features {spirantized}, {stop}, and {nasal} all require a place specification and are implemented with an oral stop closure (the place-specified closure is delayed for {spirantized} relative to the frication production according to the pertinent IRF properties). This feature also corresponds to the same phonemic sequences in the coda (*e.g.* 'task' /tæsk/ as opposed to 'tax' /tæk.s/ which contains a s-fix outside the core, as indicated by a dot in the phonemoidal transcription).

It should be mentioned here that in English, there are many syllabic sonorants (as in 'button', 'bottle'), that must be treated as separate syllables, even though, phonetically, there is no vowel. These are not s-fixes, since they do not satisfy the requirement for s-fixes that the voicing status must agree with that in the coda. Japanese also has many cases of phonetically nonexistent (or devoiced) high vowels. These syllables contain vocalic specifications which cause a minimal distinction between /i/ and /u/ in the devocalized environment. In addition, these hidden vowels always show up when the intonation pattern requires a raised pitch, as observed toward the end of a question sentence.

One critical problem in connection with the discussion of possible syllable structures is how to define syllables as abstract phonological units. Before we discuss where syllable boundaries are in polysyllabic forms, we will first be concerned with the existence of syllables, identifying syllable nuclei which may not be phonetically apparent. Some guiding principles in identifying phonetically hidden syllables may be formulated as follows.

(1) A syllable must have at most one continuous stretch of voiced portion in the phonetic signal. If a word manifests itself with an unvoiced portion surrounded by voiced portions on both sides, there must be assumed more than one syllable. Thus the sonority principle (see Clements [1989] and Fujimura [1989]) with respect to phonetic voicing should be observed with the strongest priority (at the top of the constraint hierarchy in the sense of optimality theory, see Prince & Smolensky [in press]), and probably universally (as an absolute requirement).

(2) Consonant clusters at the left and right edges of a phonological word often contain syllable affixes, which are often but not always morphological affixes. The separable affixes (p-fixes and s-fixes) must be strongly limited in phonological feature specifications, and the phonetic voicing status continuously spreads from the onset (backward) or coda (forward) toward the word edge,

thus requiring no feature specification for voicing in affixes. If there is a change in voicing at a syllable edge, as in German initial /kn/ (in 'Knabe') and English final /nt/ (in 'tent'), the two consonantal elements must be both contained within the syllable core. There is strong phonetic evidence that a phonemic minimal pair like /tent/ and /ten.d/ must be treated differently (see Fujimura & Lovins [1978]).

In this situation, it is likely that one of the consonants as a phoneme has no paradigmatic opposition in place. In English, the final phonemic sequence nasal + voiceless obstruent must be homorganic. In German, for example, /km/, is not allowed, and therefore, the specification for the /n/ element in the cluster /kn/ is {nasal} without any place specification. In English, it can be shown that at most one place specification is allowed for the onset or coda, and none is given for s-fixes. Note again that the feature set {spirantized, labial} for the English words 'spoon' or 'grasp', for example, does not specify the place for /s/.

Likewise, the feature {lateral} in English, does not have any place specified in our analysis, allowing a distinction between 'slight' and 'flight', for example, with only one place specification for the onset (*cf.* 'smell' *vs.* 'snell' for which the place specification is for the nasal element, not for the /s/, reflecting the distributional fact that there is no opposition /s/ *vs.* /f/ in this onset environment). The feature {lateral} automatically evokes the apical gesture for an alveolar contact in onset position as an elemental gesture, by looking up a feature-gesture table. It evokes a similar coronal gesture in coda in some dialects of American English but not necessarily. The most robust inherent gestures seem to be tongue blade narrowing and body retraction (see Sproat & Fujimura [1993]). Therefore, at least in coda, the lateral in English cannot be specified with a place feature from a phonetic point of view.

While the C/D analysis treats onset and coda gestures basically as independently assigned gestures, it does use the same feature name vocabulary in onset (with a superscript *o* as needed) and in coda (with *c*).

(3) When more than one p-fix or s-fix is allowed in the language (as in English for s-fixes, *e.g.* in 'sixths'), ordering of feature specifications for the sequence of affixes is required. The inventory of phonemic segments treated as syllable affixes must be small, and their feature specifications are given parsimoniously (only a manner specification in English). Apart from this paradigmatic parsimony, affixes behave like phonemes: they form a temporal string with specified sequential ordering. They are phonetically very stable, allowing, for example, segmental waveform concatenation.

Syllable affixes (p-fixes or s-fixes) may be assumed to occur only at the edges of phonological words (or morphemes).

In the C/D model, unlike the earlier demisyllable analysis [Fujimura 1976, 1979], vowels are treated separately from consonants throughout the computational process, from the feature specification level (*i.e.* input to the converter) to the control time function level (*i.e.* input to the signal generator). For this reason, the demisyllable approach is adopted in the C/D model only with respect to consonantal features and gestures.

A minimal underspecification scheme by means of privative (unary) features is used for the input representation in the C/D model. For example, in English, the first syllable /skramp / of 'scrumptious' has an onset specified as {dorsal, stop, spirantized, rhotacized}, and a coda {labial, nasal, stop} (no meaningful ordering of features intended). The voicelessness for onset or coda is not specified because obstruent manner features without {voiced} implies unvoiced, implemented as a voice cessation (vocal fold abduction) at the edge of the syllable.

CONSONANTAL TIMING

According to the original first-approximation scheme presented in previous publications, the C/D model specifies that the internal timing relation between the initial and final gesture peaks

will remain the same regardless of the magnitude of the syllable. Therefore, when the syllable is reduced, other things being equal, the duration of the acoustic vowel segment probably will increase when the margins are unvoiced, because the glottal abduction gesture will be reduced and the duration of the vocal-fold vibration will be expanded.

This particular difficulty could be resolved by assuming that the default condition for voicing was unvoiced, as in a pause, and each syllable pulse evokes in the laryngeal dimension an adduction gesture (as opposed to the scheme where voicing is the basic gesture unless obstruent features or phrase boundary features evoke voice cessation, *i.e.* laryngeal abduction). The observed voiced duration in the acoustic signal then would depend on the characteristics of the signal generator, balancing the durations of the voiceless consonants and the vowel portion in the resulting acoustic signal as dictated by the nonlinearity of the production mechanism. The vowel elongation due to syllable reduction is, of course, counterfactual, while the shortening of consonantal segments is factual. Which approach is more nearly correct as an approximation theory is an empirical issue. In either case, the total syllable duration, or more exactly, the time interval between contiguous syllables as represented by the syllable pulses, is distinctly shorter (proportional to the syllable pulse magnitude) when the syllable is weak.

An alternative general solution of this problem and some others can be provided by specifying a little more detail of the mechanism that evokes IRFs, without affecting the principle that all prosodic structure of speech articulation is computed *via* the time and magnitude evaluation of the syllable and boundary pulses. The current idea is as follows.

A syllable pulse generates separate pulses for the onset, the coda, and each of the affixes. Each of these subsidiary pulses (which may be called *pocs* pulses, standing for p-fix, onset, coda, and s-fix) evokes the IRFs. The pocs pulses inherit the magnitude of the parent syllable pulse. Pocs pulses have shadows which extend only outward from the syllable pulse, the center being the syllable pulse under discussion which excites vocalic and prosodic gestures directly.

The onset pulse is erected at the external (*i.e.* left) end of the left-hand shadow of the syllable pulse, and the coda pulse is erected at the external (right) end of the right-hand shadow of the syllable pulse. The most internal p-fix requires a p1-pulse erected at the external end of the shadow of the onset pulse, and the next external p2-pulse stands at the external edge of the p1-pulse. The s1-pulse, s2-pulse, *etc.* for the s-fixes are similar, forming a mirror image. The most external edge of the shadows to the left or right of the most external component of the syllable determines the temporal limit of the syllable in question as a whole, and the contiguous syllable or boundary pulse is placed to make this limit coincide with its associated external temporal limit, *i.e.*, the edge of its most external pulse shadow.

The pocs pulses generally delimit the time domain in which articulatory gestural activities of the pertinent syllable component (p-fix, onset, coda, or s-fix) are primarily contained. The syllable pulse covers primarily the vocalic activities corresponding to either vocalic or consonantal features, but tense consonantal gestures tend to invade into this vocalic time interval. Note that the IRFs are continuous time functions and never exhibit any sharp boundaries for activities. The segmental discontinuity as observed in the acoustic signals arise due to nonlinearity of the signal generating process. The onset pulse is the pulse that triggers IRFs of onset elemental gestures, the values of the IRFs being subdued and their acoustic effects tending to be invisible beyond the shadow edges particularly if any gesture of the next external component (tautosyllabic or heterosyllabic) manifests predominating effects. The most internal s-fix pulse (S1) marks the nominal end of the coda gesture activities, and the next external s-fix pulse (S2) marks the nominal end of the internal s-fix. The p-fix situation is a mirror image.

Some readers may find it awkward to see a response of the triggering pulse

temporally before the latter occurs: the IRFs we consider in this description are not physically realizable. This is a matter of convenience of the description. There is always a considerable delay between the cortical motor control planning and the physical execution, even if we take this model to represent a direct computational simulation of the physiological process of speech production, the actual triggering pulses must occur well ahead of the hypothetical time values of syllable or boundary pulses. Shifting the time values of all pulses by a sufficiently large and universally fixed time interval as a constant delay of responses resolves this seeming contradiction.

This revised scheme of timing computation using pocs pulses makes the intrasyllabic temporal relation between initial and final consonantal gestures more directly sensitive to the syllable pulse magnitude in general. Also, since the slopes of shadows for syllables are sensitive to the internal structure, reflecting the syllable type, the apparent vowel duration may vary not only reflecting the prominence condition and speed of utterance, but also whether the syllable is specified for a long or short vowel. In our analysis, a phonologically long vowel is specified with a "monophthongal" glide *i.e.*, the elongation feature {long^c}), and a diphthongal vowel with a more conventionally recognized glide ({palatalized^c}, *etc.*).

One distinct advantage of this pocs approach is the differential treatment of vocalic gestures from consonantal gestures. As Sproat & Fujimura [1993] pointed out, lateral and nasal consonants exhibit different intrasyllabic timing behaviors between what may be considered vocalic vs. consonantal gestures, while, phonologically, both {nasal} and {lateral} are consonantal manner features. Vocalic gestures, such as velum lowering and tongue body retraction, seem more closely linked to the center of the syllable, while tongue tip or lip gestures are linked to the margins of the syllable. Assuming that vocalic gestures are evoked by the syllable pulse while consonantal gestures are evoked by the onset or coda pulse, the correlation between the relative timing difference between the consonantal and vocalic gestures to prosodic conditions as observed in the articulatory studies (see also Krakow [1989] may be accounted for by a general phonetic principle as prescribed by the C/D model.

There are many additional details of the model that have to be worked out. The comparison of prediction with observation is not easily achieved, but has to be approached step by step in successive approximation comparing data and the updated tentative descriptive framework for interpreting data. The signal generator brings the generative description of this theory closer to direct modeling of the speech production process.

ACKNOWLEDGMENT

The author wishes to express his gratitude for the support he has been given by the ATR Interpreting Telephony Research Laboratories and the ATR Human Information Processing Research Laboratories. It is also acknowledged that this article is based in part on a longer and earlier manuscript for the Proceedings of LP'94 -- Item Order in Natural Languages, edited by B. Palek and P. Janota, to be published from Charles University, Prague, with the editors' permission.

REFERENCES

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

Chomsky, N. & Halle, M. (1968). *Sound Pattern of English.* Cambridge MA: MIT Press.

Clements, G.N. (1989). The role of sonority cycle in core syllabification. In J. Kingston and M.E. Beckman (eds.), *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech*, (pp. 283-333). Cambridge, UK: Cambridge University Press.

Fujimura, O. (1970). Current issues in experimental phonetics. In Jakobson, R. & Kawamoto, S. (eds.) *Studies in General and Oriental Linguistics* (pp. 109-130). Tokyo: Tec Co.

Fujimura, O.(1976). Syllable as concatenated demisyllables and affixes. *J. Acoust. Soc. Am.*, **59**, Suppl. 1, S55.

Fujimura, O. (1979). An analysis of English syllables as cores and affixes. *Z. f. Phonetik, Sprachwiss. u. Komm.*, **32**, 471-476.

Fujimura, O. (1989). Demisyllables as sets of features: Comments on Claimants' paper. In J. Kingston and M.E. Beckman (eds.), *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech* (pp. 334-340). Cambridge, UK: Cambridge University Press. Fujimura, O. (1990). Methods and Goals of Speech Production Research. *Language and Speech*, **33**, 195-258.

Fujimura, O. (1992). Phonology and phonetics- A syllable-based model of articulatory organization. *J. Acoust. Soc. Japan* (E), **13**, 39-48.

Fujimura, O. (1994a). Syllable timing computation in the C/D model. *Proceedings of the Third International Conference on Spoken Language Processing, Yokohama*.

Fujimura, O. (1994b). C/D model: a computational model of phonetic implementation, in E. Ristad (ed.), *Language Computations* (pp. 1-20). Providence, RI: Am. Math. Soc.

Fujimura, O. (in press). The syllable: its internal structure and role in prosodic organization, Palek, B. & Janota, P. (eds.), *Proceedings of LP'94: Item Order in (Natural) Languages.* Prague: Charles Univ.

Fujimura, O. & Lovins, J. (1978). Syllables as concatenative phonetic units. In A. Bell and J.B. Hopper (eds.), *Syllables and Segments* (pp. 107-120). Amsterdam: N. Holland.

Fujimura, O., Erickson, D. & Wilhelms, R. (1991). Prosodic effects on articulatory gestures-a model of temporal organization. *Proceedings of XIIth International Congress of Phonetic Sciences, Aix en Provence.* Vol. 2, 26-29.

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* (pp. 347-356). New York: Raven Press.

Krakow, R.A. (1989). *The Articulatory Organization of Syllables: A Kinematic Analysis of Labial and Velar Gestures.* PhD diss. Yale University.

Liberman, M. & Prince, A. (1977). On stress and linguistic rhythm. *Linguistic Inquiry*, **8**, 249-336 .

Öhman, S. E. G. (1966). Coarticulation in VCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America*, **39**, 151-168.

Pierrehumbert, J. B. (1980). *The Phonology and Phonetics of English Intonation.* PhD Dissertation, MIT (distributed by the Indiana University Linguistics Club).

Prince, A. & Smolensky, P. (to appear). *Optimality Theory,* Cambridge MA, MIT Press.

Sproat, R. & Fujimura, O. (1993). Allophonic variation in English /l/ and its implications for phonetic variation. *J. Phonetics* **21**, 291-311.

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

ARGUMENTS FOR A NONSEGMENTAL VIEW OF SPEECH PERCEPTION

Sarah Hawkins, Department of Linguistics, University of Cambridge, U.K.

ABSTRACT

Systematic acoustic variation reflects vocal tract dynamics; it provides the acoustic coherence that makes a signal sound like speech. It is thus basic to speech perception, defined as lexical access. Implications of this argument are that the perceptual system maps an informationally-rich signal directly onto lexical forms that are structurally rich, and that phonemes are not essential for lexical access. Some properties of such a view of speech perception are discussed.

1 INTRODUCTION

In this paper, I argue that speech perception takes place by reference to a mainly nonsegmental phonetic structure. I discuss first some obvious shortcomings of the standard view of phonetic structure, in which prosody and a linear sequence of phonemes form two largely separate strands. Next, I argue that models of phonetic structure and of perception should include detailed information about the dynamics of vocal tract behaviour, since these details contribute coherence and systematic information to the signal. Finally, I outline the main properties I think a nonsegmental phonetic model of speech perception should have.

2 THE STANDARD VIEW OF PHONETIC STRUCTURE

The standard view of phonetic structure is of a linear sequence of so-called segments superimposed on a rather independent prosodic strand. Most people acknowledge that this view is a vast oversimplification, but nevertheless it underlies almost all the most influential phonetic models of speech production and perception. Explanations of the relationship between this abstract picture and reality are vague. Relationships between segments and prosody are poorly understood and not well studied: the two tend to be analysed separately, even though we know they are really not separable. Timing, for example, contributes crucially to both segmental identity and prosody. And formal relationships between these phonetic and phonological constructs and the other constructs of linguistics, such as grammar, are almost nonexistent.

Segments, for most people, seem to be closely tied to phonemes, even though, as I understand it, the term segment is typically used precisely to avoid the term phone or phoneme. At its least theoretical, a segment is an 'acoustic segment', i.e. that part of the acoustic signal that corresponds most closely to the 'main properties' for a particular phoneme.

Segments that are easiest to identify have abrupt acoustic boundaries. Many correspond to changes in excitation source, and/or to spectral steady states. They are usually clearly visible in spectrograms: turbulence noise of fricatives, silence associated with oral stops (often with a noise transient), periodicity of sonorants, the steady states (and sometimes transitions) of vowels, nasals, prevocalic /l/, and so on.

Everyone acknowledges that these acoustic segments are not phonemes, or even phones. But there seems to be a willingness to let the relationship between the two remain murky, partly perhaps because the linear model is so neat, and, in these clearcut cases, there *is* a strong connection between acoustic segments and phonemic identity. However, the term 'segment' is extended to other sounds as well: /w/, /j/, various types of /r/, and postvocalic /l/ are also segments, and we say that they are 'hard to segment' because their boundaries are only arbitrarily distinguishable from those of neighbouring 'segments'.

Descriptions of coarticulation are also often vague about how it arises, although coarticulation is integral to recent models e.g. Articulatory Phonology, and [1].

Intonation has tended to be seen as having the opposite problem. The challenge has been not so much to find the acoustic correlates of a predefined set of discrete units in the more continuous, measurable f0 contour, as to establish what the discrete units should be.

In reality, acoustic correlates of linguistic units are typically complex, spread over relatively long sections of the signal, simultaneously contribute to more than one linguistic unit, and do not cluster into discrete bundles.

3 COHERENCE & SYSTEMATIC VARIATION IN SPEECH

Models of human speech perception have typically incorporated linguistic-phonetic constructs fairly uncritically. Thus they assume that the main challenge is to map the acoustic signal onto the discrete sequence of segments that correspond to phonemes. Intermediate stages such as distinctive features may or may not be included, and prosody has usually been neglected. The nondiscrete nature of the signal has been ignored or seen as a problem of noise (but cf. [2]). If we were to take the opposite approach, and use what we have learned about speech and speech perception to help define the properties that a model of phonetic structure should have, we might come up with something rather different.

3.1 Natural speech

When humans speak, there is a tight relationship between the behaviour of the vocal tract and the acoustic properties of the emitted sounds. Thus natural speech is acoustically coherent: it contains all sorts of acoustic-phonetic fine detail that reflects vocal tract behaviour. This fine detail, and consequent acoustic coherence, is found in all aspects of speech. For example, it is found in correlations between the mode of glottal excitation and the behaviour of the upper articulators, especially at abrupt segment boundaries; in the amplitude envelope governing, for instance, perceptions of rhythm and of 'integration' between stop bursts and following vowels; and in coarticulatory effects on formant frequencies. (Other modalities, such as vision, can also contribute coherence, as I discuss briefly below. For simplicity, I restrict the present discussion to the acoustic signal.) All these types of effect contribute to acoustic variability. But the variation they contribute is *systematic*, or *lawful* variation, and adds information rather than noise to the signal.

We could say that systematic variation will only be called variation if we are bound to a view of speech as a linear sequence of phonemes that have a canonical, or pure, form, with clearcut temporal boundaries, and in which phonemes, excitation source, and prosodic variables are thought of as independent. These conceptually distinct strands are not separable in reality, and although there are reasons within linguistic theory to analyse them separately, maintaining a rigid separation may unnecessarily distort our thinking about speech perception and synthesis.

There is evidence from at least three fields of enquiry that coherence (or naturalness) of the speech signal is crucial to speech perception: from auditory psychophysics of the way the auditory system organises sounds into patterns; from speech synthesis by rule; and from speech perception itself.

3.2 Auditory psychophysics

Experiments show that when sounds have certain temporal and spectral relationships to one another, humans group them so that they form coherent patterns, such as alternating single notes and chords, or particular rhythms of a single tone. This phenomenon is called auditory streaming [4]. An auditory stream is perceived as coming from a single source. To use an older term from psychology, the sounds form a Gestalt. To cohere as an auditory stream, the sounds must be somewhat similar in frequency and timbre: a rhythmically-structured series of tones that differ greatly in pitch, say, is more likely to be heard as two independent streams. Changes in frequency or temporal relationships can drastically change the percept. Depending on the change made, the sounds may be heard as another pattern in the same stream, or they may break into a different number of streams, each with its own pattern, or as a chaos of unrelated events. In short, whether or not a time-varying signal (like speech) is heard as coming from a unitary source depends on tight spectral and temporal relationships between its various events. An example in speech is that we use the continuity of f0 to distinguish simultaneous vowels from each other [5].

3.3 Speech synthesis by rule

Those who work with synthetic speech have experienced the sense of incoherence due to inappropriate changes

in, e.g. f0 or amplitude. Synthetic speech today is generally good enough to avoid the worst cases. Less effort has been put into increasing coherence beyond these obvious cases, yet we all know that some sound sequences are much more acceptable than others that are just as intelligible. Auditory streaming strongly suggests that to produce robust synthetic speech, we must pay attention to the fine detail of the acoustics: to the variation that has typically been ignored as not essential to phoneme identification.

The popularity of concatenated natural speech segments over formant-based synthesis supports this argument. The phonetic quality of formant-based synthetic speech is not much better than it was a decade ago, and many applications continue to use concatenated natural speech. In formant synthesis, the most stringent measures of segmental intelligibility, such as sound identification in isolated syllables, reach a ceiling above which it is difficult to make significant improvements. Well done, concatenated speech has at least two advantages over formant synthesis: it contains all the short-term systematic variation (e.g. at segment boundaries) of natural speech, and at least some of the longer-term variation. Typically, formant synthesis mimics only some of these relationships, mainly those that most clearly underpin phoneme identification and, to some extent, speech rhythm and intonation. When more subtle properties like vowel-to-vowel coarticulation are included in formant synthesis, it sounds better and is significantly more intelligible, especially in difficult listening conditions [6,7].

3.4 Speech perception by humans

A wide range of work in speech perception converges to emphasize that systematic variation is central to the speech signal. The motor theory [8] has obvious relevance. One does not need to espouse such theories in their entirety to acknowledge the importance of their central tenet: that the listener's knowledge of the relationship between vocal tract behaviour and sound profoundly influences his or her understanding of the speech signal. Sounds that could come from a vocal tract are perceived as speech; sounds that the vocal tract cannot produce are less likely to be heard as speech. Sounds that cannot come from a vocal tract but can nevertheless be interpreted, like sine wave speech [9], seem to be understood because they mimic fundamental properties of speech, and hence of vocal tract movement: achieving the right timing, frequency and amplitude relationships is crucial. No one claims that sine wave speech sounds natural, nor that it is easy to understand: these requirements demand that fine acoustic detail is added. And this detail follows the systematic variation caused by the way the vocal tract works.

Theories based on the acoustic signal incorporate vocal tract dynamics at least implicitly to the extent that they refer to time-varying properties. The theory of acoustic invariance, for example, stresses effects of the changing shape of the vocal tract that are reflected in constancy of relationships in frequency or amplitude across acoustic boundaries [10]. Postulated invariant properties transcend systematic variation in the signal, yet include it because the variation is part of each measure of invariance. The variation can be responded to as information rather than noise if we assume that the perceptual process continuously assigns probabilities rather than binary values to features, as I suggest below.

Experiments from other theoretical approaches also support the importance of vocal tract dynamics. The large literature on the importance of consonant transitions to phoneme identification is a prime example, while others show that the more subtle systematic variation also contributes to perceptual decisions. Some of these are mentioned below.

4 TOWARDS A NONSEGMENTAL MODEL OF SPEECH PERCEPTION

This section considers what the above arguments, if accepted, could entail for a model of speech perception.

4.1 The task

In the phonetic literature, the term speech perception often seems to mean the identification of phonemes or syllables in simple contexts. I see this interpretation as narrow, and prefer to define the task of speech perception as to understand the meaning of what someone has said. That task is too large for study however; I see the immediate phonetic task as to identify words, meaningful or not. Psychologists call this lexical access.

4.2 Modality

A historian might be forgiven for concluding that one must decide on whether the modality of interpretation is motoric or acoustic/auditory e.g. [11]. I believe that the sharp division that has been drawn between these approaches is one of philosophy rather than of evidence—the differences are often smaller than has been suggested [11] and the theoretical approach can influence the experimental design and analytic method to create spurious differences [12]. Rather, consistent with the preceding argument, I assume that *all* relevant sensory information is usable. Modality is not crucial, but the input must seem to have come from a vocal tract.

4.3 What constitutes perceptual information?

I make three assumptions about what constitutes perceptual information: that all speech-relevant information is potentially salient; that sensory input is interpreted in relational terms; and that the signal varies in the amount of information carried per unit time.

The assumption that all speech-relevant information is potentially salient to the perceptual mechanism does not entail the claim that it is always all used. Whether it is used, and the extent to which it is used, depends on its quality and on what other information is available. Evidence supporting this view comes in many forms, including acoustic cue trading, the many demonstrations of the influence on sound or word identification of higher-order linguistic factors such as vocabulary size, predictability from context, and lexicality, and cross-modal influences on speech perception e.g. [13]. Of these, the last is perhaps most worth discussing.

In [13], /baba/ and /gaga/ were cross-matched such that listeners heard /baba/ synchronised with a video of a mouth saying /gaga/, or vice versa. Responses were asymmetrical: the visual stimulus has a profound influence when the heard stops are bilabial, but when they are velar, the visual influence (of /baba/) is smaller and less consistent. The explanation rests in what the listener knows about the relative quality of each sensory channel. Acoustically, velar stops are fairly distinctive, whereas bilabials are not [14,15] and can easily be misheard. Clear sight of a closure being made inside the mouth can apparently cause the weak and hence potentially unreliable acoustic properties of a bilabial stop to be disregarded in favour of the more reliable visual information: when /baba/ is heard but /gaga/ is seen, the visual input dominates. When information from both channels is clear and hence reliable, the perceptual system gives weight to both, and produces combination g-b responses.

Evidence for the second assumption, that acoustic and visual information is interpreted in relational terms, is also widespread. Auditory streaming attests to its importance. Timing, by its nature, involves relational properties, as do aspects of perceived phone identity such as stop voicing and schwa identity. That relational properties are fundamental suggests that normally, sounds or features can only be interpreted in context. While that is a relatively new idea in acoustic studies of speech perception, it is not new in linguistic theory: relational properties underpin the entire phonetic and phonological structure. When the salient sensory cues are also expressed in relative terms, we have a consistent contrastive structure from sensory input up to the lexicon.

The third assumption, that the signal varies in the amount of information conveyed per unit time, requires no justification, but it does have important consequences for our thinking about perception. Let us first consider regions of the signal that are rich in information. These are sometimes called islands of reliability. They feature (with different names) in a number of theoretical approaches, including invariance theory [10], quantal theory [16] and robust features [17]. While work on acoustic invariance has tended to emphasize dynamic properties, robust features are typically characterised in terms of properties that are constant for the duration of at least the major portion of a phone-sized acoustic segment. It is not clear that we need to choose between these approaches. While acoustic invariance seeks short-term properties that are minimally sufficient to provide evidence for a particular feature, each

robust feature must last long enough for the phone it underpins to be recognizable. The two approaches reflect different consequences of articulatory movement, and so both contribute to the signal that the perceptual system tracks.

If we accept this reasoning, then our perceptual model must effectively operate with at least two time windows, a short one for rapid events, and a longer one for more continuous properties. And, since different features are recognized at different times, they will not naturally fall into the neat bundles of standard phonology. These consequences are consistent with data showing that the temporal structure of both spectral change and steady states is critical for the correct identification of most sounds cf. rate of change of formant transitions (stop vs approximant), and the duration of frication noise, which, when short, can contribute to the percept of place of stop articulation [14,18], and when long is heard as fricative or affricate. Anecdotally, I need to hear quite a lot of the vowel in a CV syllable before it takes on the right quality. At shorter durations, I hear one or more other English vowels.

Evidence for the contribution of regions of the signal that are not rich in information is more sparse than that for islands of reliability, but that may be due partly to fashions of inquiry. Some regions of the signal indisputably demand more inference about the message than others e.g. some phones are inherently not robust [19]. Nevertheless, regions of low information can contribute valuable perceptual information. Under some conditions, natural variation in formant frequencies that is engendered by consonants can spread throughout adjacent vowels and even to nonadjacent ones. Experiments in progress in my laboratory show that listeners can use such weak acoustic cues to identify phonemes in natural and synthetic speech (cf. [6]). Gating experiments illustrate the use of both weak coarticulatory information and islands of reliability [20].

4.4 Is the phoneme necessary for speech perception?

I have argued that there is reason to suppose that the perceptual system closely tracks the detailed acoustic signal, along with other sensory information such as sight of the speaker's face, if available. I have also argued that all input provides potentially valuable information, that its quality is evaluated during the process of making perceptual decisions, and that relational (context-dependent) properties are fundamental. These arguments lead me to question whether a phonemic stage is necessary to lexical access. Why not map more detailed properties of the signal directly onto words? This proposal is not original (cf. [21,22]), but some reasons for making it are worth examining, in addition to those made by e.g. [18] that acoustic cues are not always straightforwardly combined into phonetic features, nor features into phonemes.

An obligatory phonemic stage must be intermediate between the acoustic signal and the lexicon. An intermediate classification seems only worthwhile if it reduces processing load: it must be reasonably error-free, and allow information to be thrown away. But acoustic information seems to be held until quite late in the identification sequence. For example, listeners can back-track to reinterpret acoustic information quite a long time after a misperception, reconstructing an entire phrase and seeming to 'hear' that the reconstruction is more satisfactory than the original interpretation (see [23]).

Another argument is that some phonemic sequences map uniquely onto words only after the acoustic offset of some candidates e.g. 'plum' vs 'plumber' in *I saw the plum on the tree* [24]. The listener seems able to keep both lexical options available [25], but it seems risky to keep only phoneme strings without detailed sensory information, and contrary to evidence of late integration of different sources of information e.g. [26].

A less commonly made argument comes from language acquisition. Children seem to learn to talk by imitating the sound pattern of what they hear, without a complete phonological (or syntactic) analysis [27,28]. If that is the case, then presumably they operate without a fully systematic phonemic inventory, and if children start by doing that, it is difficult to see that they should be obliged to change as they get older. I suggest that it is possible but not obligatory to interpret the signal in terms of phonemes before lexical access. Normally, the phonemic interpretation will come after lexical access, perhaps to the extent that the person is literate.

Modelling phonemes as only an optional route resolves the conundrum in which allophonic information is crucial to feature identity and word segmentation but must be ignored in order to assign phoneme status. In a model in which phonemes are not central, we preserve the perceptual cueing value of variation due to phonetic context and connected speech processes by relating the input directly to phonological structure. Thus we preserve the information about syllable-dependent variation in the spectral and temporal properties of phones that is crucial to lexical access. This can have interesting phonological implications. Take the patterns of clear vs dark vs vocalized /l/ found in several varieties of English. Thus *lull* is [lʌɬ] in standard Southern British English, but [lʌʊ] is rapidly gaining ground as a stylistic option for some speakers, and is the only option for others. In standard phonological theory, all these accents are said to have a phoneme /l/ which can fall either before or after a syllabic nucleus. But are these 'l's the same for speakers who have only the vocalized version syllable-finally? For such speakers, the vocalized version is subject to linking phenomena which cannot occur in the dark /l/ version: consider *legal fees*, [ligʊfiz], but *legal aid*, [ligəweɪd]. This suggests to me that, in these contexts, vocalized and word-initial /l/ are phonologically distinct. An important consequence of distinguishing syllable position of phones or features is that syllabic constituency is not only signalled, but preserved throughout the interpretation. Correct assignment of syllabic constituency seems basic to correct word segmentation, but this information is lost in a phonemic string unless phonotactic constraints are violated.

4.5 Outline of a nonsegmental model of speech perception

The model I suggest follows that proposed by [23]. Here, I develop some nonsegmental aspects of the model, for a bottom-up channel from the sensory signal to words. The role of higher-order knowledge and the model's interactive aspects are neglected here due to space constraints; they are discussed in [23].

One consequence of seeking a model that is closely tied to vocal tract dynamics is that it will use a rich acoustic structure with many redundancies. Another is that time must be explicitly represented, with both rapid events and more slowly changing information contributing to perceptual decisions. The model assumes that all speech-relevant information is used, weighted according to its apparent value. The sensory input is interpreted in terms of linguistically-relevant units, and lexical items are represented as complex structures involving those same units.

These lexical structures comprise syllables and their constituents, together with information that maps onto higher-order structures of prosodic and grammatical trees. Thus intonation and rhythm guide decisions and focus attention onto stressed syllables [29]. Lexical structures include some set of features as terminal elements. These features are unconventional: they take probability rather than binary values, and are distributed across time rather than bundled into units with discrete boundaries. Probabilities attached to features can change within as well as between syllabic constituents. Thus weak cues from small coarticulatory effects are represented. For example, the probability of a feature [high] is significantly greater than zero in the nucleus of the syllable before a high vowel, but it is higher still (normally 1) in the nucleus of the syllable containing that high vowel.

A model that assumes feature probabilities but neglects the weak cueing function of coarticulatory effects can assume that the lexical representation is in terms of resting levels, thresholds, and supra-threshold activation; any input value greater than the threshold activates the feature. But to accommodate coarticulatory cues, it seems necessary to limit the range of expected probabilities for each feature. When the effect of interest is weak (e.g. slight vowel raising due to coarticulation with a high vowel in the next syllable) both lower and upper limits of the range will be less than 1 for the relevant feature, here [high]. When the effect of interest is the primary

property of that part of the signal (e.g. a high vowel), then the upper limit will be 1, and the lower limit will be determined by contextual influences from other parts of the utterance. There must also be knowledge of *relative* probability of features across acoustic segments [23].

The input signal is represented as a set of prelexical features (or possibly loose clusters of features) whose values are also represented as probabilities. The signal is continuously monitored for information on each unit, giving rise to continuous modulation of probability levels of pre-lexical features, which in turn affects activation level of lexical items. Thus the model tracks time functions and hence vocal-tract dynamics (and their acoustic consequences), rather than only event sequences.

Lexical access involves taking the best match between input and stored probabilities for features. Unambiguous stimulus input is given great weight, and can be in any modality. But because the system is based on choosing the most probable answer, a signal can produce a clearcut response even if acoustic cues are relatively poor, as long as they are consistent for long enough and there is no strong contrary information.

The model preserves the relational, hierarchical structure of contrasts from input through to the highest levels of linguistic interpretation. No one unit is of prime importance, nor can it be functionally separated from the others in the structure of which it is part. (It can be analyzed independently.) In other words, acoustic information feeds several units simultaneously, and each unit uses several types of acoustic information.

Since the entire signal is potentially represented, the model system is 'holistic': it is not divided strictly into discrete segments, nor into segmental and prosodic strands. Such distinctions can be made, but they need not be, and possibly they are not normally made. In traditional terms, allophonic information and coarticulation are represented as central properties of the system, rather than as secondary or intermediate stages relative to phonemic information. Additionally, phoneme strings need not be identified before lexical access, although there is nothing to stop them being so identified, assuming they are represented in the lexical structures and available to the listener.

A rich and redundant structure allows flexibility in the units used and how the signal is segmented. This provides one source of individual differences in speech production and perception. In speech production, the route the child first learns for a particular articulatory manoeuvre stands a good chance of being perfected. It will be changed only if subsequent learned patterns conflict with it. Likewise for perception: some people pay more attention to one set of cues, others to another. Thus there is room in the model for experience of the individual child to underlie individual differences in adulthood.

Individual differences in experience may mean that people do not have maximally systematic representations of language in their brains. Informal evidence suggests that some people operate throughout life without a complete phonological and syntactic system as a linguist would recognize them. Take /aid ɔv laɪkt tɔ gʊʊ/, frequently expanded even by adults as *I would of* rather than *I would have*.

A relatively direct mapping from signal to lexicon seems to be consistent with the general approach of the more successful speech recognition by machine systems, whose impressive recent success has depended on the use of all acoustic information over long domains, for example an entire sentence, using fairly minimal linguistic information [30]. The general pattern-matching approach of statistical solutions to speech recognition is almost certainly germane to human speech perception. Possibly incomplete linguistic structures could be built up from statistical evidence of recurrent patterns, together with appropriate hardwiring in the brain. I have tried to show that this might be possible.

REFERENCES

[1] Fujimura. O. (1994) The syllable: Its internal structure and role in prosodic organisation. Ms.

[2] Massaro, D.W. (1987) Categorical partition: A fuzzy-logical model of categorization behaviour. In [3] 254-283.

[3] Harnad, S. (1987) *Categorical Perception*. Cambridge: CUP.

[4] Bregman, A.S. (1990) *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: MIT Press.

[5] Assmann, P.F. and Summerfield, Q. (1990) Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies. *J. Acoust. Soc. Am.* 88, 680-697.

[6] Hawkins, S. and Slater, A. (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.

[7] Local, J. (1993) Segmental intelligibility of a nonsegmental synthesis system. *York Res. Papers in Linguistics.*

[8] Liberman, A.M. & Mattingly I.G. (1985) The motor theory of speech perception revised. *Cognition* 21, 1-36.

[9] Remez, R.E., Rubin, P.E., Pisoni, D.B. & Carrell, T.D. (1981) Speech perception without traditional speech cues. *Science* 212, 947-950.

[10] Blumstein, S.E. & Stevens, K.N. (1979) Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *J. Acoust. Soc. Am.* 66, 1001-1017.

[11] Parker, E.M., Diehl, R.L. & Kluender, K.R. (1986) Trading relations in speech and nonspeech. *Perc. & Psychophysics* 39, 129-142.

[12] Andrew, J. (1989) The theoretical interpretation of context effects in categorical perception. Unpublished ms, Dept. of Linguistics, Univ. Cambridge.

[13] McGurk, H. & MacDonald, J. (1976) Hearing lips and seeing voices. *Nature* 264, 746-748.

[14] Kewley-Port, D. (1983) Time-varying features as correlates of place of articulation in stop consonants. *J. Acoust. Soc. Am.* 73, 322-335.

[15] Hawkins, S. & Stevens, K. N. (1987). Perceptual and acoustical analyses of velar stop consonants. *Proc. XI Int. Congr. Phon. Sci.* Academy of Sciences, Estonian SSR, 5, 342-345.

[16] Stevens, K.N. (1989) On the quantal theory of speech. *J. Phonetics* 17, 3-45.

[17] Zue, V.W. (1985) The use of speech knowledge in speech recognition. *Proceedings IEEE 73*, 1602-1615.

[18] Klatt, D.H. (1989) Review of selected models of speech perception. In W. Marslen-Wilson (ed.) *Lexical Representation and Process*. Cambridge: MIT Press 169-226.

[19] Miller, G.A. & Nicely, P.E. (1955) An analysis of perceptual confusions among some English consonants. *J. Acoust. Soc. Am.* 27, 338-352.

[20] Warren, P. and Marslen-Wilson (1987) Continuous uptake of acoustic cues in spoken word recognition. *Perc. & Psychophysics* 41, 262-275.

[21] Klatt, D.H. (1979) Speech perception: A model of acoustic-phonetic analysis and lexical access. *J. Phonetics* 7, 279-312.

[22] Stevens, K.N. (1988) Phonetic features and lexical access. In *The Second Symposium on Advanced Man-Machine Interface through Spoken Language*, 10-1 - 10-23.

[23] Hawkins, S. & Warren, P. (1994). Implications for lexical access of phonetic influences on the intelligibility of conversational speech. *J. Phonetics* 22, 493-511.

[24] Grosjean, F. & Gee, J.P. (1987) Prosodic structure and spoken word recognition. U.H. Frauenfelder and L.K. Tyler (eds.) *Spoken Word Recognition*. Cambridge, MA: MIT Press. 135-155.

[25] Cutler, A. (1986) Forbear is a homophone: lexical prosody does not constrain lexical access. *Language & Speech* 29, 201-220.

[26] Munhall, K.G., Gribble, P., Sacco, L. & Ward, M. (1994) Temporal constraints on the perception of the McGurk effect. *ATR Technical Report* TR-H-112.

[27] Eimas, P.D., Miller, J.L. & Jusczyk, P.W. (1987) On infant speech perception and the acquisition of language. In [3] 161-195.

[28] Ferguson, C.A., Menn, L. & Stoel-Gammon, C. (1992) *Phonological Development: Models, Research, Implications*. Timonium, MD: York.

[29] Cutler, A. (1990) Exploiting prosodic probabilities in speech segmentation. In G.T.M. Altmann (ed.) *Cognitive Models of Speech Processing*. Cambridge: MIT Press. 105-121.

[30] Young, S.J., Odell, J.J. & Woodland, P.C. (1994) Tree-based state tying for high accuracy acoustic modelling. *Proc. ARPA Human Lang. Tech. Conf.* Princeton: Morgan Kaufmann.

SEGMENTAL PHONOLOGY AND NON-SEGMENTAL PHONETICS

Patricia A. Keating
Phonetics Lab, Linguistics Department,
University of California, Los Angeles, USA

ABSTRACT

In this contribution I present the view that there is no fundamental problem in relating segmental, non-dynamic phonological representations to non-segmental, dynamic phonetic representations of speech, and that other kinds of theories of phonological representation are less suited to dealing with prosodically-conditioned variation.

INTRODUCTION

The question posed in this session is, "What benefits/problems flow from taking a dynamic/non-segmental approach to phonetics?". This question arises because most, though not all, phonologists have traditionally assumed that the segment is one basic level of phonological representation, and because most, though not all, phoneticians assume that there are no phonetic segments. Thus there is a mis-match between the two levels, which would seem to be undesirable. An apparent solution to this apparent problem has become increasingly popular: that the phonology should match the phonetics in being dynamic/non-segmental. In what follows I will present my own view on this issue: first, that there is no real problem crying out for a solution, and second, that a dynamic/non-segmental phonological representation creates new problems for phonetics, because it contains too much specific information.

In addition, I should note that I believe that there is good reason to assume that people have a tendency to construct a psychological representation in terms of segments. In short, I share the view that the widespread success of phonemic alphabets reflects (because it depends on) the ability of humans to readily construct segmental representations of words. (This is not to say that a person must have a complete and final segmental representation before learning to read, or that the orthography has no influence on the segmental representations.) I will not have time to defend that component of my position here, but it is certainly a motivating principle in what follows, because it means that some "benefit flows" from having at least a quasi-segmental phonological representation. Following Goldsmith [1], then, I see the language learner as being faced with the task of making structural sense of the speech signal by abstracting segments (and other higher-level units) from it.

"DYNAMIC" AND "NON-SEGMENTAL"

First, the terms under discussion, "dynamic" and "non-segmental", require some clarification, especially since they are not the same, and therefore might bring different benefits and problems into play. The question of whether representations are chunked into units is separate from the question of whether those units, or their components, are dynamic.

"Dynamic" seems the easier term of the two: if some component is dynamic, it is directly and inherently specified as time-varying. The phonological alternative is "static": no claim is made about how long a given property persists, how fast it comes on or turns off, etc. Static phonological representations (though non-segmental) are defended by e.g. Local [2].

"Segmental" is much the murkier term. (See Abercrombie [3] for an interesting historical review; see Pike [4] in particular for a full program of phonetic segmentation.) Phonetic segmentation, for example of spectrograms, usually refers to a strict division of the speech signal into discrete, non-overlapping, temporal slices which exhaustively parse the entire signal. Certainly this can be done up to a point: acoustic records of speech show a quasi-segmental character based on changes in the major manner features of the primary articulatory constriction. Thus conventions for acoustic measurement of "segment" durations are typically based on source characteristics

of the constriction (stop, fricative, approximant), rather than on voicing, nasalization, place of articulation (formant frequencies), or secondary constrictions. (This manner-based division of the signal is somewhat along the lines of McCarthy [5], where the segmental root node consists of the features Consonantal and Sonorant.) One non-segmental aspect of speech, then, is that other features of a segment do not have to line up with this basic manner division either grossly (e.g. nasalization may begin well before a nasal consonant) or in detail (e.g. voicing may continue for one or two pitch periods after the closure for a voiceless stop).

Another non-segmental aspect of speech is that when these manner features do not change, most notably in a sequence of resonants, no obvious segmentation emerges and our acoustic criteria are quite arbitrary. Put another way, acoustic signals may suggest some segmentation and perhaps support further segmentation, but not always corresponding to a phonological segmentation.

Hertz [6] takes an intermediate position on acoustic segmentation: phones are quasi-steady-state portions of the signal, while transitions are specific time intervals that come between phones. F2 is used as the primary basis for segmentation. Hertz and colleagues show that interesting phonetic generalizations can be made on the basis of this segmentation, e.g. that phones and transitions pattern differently in terms of durational changes.

Phonological segmentation is by no means the same thing as dividing up a spectrogram. The job of phonological segments is at least twofold: to indicate phonological precedence (which features come roughly at the same time vs. which are clearly in sequence), and to give a gross indication of notional time (a segment's worth of time). The same jobs are done by higher-level units too, of course; the segment is simply one level of such organization.

In fact, most phonological representations are what we might call "semi-segmental". They are basically segmental in that there are segment slots of some kind (root nodes, Xs, whatever) which do these jobs of segments, but they are non-segmental to the extent that features are autosegmentalized, that is, features can belong to no segment slots (as in floating morphological features), or to more than one (as in geminates), and more than one value of a feature can belong to a single segment (as in affricates).

Finally, note that for the purposes of thinking about the relation between phonological and phonetic representations, it doesn't matter whether segmental phonological representations are underlying (as most phonologists assume) or derived (e.g. Archangeli and Pulleyblank [7]).

SEGMENTAL AND NON-DYNAMIC PHONOLOGY: IT'S NOT A PROBLEM

The traditional class of models of the relation between phonology and phonetics is known as "target and interpolation" models. That is, these are models that provide targets and interpolations between targets. Individual phonological feature values associated with segments (or, in some speech synthesizers, unanalyzed whole segments) specify "targets" in articulatory and/or auditory-acoustic domains. The targets are aimed at by the speech producing mechanism, which moves, or "interpolates", from target to target. Examples of work in this framework include [6], [8], [9], [10], [11], [12], [13], [14], and [15].

In this approach, then, there are three steps in getting from a discrete phonological representation to a continuous phonetic representation.

The first step is a general one, and the other two are done for each utterance.

Step 1: relate each feature to one or more parameters (articulatory or auditory-acoustic, as relevant). To some extent this correspondence will be the same across languages: some one parameter is most basic for a given feature; but to some extent this correspondence will vary across languages, because other parameters may also be used, or not. These expressions of features can be quite complex, but that is the nature of speech, not of the theory itself (contra Zsiga [16]).

Every theory must grapple somewhere with the dual facts that articulatory-acoustic correspondences are complex and that different articulations can be used together to produce or enhance a given acoustic end. For example, the feature Strident needs to control parameters of tongue shape, jaw (tooth) position, glottal opening size, and velic opening size. Task Dynamics theory (e.g. [17]) does this in terms of Coordinative Structures (where the example of Strident is more complicated than ones usually discussed in that framework), and Enhancement theory (e.g. [18]) does this in terms of redundant feature specifications (for which again Strident is a complex example).

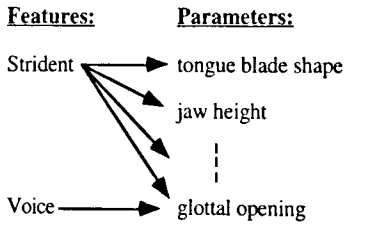


Figure 1. Features affecting multiple phonetic parameters.

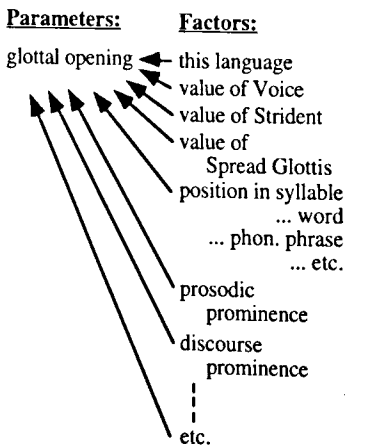


Figure 2. Multiple factors affecting the setting of a phonetic parameter.

Step 2: interpret any given (discrete) value of that feature as a (continuous) value in two dimensions: along the relevant parameters, for some interval of (notional) time. This continuous value will depend on many factors besides the phonological feature value, including values of other features in the same segment, and prosodic variables. For some parameters, more than one feature will determine the ultimate value. For example, Strident, which wants an open glottis for high airflow, competes with Voice, which wants approximated vocal cords for vibration, for control of glottal opening in a voiced strident, which makes voicing breathy, and/or difficult to sustain. Assignment of target values is called target evaluation by Pierrehumbert [8].

Step 3: connect successive values according to some mathematical function; this function may differ for different target values, parameters, languages, but is most commonly treated as linear.

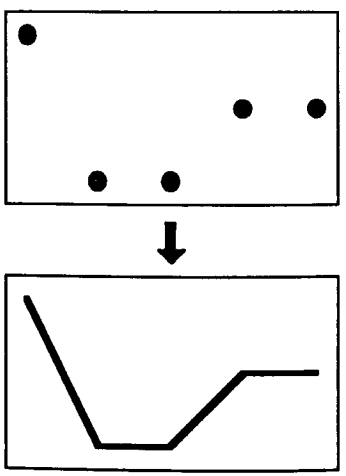


Figure 3. Interpolation between successive targets on some paramter

The evaluation for each feature is temporally independent in the sense that, for example, the different articulations for a single segment can begin and end at different times. That is, target and interpolation models require a theory of target alignment in the same way that Articulatory Phonology requires a theory of gestural phasing. The difference is that with a segmental phonology, these alignments are not considered to be lexically specified.

A property of both target and interpolation models and Articulatory Phonology is what I call phonetic underspecification. Phonetic underspecification means that not every segment has to have a specification, or target, for every feature. This has a strong effect on speech when interpolation functions do not care whether adjacent featural target specifications are from adjacent segments or not; targets are connected up through an empty time interval between them. This means that the effects of target specifications will extend further in time than the time interval occupied by the targets themselves. This is a way of getting dynamic effects without having the targets themselves be dynamic.

The diagnostic for phonetic underspecification, then, is variability across contexts. If there is no specification, then what you see will depend entirely on the surrounding specifications, which will trigger interpolation through the unspecified span in a temporally-gradient fashion.

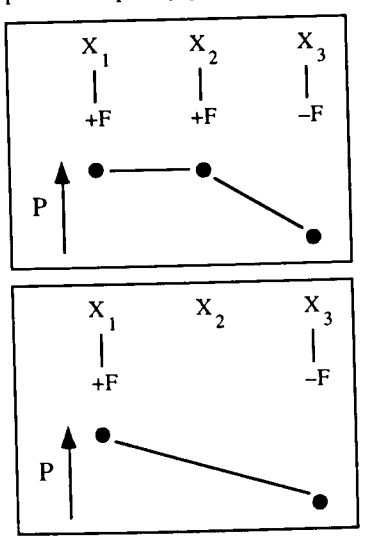


Figure 4. Interpolation between specified (top) and underspecified (bottom) values of some parameter.

On this account, then, much allophonic variation, especially variation that is coarticulatory or assimilatory in character, is generated by the quantitative operations (target evaluation and interpolation) just described. It can be seen that many allophones that have been previously described featurally (that is, by the change or acquisition of a feature value in a segment) are not described featurally here. As just one example, many vowel allophones that might be noted in a narrow phonetic transcription can be derived by interpolation, not by feature changes (e.g. Choi [15] on Marshallese). This does not mean that phonological feature spreading or changing cannot occur, but it certainly means that it does not occur as often as has been posited in the past. *The position that a segmental phonology means that every case of allophonic variation must involve segmental allophones is a straw man.*

Next, note that targets come from -- are projected by -- phonological feature values. Thus, when there is a phonological contrast, and therefore feature specification, there must be one or more phonetic targets (depending on how many parameters implement a feature). The targets will be the main influence on the parameter contour at that time. But if at some point in time there is no contrast that uses a given parameter, there will be no target at that time on that parameter, and the influence of context will obviously be strong. That is, contrast can restrict contextual variability while lack of contrast can give rise to contextual variability.

I developed this idea as the "window" model of surface phonetics. In this model, targets are not single values. Instead, they are ranges of permitted values. For articulation, you can think of these as constraints that say how much it matters how precise an articulation is. Some targets are very narrow ranges or windows; they permit little variation. Other targets are wide ranges or windows; they permit correspondingly more variation. In effect, windows turn phonetic underspecification from an all-or-none proposition to a gradient proposition.

With target ranges rather than points, interpolation becomes a more complicated function. For articulation, the general idea is to go as slow as possible while still making it into the required range. Guenther [19] has implemented a neural

net model of articulation that incorporates a windows-like idea. Guenther's ideas differ from mine in certain respects but his model shows that interpolation between articulatory target ranges in accord with motor control principles is possible.

Interestingly, Guenther is developing his target ranges as an implementation of Lindblom's H&H dimension [20]. A small window is a kind of hyperarticulation because it requires more careful speech to reach the small target and it limits coarticulation. So the target sizes encompass all styles of speech, but slower speech, and more careful speech, would be modeled as a shrinking of the targets, while faster/less careful speech would use the fully expanded targets. He also proposes to follow up a result of deJong et al. [21] and deJong [22], that phrasal prominence results in a decrease in contextual variation and thus involves hyperarticulation. This can be modeled straightforwardly as a decrease in target range of the head of a prosodic domain (though his model as it stands needs some modification to generate more extreme articulations under hyperarticulation). Not only are vowel articulations hyperarticulated under stress, but onset consonants show hyperarticulation effects of stress on their oral [22] and glottal [23] gestures.

I would take this approach even farther. It seems likely that hyperarticulation characterizes not only prosodic heads, but also at least some prosodic edges. Consider how edges of words are treated in English (and plausibly other languages). Wordfinally, lengthening is the demarcative property (e.g. [24], [25]), specifically lengthening of the closing (VC) gesture of the final rime (e.g. [26]) rather than the final consonant constriction interval, which is in fact shorter [27]. Word-initially, on the other hand, the initial consonant's constriction interval is lengthened [27] and both oral and velic gestures are more constricted ([28], [29]--see review by Browman and Goldstein [30]) and glottal opening is greater (again, reviewed by [30]). Most strikingly, and more generally, the size of the glottal opening for /h/ and for aspiration gets bigger the larger the prosodic domain in which the consonant is initial ([23], [31]). Thus it is not only heads of prosodic domains that are hyperarticulated; initial edges are also hyperarticulated, even above the word level.

This means that prosodic structure, not only at the syllable and word levels, but also at different phrasal levels, probably plays an enormous role in determining what we think of as purely "segmental" characteristics, like degree of stricture, as well as what we think of as "suprasegmental" characteristics, like duration.

THE ALTERNATIVE HAS A PROBLEM

We have seen how a distinction can made between phonological feature values, which characterize physical properties only very abstractly, and phonetic targets, which have specific spatial and temporal quantitative values as a function of many factors, including but not limited to the phonological feature values. Suppose instead we want phonological representations to include much of this detail: some indication of how fast and how long a movement will be, a more explicit indication of its exact spatial goal or target, more information about the relative alignments of different movements -- say, a theory like Articulatory Phonology. How will all the prosodic variation discussed above be dealt with?

Browman and Goldstein [30] broach this issue. However, they do so by focusing on different kinds of variation at different levels. Variation in gestural parameter specifications and in phasing are both considered at the word level, where lexical stress, position in syllable, and position in words are all relevant variables. As long as such variation occurs within the word, it can be incorporated into the lexical representation itself, where it is redundant information, but useful in specifying how the word is to be actually pronounced. Phrasally, however, Browman and Goldstein consider differences in phasing only as they occur between words. These different phasings leave the lexical representations intact and simply slide them around relative to one another. Even American English flapping of final alveolars before vowel-initial words is said to involve no change in the word-final alveolar gesture itself, presumably because if it did, that would be harder to account for.

Yet within-word variation in gesture parameters and possibly in phasing does occur as a function of postlexical structure, and in fact may be completely pervasive. In that case, we cannot say that lexical specification tells us how to pronounce a word, only how to pronounce it in some particular context. Which prosodic position should be taken as the basis for the lexical representations? Should some position be taken as canonical, and other variants derived from it by some kind of readjustment? This would go against the spirit of the whole endeavor, because the lexical information would be misleading just because it is precise. Or should a list of all possible alternatives be precompiled, along with indices so you select the right one for any given occasion? Not only does this again go against the spirit of the theory, but it requires that there be some finite number of possibilities. Or should gestures in lexical representations indicate only ranges of spatial and temporal variation, with more precise values to be determined postlexically? Or should lexical representations be segmental and non-dynamic, as they are in Zsiga's [16] version of Articulatory Phonology? In these last two cases, the segmental-and-nondynamic and non-segmental-and-dynamic theories will turn out to be much more alike than they now seem.

REFERENCES

[1] Goldsmith, J. (1976), *Autosegmental phonology*, PhD dissertation, MIT.

[2] Local, J. (1992), "Modeling assimilation in nonsegmental, rule-free synthesis", *Papers in Laboratory Phonology II* (eds. Docherty & Ladd):190-223.

[3] Abercrombie, D. (1991), *Fifty Years in Phonetics*, Edinburgh University Press.

[4] Pike, K. L. (1949), *Phonetics*, University of Michigan Press.

[5] McCarthy, J. (1988), "Feature geometry and dependency: a review", *Phonetica* 45:84-108.

[6] Hertz, S. R. (1991), "Streams, phones, and transitions: toward a new phonological and phonetic model of formant timing", J. Phon., 19(1):91-109.

[7] Archangeli, D. & D. Pulleyblank (1994), *Grounded Phonology*, MIT Press.

[8] Pierrehumbert, J. B. (1980), *The phonology and phonetics of English intonation*, Ph.D. dissertation, MIT.

[9] Pierrehumbert, J. & M. Beckman (1988), *Japanese tone structure*, Linguistic Inquiry Monograph, MIT Press.

[10] Clements, G. N. & S. R. Hertz (1991), "Nonlinear phonology and acoustic interpretation", *Proceedings of the XIIth ICPhS* , vol 1:364-373.

[11] Clements, G. N. & S. R. Hertz (1995), "An integrated representational system for phonology and acoustic phonetics, with a case study of English vocalic nuclei", MS.

[12] Huffman, M. (1990), *Implementation of Nasal: timing and articulatory landmarks, UCLA Working Papers in Phonetics* 75.

[13] Cohn, A. (1990), *Phonetic and Phonological Rules of Nasalization, UCLA Working Papers in Phonetics* 76.

[14] Cohn, A. (1993), "Nasalization in English: phonology or phonetics", *Phonology* 10:43-81.

[15] Choi, J. D. (1992), *Phonetic Underspecification and Target Interpolation: an acoustic study of Marshallese vowel allophony, UCLA Working Papers in Phonetics* 82.

[16] Zsiga, E. (1993), *Features, gestures, and the temporal aspects of phonological organization*, Ph.D. dissertation, Yale U.

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

[18] Stevens, K. & S. J. Keyser (1989), "Primary feature and their enhancement in consonants", *Language* 65:81-106.

[19] Guenther, F. H. (1994), Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psych. Rev.*, in press.

[20] Lindblom, B. (1990), Explaining phonetic variation: A sketch of the H & H theory, in W. J. Hardcastle and Alain Marchal (eds.) *Speech Production and*

Speech Modelling, Kluwer Academic Publishers, pp. 403-439.

[21] deJong, K., M. Beckman, and J. Edwards (1993), The interplay between prosodic structure and coarticulation, *Language and Speech* 36:197-212.

[22] deJong, K. (1995), "The supraglottal articulation of prominence in English: linguistic stress as localized hyperarticulation", *JASA* 97(1):491-504.

[23] Pierrehumbert, J. & D. Talkin (1992), "Lenition of /h/ and glottal stop", *Papers in Laboratory Phonology II* (eds. Docherty & Ladd):90-116

[24] Oller, D. K. (1973), "The effect of position in utterance on speech segment duration in English", *JASA* 54:1235-1247.

[25] Wightman, C. W., S. Shattuck-Hufnagel, M. Ostendorf, & P. J. Price (1992), "Segmental durations in the viscinity of prosodic phrase boundaries", *JASA* 91:1707-1717.

[26] Beckman, M., J. Edwards, & J. Fletcher (1992), "Prosodic structure and tempo in a sonority model of articulatory dynamics", *Papers in Laboratory Phonology II*, (eds. Docherty & Ladd):69-86.

[27] Byrd, D. M. (1994), *Articulatorry timing in English consonant sequences, UCLA Working Papers in Phonetics* 86.

[28] Krakow, R. A. (1989), *The articulatory organization of syllables: a kinematic analysis of labial and velar guestures*. PhD dissertation, Yale U.

[29] Vaissiere, J. (1988), "Prediction of velum movement from phonological specifications", *Phonetica* 49:48-60.

[30] Browman, C. & L. Goldstein (1992), "Articulatory Phonology: an overview", *Phonetica* 49:155-180.

[31] Jun, S. (1993), *The phonetics and phonology of Korean Prosody*, PhD dissertation, Ohio State U.

THE EARS HAVE IT: THE AUDITORY BASIS OF SPEECH PERCEPTION

Steven Greenberg
Department of Linguistics
International Computer Science Institute
University of California, Berkeley, CA 94720 USA

ABSTRACT

Two neural systems - the auditory pathway and the thalamo-cortical association areas - set constraints on the acoustic properties of all vocal communication systems, including human language. This paper discusses the manner in which the two systems, operating in concert, may shape the spectro-temporal properties of human speech.

INTRODUCTION

The acoustic nature of speech is typically attributed primarily to constraints imposed by the human vocal apparatus. The tongue, lips and jaw can move only so fast and so far per unit time, the size and shape of the vocal tract set limits on the range of realizable spectral configurations, and so on. Although such articulatory properties doubtless impose significant constraints, it is unlikely that such factors, in and of themselves, entirely account for the full constellation of spectro-temporal properties of speech. For example, there are sounds which the vocal apparatus can produce, such as coughing and spitting, which do not occur in any language's sound inventory. And although speech can readily be whispered, it is only occasionally done. The vocal tract is capable of chaining long sequences of vowels or consonants but no language relies exclusively on either basic segment form, nor does speech contain long sequences of acoustically similar segments.

In this paper it is proposed that auditory system imposes its own set of constraints on the acoustic nature of speech, and that these factors are crucial for understanding how information is packaged in the speech waveform. This packaging is designed largely to insure robust and reliable coding of phonetic information by auditory mechanisms operating under a wide range of potentially adverse acoustic conditions, as well as to integrate these phonetic features with information derived from neural centers responsible for visual and motor coordination.

SPECTRO-TEMPORAL PROPERTIES OF SPEECH

The discussion focuses on the following parameters of auditory function:

(1) the range of frequency sensitivity
(2) the frequency resolving capabilities of peripheral auditory neurons
(3) the limits of neural periodicity coding
(4) the time course of rapid adaptation
(5) the temporal limits of neural coincidence detection
(6) the modulation transfer characteristics of brainstem and cortical auditory neurons.

These parameters account for a number of important acoustic properties of speech, including:

(1) an absence of spectral energy above 10 kHz
(2) a concentration of spectral information below 2.5 kHz
(3) preference for encoding perceptually relevant information in the spectral peaks
(4) sound pressure levels for segments ranging between 40 and 75 dB
(5) rapidly changing spectra
(6) a prevalence of phonetic segments with abrupt onsets and transients
(7) a preference for quasi-periodic waveforms whose fundamental frequencies range between 75 and 330 Hz
(8) temporal intervals for integrative units (syllables and segments) ranging between 50 and 250 ms

These acoustic properties are essential for the robust encoding of speech under uncertain (and often noisy) acoustic conditions and in circumstances where multiple vocal interactions may occur (e. g., the proverbial cocktail party).

Each of these properties is examined in turn.

Spectral Energy Distribution

a. Energy entirely below 10 kHz.

No speech segment contains significant energy greater than 10 kHz, and most speech segments contain little energy above 4 kHz [1,2].

b. Concentration of energy below 2.5 kHz.

The long term power spectrum of the speech signal is concentrated below 2.5 kHz, as a consequence of the temporal domination of vowels, glides and liquids in the speech stream [1,2].

Moderate to High Sound Pressure Levels

Although we are certainly capable of talking at very high or very low amplitudes, we rarely do so. Rather, speech is spoken at a level that is approximately 60-75 dB for vocalic segments and about 20-30 dB lower for stops and fricatives [3].

Rapid Changes in Spectral Energy over Time

a. Significant changes in spectral energy maxima over 100-ms intervals.

The spectral energy in speech moves rapidly through time. There is virtually no segment in the speech signal where the formant pattern remains relatively stationary for more than 50 ms [2].

b. Prevalence of abrupt onsets (e. g. stop consonants, clicks, affricates).

There is a tendency for words to begin with segments having abrupt onsets, such as stops and affricates and it is uncommon for words to begin with gradual onset segments such as vowels [4].

Temporal Modulation Characteristics

a. Micro-modulation patterns with periodicity between 3 and 12 ms

Most of the speech signal is produced while the vocal folds are vibrating. The acoustic result is that the speech waveform is modulated at a quasi-periodic rate ranging between 3 (330 Hz) and 12 ms (85 Hz). The lower limit is characteristic of the high range of female voices while the upper limit is typical of the low end of the male range.

b. Macro-modulation patterns on the time scale of 50 to 250 ms

Riding on top of this micro-modulation is a longer periodicity associated with segments and syllables. The durational properties of syllables is a property of the permissible sound sequences in the language [5].

It is likely that all spoken languages are characterized by these properties, despite the differences in their phonetic inventories. It is these "universal" macro-properties of the speech signal that form the focus of the discussion below.

AUDITORY MECHANISMS

The auditory bases for the spectro-temporal properties described above are varied and complex, reflecting general constraints imposed by the mammalian acoustic transduction system. However, these general mammalian properties have special consequences for the nature of speech as a consequence of the unique fashion in which sensory-motor and cognitive information is integrated in the human brain.

Spectral Energy Distribution

Human listeners are sensitive to frequencies between 0.05 and 18 kHz [6]. However, the truly sensitive portion of this range lies between 0.25 and 8 kHz [6], setting approximate limits to the bandwidth of the acoustic communication channel.

The precise shape of the human audibility curve is conditioned by several factors. The lower branch of the audibility curve reflects the impedance characteristics of the middle ear [7]. The coupling of the ear drum to the oval

window of the cochlea via the ossicular chain is much more resistive to low frequencies than to high.

The upper branch of the audibility range is determined by the mechanics of the inner ear, which in turn is accounted for by macro-evolutionary factors pertaining to the ability to localize sound. The upper audibility limit pertains to the frequency range over which interaural intensity cues are available for precise localization of sound in the azithmuthal and vertical planes. Because of the relatively large diameter of the human head (25 cm), it is possible to extract reliable localization information based on differential intensity cues for frequencies as low as 4-6 kHz [8]. This is an important limit, because the upper boundary of audibility for mammals is conditioned largely by the availability of these cues. For a small headed mammal, such as a mouse, comparable cues are available only in the human ultrasonic range, well above 20 kHz. Small headed mammals tend to be sensitive to far higher frequencies than their larger-headed counterparts [9]. Humans and other large-headed mammals need not be sensitive to the high-frequency portion of the spectrum since they can exploit both interaural time and intensity cues at the lower frequencies. in view of the limited number of neural elements available for frequency coding, an increase in bandwidth sensitivity necessarily reduces the proportion of tonotopic real estate focused on the lower portion of the spectrum. The comparative evidence suggests that there is a preference for focusing as much of the computational power of the auditory pathway on the lower end of the spectrum as possible, and that sensitivity to the high-end of the frequency range is a drawback except for the rather necessary function of source localization and characterization. There is a further implication, as well, that there is something special about the low-frequency portion of the spectrum, as discussed below.

Thus, it is no mystery why speech sounds contain little energy above 10 kHz. The human ear is relatively insensitive to frequencies above this limit because of the low-frequency orientation of the inner ear. But what precisely accounts for this low-end spectral bias, and could this account for the concentration of speech energy in the speech signal below 2.5 kHz?

One of the common properties of all vertebrate hearing systems is the ability of auditory neurons to temporally encode low-frequency information. This encoding is performed through a process referred to as "phase-locking," in which the time of occurrence of a neural discharge is correlated with the pressure waveform driving the cell. The limits of this temporal coding in the auditory periphery are a result of the modulation of neurotransmitter released by inner hair cells and the uptake of these chemicals by auditory-nerve fibers [10]. Auditory-nerve fibers phase-lock most effectively to frequencies up to 800 kHz, progressively diminishing in their temporal encoding potential with increasing frequency. The upper limit of robust phase-locking is ca. 2.5 kHz [11], although a residue of phase-locking persists up to 4-5 kHz [11].

What is the significance of phase-locking for encoding speech-relevant information? It provides a means of encoding spectral information in a robust, fault-tolerant fashion that is important for signal transmission in uncertain and potentially adverse acoustic conditions. In order to understand how this occurs, it is helpful to first consider another means of encoding frequency information in the auditory system.

The discharge rate of a peripheral auditory neuron is roughly proportional to the energy driving the cell over a limited range of sound pressure levels. Because of the filtering action of the cochlea, it would be possible to encode the spectrum of a complex signal entirely in terms of the average discharge rate of spectrally tuned neurons across a tonotopically organized neuronal array if the dynamic range of these cells was sufficiently large. However, the range over which most auditory neurons increase their firing rate is only about 20-30 dB [12], far too small to effectively encode spectrally dynamic features. As a consequence, this "rate-place" representation of the acoustic spectrum becomes progressively blurred at higher sound pressure levels, despite robust perceptual decoding [13].

A significant problem with the rate-place code is its vulnerability to acoustic interference. Because the magnitude parameter is based simply on average rate it is impossible to distinguish neural activity evoked by a target signal of interest and extraneous background noise. There is no effective means of segregating the two sound sources on the basis of neural "labeling." A rate-place representation is particularly vulnerable to acoustic interference since its only measure of spectral identity is the location of activity in a topographically organized plane. Any extraneous source which intrudes into the place of the target signal could potentially disrupt its representation.

In contrast, phase-locked responses do provide a considerable degree of noise-robustness since the neural response is effectively labeled by the driving stimulus. A 500-Hz signal evokes a temporal distribution of nerve impulses rather distinct from one centered at 900 Hz, or even 550 Hz. This temporal information allows the auditory system to successfully segregate signals derived from disparate sources. In this fashion the temporal code provides a measure of protection against background sounds.

Phase-locking also provides a means to reduce the background noise as a consequence of its limited dynamic range. Frequencies more than 10-15 dB below the spectral peak driving the cell are rendered effectively invisible, since they do not affect the temporal discharge pattern [14]. This property effectively acts as both an automatic gain control [15] that suppresses background noise and enhances local peaks in the spectrum. This peak enhancement is combined with a broadening of the cochlear filters at moderate to high sound pressure levels to spread the labeled low-frequency information over a wide tonotopic range of neurons. As a consequence, there are many neural channels carrying similar temporal information, and this redundancy of central nervous system input provides for a large measure of reliability in encoding such phase-locked information. Noise tends to be concentrated in particular frequency bands, and therefore its potentially disruptive effect minimized by virtue of the important information distributed over a large number of channels.

Because robust phase-locking only occurs for frequencies below 2.5 kHz, there is an advantage in using the lower portion of the spectrum for encoding information that needs to be transmitted over noisy acoustic channels. Thus, there is a real incentive from an information reliability perspective to pack as much information in the lower spectral bands as possible. Disruption of the representation can be detected and patched up because it is possible to associate similarly labeled activity. Furthermore, the limited dynamic range of phase-locking makes it more difficult for noise to disrupt the encoding of low-frequencies. The signal to noise ratio must be exceedingly low before auditory neurons lose the ability to phase-lock to the foreground signal.

Enhancement of spectral peaks in the neural activity pattern has the effect of center clipping in the frequency domain providing considerable incentive to concentrate communicationally relevant information in the low-frequency spectral peaks, particularly at high sound pressure levels.

Moderate to High Sound Pressure Levels

The energy in the speech signal is distributed as follows.

The voiced speech sounds, particularly the vowels and glides typically possess a considerable amount of energy, in the range of 60-75 dB SPL at the receiver's ear [4]. And most of these voiced segments concentrate their energy below 2.5 kHz [5]. The neural signature cast by these segments spread their spectral shadow over much of the auditory nerve, recruiting high-frequency

neurons to their temporal discharge, thereby rendering the information encoded relatively impervious to acoustic interference [16].

Segments with energy concentrations in the mid- and high-frequency (3-8 kHz) portions of the spectrum are typically of much lower amplitude (30-50 dB). This is probably a consequence of two factors. This is the most sensitive portion of the human ear's audible range. However, the gain in sensitivity is modest (10 dB) relative to frequencies between 0.5 and 2 kHz. A second factor is perhaps of more significance. This is the portion of the spectrum above the limits of neural phase-locking. Because of the limited dynamic range of auditory neurons the spectral contours associated with these mid-frequencies are most clearly delineated in the rate-place representation at low sound pressure levels. For this reason, these segments are more intelligible at low-to moderate intensities than at higher SPLs.

One may therefore surmise that the amplitude of individual classes of speech segments depends at least partially on the robustness of the resulting auditory representation, and not just on purely articulatory considerations. Low sound pressure levels are required for optimum encoding of high frequency spectra as they are dependent on rate-place auditory code which has a restricted dynamic range. Low-frequency spectra, on the other hand, are most robustly encoded at moderate-to-high sound pressure levels as a result of the spread of phase-locked excitation at these intensities.

Rapid changes in the Spectral Distribution of Energy over Time

a. Significant changes in spectral energy maxima over 100 ms intervals

One of the most salient acoustic properties of speech is its constantly changing character. Stop consonants come on abruptly. The formant patterns of liquids, glides and even vowels are constantly shifting. In continuous speech, spoken at a normal rate, it is difficult to locate steady-state segments. This dynamic property of the speech signal is commonly interpreted as a reflection of a vocal tract pattern constantly in motion. And yet it is possible to produce quasi-steady-state speech-like segments in a continuous fashion, such as done in many forms of vocal music. But we don't typically communicate in chant.

So what factors account for the pace of spectral movement in the speech signal?

One consideration concerns the impact steady-state spectra have on the activity level of auditory neurons. A salient property of auditory neurons is adaptation. Turn on a signal and an auditory nerve fiber will fire at a very high rate for 5 to 15 ms [17], diminishing its activity level steadily over the next 100 ms or so. At higher levels of the auditory pathway many cells will fire only at signal onset. But this adaptation is highly frequency selective [18]. Change the spectrum of the signal and formerly quiescent neurons will turn on. In this fashion dynamic spectral properties could be used to maintain a high neural activity level. This preference for spectral dynamics is enhanced at the higher stations of the auditory pathway, where many cells respond only to spectrally changing stimuli, not to steady state signals.

A second factor pertains to the modulation transfer function of auditory cortical neurons, as discussed below. And a third factor pertains to the rate of information transmission through the auditory pathway into the central cortical areas, also discussed below.

b. Prevalence of abrupt onsets (e. g. stop consonants, clicks, affricates)

Abrupt onsets act in a manner similar to spectrally dynamic signals in that they tend to evoke a high rate of neural activity. This is a consequence of the fact that many cells in the central auditory pathway receive inputs from a wide tonotopically organized array of input neurons [19]. These cells act like "coincidence" detectors, responding to stimulation only when a sufficient proportion of their inputs fire within a very small interval of time, typically 250 μs or less. Among the more effective stimuli for simultaneous activation of a large neural array are transients associated with stop and affricate consonants, and as such are relatively more reliably encoded under adverse conditions.

Temporal Modulation Characteristics

a. Micro-modulation patterns with periodicity between 3 and 12 ms.

A significant proportion of speech, perhaps as much as 80 percent [3], is produced while the vocal folds are vibrating. And yet this predominance of voicing is not necessary for sustained vocal production, as evidenced by whispered speech.

The fundamental frequency of adult human speech ranges from 75 Hz for a low, adult male to 330 Hz for a high female voice [5]. Although this range reflects to a certain degree the length and mass of the vocal folds, it is possible for the human voice to go well above this range, as attested by operatic performance. What is there about the normal f_0 range that makes it special with respect to the auditory coding of speech?

If we look at the ability of auditory neurons to encode the waveform periodicity of spectrally complex signals such as speech, phase-locking to this temporal parameter of the speech signal is most robust among central auditory neurons in the range 75-300 Hz [20] and is the region of the most acute modulation discrimination [21].

The significance of encoding waveform modulation becomes apparent when we consider how the auditory system would track a sound source through time without some equivalent form of cohesive force to bind disparate spectral elements together into a single sound source. Because speech and other communication signals are typically broadband, the system needs to know that the activity in the low-frequency channels is related to that evoked in the higher channels. Common periodicity provides a cohesive cue that enables the system to attribute disparate neural activity to the same source.

Periodicity tracking in the range of the human voice is a particularly effective means of robust encoding of information [22]. At the level of the auditory nerve, fibers are capable of firing at sustained rates up to 250 spikes per second [23]. And at the level of the cochlear nucleus some cells can fire at rates up to 1000 spikes per second [24]. This phase-locked response to the modulation cycle enables each glottal cycle of the speech waveform to be marked with a burst of excitation that enables the system to track the signal across frequency and time. [25].

b. Macro-modulation patterns on the time scale of 50 to 250 ms

In addition to the modulation of the speech waveform imposed by vibration of the glottis, is a slower amplitude modulation that is correlated with the passage of individual segments and syllables. A normal speaking rate (at least for English-speaking Americans) is approximately 4 syllables per second [4]. And, in English, there are approximately 2.5 segments per syllable [3]. Thus, the average length of a phonetic segment is ca. 100 ms and that of a syllable, 250 ms.

At the higher stations of the auditory pathway, principally the auditory cortex, neurons generally respond at rates between 5 - 20 Hz (50 - 100 ms) [26]. Each cell in this region acts as an integrative center, its response reflecting the activity of hundreds, if not thousands of cells at more peripheral stations. It appears likely that the syllable and segment rate of spoken discourse is at least partially conditioned by the firing rates of these cortical neurons. These cells can phase-lock to the slow modulations of energy within their response areas and thereby provide an accurate representation of both syllabic and gross spectral contours. The gross amplitude modulation cues can be used to temporally bind neural activity driven by different portions of the spectrum, but having common onsets and offsets.

CORTICO-THALAMIC MECHANISMS

The brain utilizes specific strategies to integrate auditory information into linguistically meaningful units. It is rather unlikely that speech is processed phoneme by phoneme like so many "beads on a string." Rather the acoustic components of speech appear to segregate into syllable or mora-length units, which in turn are integrated into words and higher level semantic units.

What are the neural bases for this syllable timing properties of speech?

Many properties of auditory function appear to be governed by a 200-ms time constant, including temporal masking, intensity integration and loudness summation [6]. It is of interest that a similar time constant figures prominently in visual and motor function as well [27].

These observations suggest the existence of a quantal unit common to the sensory and motor systems, a unit of time over which sensory information is analyzed and correlated with the relevant motor systems, possibly through the reticular nuclear complex of the thalamus and the neo-dentate nucleus of the cerebellum.

Thus, the syllable may serve as the temporal unit for integration of auditory information into higher-level linguistic units.

During speech production the motor system controlling the vocal apparatus is almost surely in close communication with the output of the auditory system. The syllable may thus serve as the temporal unit for which the auditory and articulatory components of speech are synchronized, and also serve as well as the basic unit for higher level integration into semantic units.

Information encoded in syllable packets places a temporal constraint on linguistic information. It establishes this time frame as one in which a minimum amount of information needs to be fit for higher level integration.

These observations suggest the existence of a quantal unit common to the sensory and motor systems, a unit of time over which sensory information is analyzed and correlated with the relevant motor systems, probably through the reticular nuclear complex of the thalamus.

SUMMARY AND CONCLUSIONS

The human vocal apparatus is likely to have evolved under conditions optimized to produce communication signals possessing properties that exploit the auditory system's ability to encode information in a robust, fault-tolerant fashion.

The speech spectrum is biased towards the low-frequencies which are particularly resistant to disruption from background noise. The sound pressure level of most speech is sufficiently high as to insure that low-frequency spectral information is spread across a wide array of auditory frequency channels. Glottal periodicity insures that the system is able to track speech in noisy, acoustically adverse conditions, and syllable length modulation helps the brain bind together disparate spectral entities into meaningful units.

Within this framework, the importance of the auditory system for speech is that it preconditions the neural representation for maximum reliability and rate of information transmission. It does this by creating a sparse representation of the signal, consisting mainly of changes in spectral peaks and temporal parameters. The brain therefore only needs to keep track of novel features in the waveform, confident that only these encode important information.

Is this correlation between auditory properties and the speech waveform sufficient to fully account for the acoustic properties of human language? Probably not. Although the auditory system necessarily provides the sort of robust, efficient form on information representation required for higher level linguistics integration, it fails to fully specify why speech occurs in syllable and word level units.

Other brain centers, such as the thalamus and cortical association areas are undoubtedly involved in the transformation of this acoustic information into a complex symbolic system.

REFERENCES

[1] Pickett, J. (1980), *The sounds of speech communication*, Baltimore: University Park Press.

[2] O'Shaughnessy, D. O. (1987), *Speech communication: human and machine*, Reading, MA: Addison-Wesley.

[3] Miller, G. (1951), *Language and communication*, New York: McGraw-Hill.

[4] Fletcher, H. (1953), *Speech and hearing in communication*, New York: van Nostrand.

[5] Lehiste, I. (1970), *Suprasegmentals*, Cambridge, MA: MIT Press.

[6] Moore, B. C. J. (1989), *Introduction to the psychology of hearing* (3rd ed.), London: Academic Press.

[7] Dallos, P. (1973), *The auditory periphery: biophysics and physiology*. New York: Academic Press.

[8] Erulkar, S. D. (1972), "Comparative aspects of spatial localization of sound", *Physiological Reviews*, vol. 52, pp. 237-360.

[9] Masterton R. B.. (1974), "Adaptation for sound localization in the ear and brainstem of mammals", *Federation Proceedings*, vol. 33, pp. 1904-1910.

[10] Palmer, A. R and Russell I. J. (1986), "Phase-locking in the cochlear nerve of the guinea-pig and its relation to the receptor potential of inner hair-cells", *Hearing Research*, vol. 24, pp. 1-15.

[11] Johnson, D. (1980), "The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones", *J. Acoust. Soc. Am.*, vol. 68, pp. 1115-1122.

[12] Palmer A. R. and Evans E. F. (1980), Cochlear fibre rate-intensity functions: no evidence for basilar membrane nonlinearities, *Hearing Research*, vol. 2, pp. 319-326.

[13] Sachs M. B. and Young E. D. (1979), "Encoding of steady-state vowels in the auditory nerve: representation in terms of discharge rate", *J. Acoust. Soc. Am.*, vol. 66, pp. 470-479.

[14] Greenberg S, Geisler C. D and Deng, L. (1986), Frequency selectivity of single cochlear-nerve fibers based on the temporal response pattern to two-tone signals, *J. Acoust. Soc. Am.*, vol. 79, pp. 1010-1019.

[15] Geisler C. D and Greenberg S. (1986), "A two-stage nonlinear cochlear model possesses automatic gain control", *J. Acoust. Soc. Am..*, vol. 80, pp. 1359-1363.

[16] Young E. D. and Sachs, M. B. (1979), "Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers", *J. Acoust. Soc. Am.*, vol. 66, pp. 1381-1403.

[17] Smith R. L. (1977), "Short-term adaptation in single auditory nerve fibers: some poststimulatory effects", *J. Neurophys.*, vol. 40, pp. 1098-1111.

[18] Harris D. M. and Dallos P., (1977), "Forward masking of auditory nerve fiber responses", *J. Neurophys.*, vol. 42, pp. 1083-1107.

[19] Rhode, W, S, Oertel D. and Smith P. H. (1983) Physiological response properties of cells labeled intracellularly with horseradish peroxidase in cat ventral cochlear nucleus. *J. Comp. Neurol.*, vol. 213, pp. 448-463.

[20] Langner, G and Schreiner C. E. (1988) Periodicity coding in the inferior colliculus of the cat. I. Neuronal mechanisms. *J. Neurophys.*, vol. 60, pp. 1799-1822.

[21] Bacon S. P. and Viemeister N. F. (1985) "Temporal modulation transfer functions in normal-hearing and hearing-impaired listeners", *Audiology*, vol. 24, pp. 117-134.

[22] Bregman, A. (1990) *Auditory scene analysis*. Cambridge: MIT Press.

[23] Kiang, N. Y.-S. (1965) *Discharge patterns of single fibers in the cat's auditory nerve*. Cambridge, MA: MIT Press.

[24] Rhode, W. S. and Smith, P. H. (1986) "Encoding timing and intensity in the ventral cochlear nucleus of the cat", *J. Neurophys.*, vol. 56, pp. 261-286.

[25] Holton, T. (1995) "A computational approach to recognition of speech features using models of auditory signal processing", *Proc. Int. Cong. Phon. Sci.*

[26] Schreiner C. E. and Urbas J. V. (1986) "Representation of amplitude modulation in the auditory cortex of the cat. I. The anterior auditory field (AAF)", *Hearing Research*, vol. 21, pp. 227-41.

[27] Kandel, E. R. and Schwartz, J. H. (1985) *Principles of neural science* (2nd ed.) New York: Elsevier.

PSYCHOPHYSICS OF SPEECH ENGINEERING SYSTEMS

H. Hermansky and M. Pavel

Oregon Graduate Institute, Portland, Oregon, USA

ABSTRACT

The paper reviews two engineering techniques, the Perceptual linear predictive (PLP) analysis and the RelAtive SpecTrAl (RASTA) processing, used in automatic speech recognition and describe their consistencies with some properties of human speech perception.

INTRODUCTION

Assuming that speech developed so that its linguistically important components are heard well, processing of speech should respect properties of human hearing. However, a blind copying of nature without deeper understanding of underlying mechanisms in hopes of "obtaining" a successful engineering solution has frequently proven to be a failure[1].

We believe that engineeringdisciplines can benefit from *selective* modelling of *relevant* characteristics of human information processing[2]. In this paper we discuss two techniques, the Perceptual linear predictive (PLP) analysis, and the RelAtive SpecTrAl (RASTA) processing, which were designed to improve performance of automatic speech recognizers. Subsequently, these techniques were found to be consistent with specific properties of human speech perception. We discuss (in italics) relevant properties of human speech perception, and before describing the two we attempt to put both techniques into historical perspective with selected engineering systems.

[1] Airplanes do not flap wings, and most of "auditory models" do not demonstrate significant advantage in engineering, and sometimes yield clearly inferior results.

PERCEPTUAL LINEAR PREDICTION (PLP)

The PLP analysis technique was designed to suppress speaker dependent components in features used for automatic speech recognition. Several basic properties of human hearing (as noted bellow, each previously used in engineering) were integrated in a speech analysis technique called PLP [1].

Root Spectral Compression

Perception of intensity appears to be consistent with a compressive type of nonlinearity. In particular, perceived loudness of a steady sound is approximately proportional to a cube root of its power [2].

Lim [3] investigated the use of different compressive functions in homomorphic analysis of speech. He concluded that the cube root compression was optimal with respect to resulting speech quality of re-synthesised speech.

Hermansky et al. [4] experimented with varying compressive functions in linear predictive analysis and found that when the short-term power spectrum of speech is compressed through cube root function, the analysis is the least affected by the fine spectral structure of voiced speech. The root spectral compression also helps in modelling spectral envelope zeros which occur in nasalized and fricative speech sounds.

Furthermore, root compressed power spectrum (root compression with exponents 2-4) appears to be optimal for processing which alleviates additive noise in the acoustic signal (see e.g. [5-7])

[2] Airplanes do not flap their wings, but their design is based on thorough understanding and use of principles of aerodynamics which allow birds to fly.

Nonlinear spectral resolution

Decreasing selectivity of human hearing with frequency is one of the best documented and least disputed properties of human auditory perception.

Bridle and Brown [8] and later Mermelstein [9], and Davis and Mermelstein [10] proposed to use cosine transform of logarithmic energies (cepstrum) from non-uniformly spaced bandpass filters with bandwidth increasing with frequency. Davis and Mermelstein proposed triangular filters with a shape which is about constant on the mel scale. Mel cepstrum is currently the dominant feature extraction technique in automatic speech recognition.

Nonuniform spectral sensitivity of hearing

For typical levels of human speech communication, hearing is most sensitive in 2-4 kHz range, therefore emphasising the second and third formant region.

A typical preemphasis in speech analysis approximates this property by 6dB/oct high-pass filtering of the signal.

To obtain more stable formant estimates, Itahashi and Yokoyama [11] proposed to warp the spectral envelope of speech (estimated by high-order LPC analysis) using a mel warping function, and weight it by an approximation of the Fletcher-Munson equal loudness curve. The resulting auditory-like spectrum was then again approximated by relatively low (6th order) LPC all-pole model.

Broad spectral integration in speech perception.

Klatt [12] *speculated that for gender normalization, larger than 1 Bark spectral resolution would be required. This notion is supported by perceptual studies that suggest that human speech perception could integrate formant peaks within 3.5 Bark interval* [13], *and therefore could merge several speech formants. Thus, frequency resolution for perception of speech signals seems to be considerably broader that the critical-band concept would suggest*

Pols et al. [14] reported that the first three (six) principal components of a set of non-uniformly spaced 1/3 octave filter bank output power explain 82% (97%) of variance in his data. Later work Pols [15] also shows that these first three principal components can be used successfully in automatic speech recognition.

The Technique

Several engineering approximations to the properties of human speech perception are used in PLP analysis of speech:

1) critical band (Bark) nonlinear frequency resolution, implemented by integrating short-term Fourier spectrum of speech under increasingly wider trapezoidal curves,

2) asymmetries of auditory filters, implemented by relatively steep (25dB/Bark) slope of the trapezoidal curve towards higher frequencies and more gradual (10dB/Bark) slope towards lower ones,

3) unequal sensitivity of human hearing at different frequencies, implemented by a fixed approximation of Fletcher-Munson equal loudness curve,

4) intensity-loudness nonlinear relation, implemented by a cube root compression, and

5) broader than critical-band integration, hypothesised in perception of speech (see e.g. [12]), implemented by an autoregressive all-pole model.

The optimal order of the PLP all-pole model was determined experimentally on cross-speaker speech recognition experiments in which training data from one speaker were used to recognise speech of another speaker. Results are shown in Fig. 1. The two-peak (5th order) model was found to be optimal.

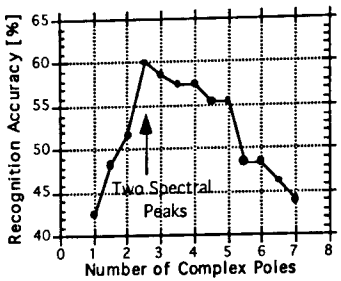


Fig. 1 Dependency of recognition accuracy in cross-speaker experiment on the maximum number of spectral peaks (model order) of PLP model

The spectrograms in Fig. 2 show that, in comparison to the conventional formant based representations, the broader spectral integration implied by low-order PLP analysis is capable of more consistent speech representations from adult and child speech.

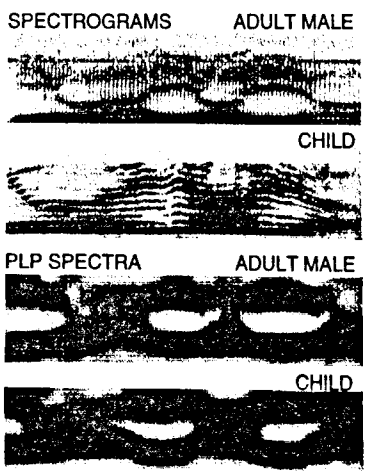


Fig. 2 Spectra of adult and child speech obtained by conventional spectral analysis and by PLP analysis

The 5th order PLP model was used successfully in speaker-independent recognition of digits [1]. For more complex tasks with a sufficient amount of training data, higher model order (7th-8th) appears to be more efficient.

RELATIVE SPECTRAL (RASTA) PROCESSING

We will next describe our engineering approach based on certain temporal properties of human hearing.

Perception of modulated signals

Since early experiments of Riesz [16] *it is known that sensitivity of human hearing to both the amplitude and the frequency modulation is highest for frequency of modulation at about 4-6 Hz. Thus, human hearing in perception of modulated signals acts as a band-pass filter.*

Drullman et. al [17, 18] *support the band-pass character of human hearing in speech perception by showing that low-pass filtering of 1/4 octave-derived spectral envelopes of speech at frequencies higher than 16 Hz or high-pass filtering it at frequencies lower that 2 Hz causes almost no reduction in speech intelligibility. They proposed that that the bulk of linguistic information is contained in modulation frequencies between 2 and 16 Hz.*

Furui [19] introduced delta features to enhance dynamic components of speech signal. To approximate the derivative of time trajectories of cepstral coefficients, Furui computed the delta features using a regression fit to a short segment of the cepstral trajectories. This operation is equivalent to band-pass filtering of the trajectory by an FIR filter with a relatively shallow (-6db/oct) low frequency slope. The optimal length of the segment for deriving the regressionfit was about 170 ms, which corresponds to a FIR bandpass filter with its maximum at about 4 Hz [20].

Rosenberg et al [21] experimented with cepstral mean subtraction in speaker recognition system using mean

computed over short-term window of variable lengths. They reported the best results for the window of 165 ms. As discussed e.g. in Hermansky and Morgan [20], the cepstral mean subtraction with the 165 ms window implies high-pass filtering with the filter cut-off frequency of about 1 Hz.

The Technique

RASTA engineering technique uses the fact that linear distortions and additive noise in speech signal show as a bias in the short-term spectral parameters. Since rate of such extra-linguistic changes is often outside the typical rate of change of linguistic components, Hermansky et al [22] and Hirsh et al. [23] have proposed filtering of temporal trajectories of speech parameters which would alleviate the extra-linguistic spectral components from the speech representation. This technique is known as RASTA speech processing. A series of recognition experiments in which the test data were linearly distorted by convolution with a simple first-order high-pass system [20] was run with different RASTA filters to determine the optimal filter structure. Results of experiments are shown in Fig. 3.

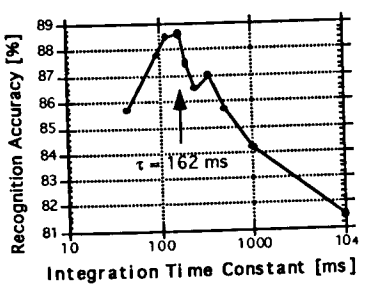


Fig. 3 Dependency of recognition accuracy in presence of linear distortions on time constant of integrator of RASTA filter.

The optimal filter for recognition of noisy speech was found to be a bandpass filter with the pass-band between about 1 Hz and 12 Hz. The time constant of the integrator in the filter was about 170 ms. RASTA processing enhances dynamic events is the signal and suppresses the slowly varying ones, as illustrated in Fig. 4.

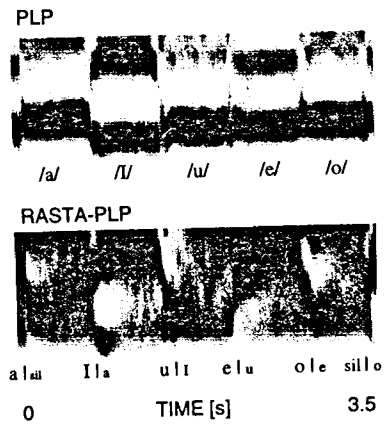


Fig. 4 Spectra of five sustained Czech vowels obtained by PLP and RASTA-PLP analyses. Note enhanced transitions resulting from RASTA processing

The RASTA band-pass filtering is typically done either on logarithmic spectrum (or cepstrum, which is a linearly transformed logarithmic spectrum) or on the spectrum compressed by *ln(const+x)* nonlinearity. However, Hermansky et al. [7] reported that RASTA filtering on root-compressed power spectrum (with filters designed from the training data) is effective for perceptual enhancement of noisy telephone speech. Filters in the frequency range with most speech energy have a maximum at about 6-8 Hz.

For speech recognition applications, we most often use RASTA processing in combination with the above described PLP technique. In this combination, RASTA filtering is performed on outputs from a critical-band analysis, i.e.,

prior to the cube root compression and loudness equalisation, and the all-pole modelling. We note that RASTA-PLP technique is rapidly gaining recognition in engineering community, especially in applications which can tolerate or even benefit from the enhanced spectral dynamics, such as the isolated phrase recognition. Alternative recognition paradigms which could capitalise on the enhanced spectral dynamics are being studied for applications of RASTA processing in recognition of continuous speech [24].

CONSISTENCIES WITH HUMAN SPEECH PERCEPTION

Although both PLP and RASTA were designed on purely engineering grounds and with a clear engineering objective in mind, they both turned out to be at least in certain aspects consistent with human speech perception.

PLP and effective perceptual second formant F2'

Fant and Risberg [25] *observed that all Swedish vowels can be simulated by synthetic stimuli with only two spectral peaks, providing that their second spectral peak F2' is in particular position, which does not necessarily coincide with any of the formants. Fant* [26] *proposes that the effective second formant F2' might correspond to a resonance frequency of the uncoupled front cavity of the vocal tract. Hermansky and Broad* [30] *showed on X-ray tracings that the front cavity appears to be less dependend on the age of the talker than the rest of the vocal tract. They speculated that speech perception (simulated by the PLP analysis) might be able to integrate detailed formant structure and extract the resonance frequency of the front part of the vocal tract.*[3]

[3] Just as is more or less accepted that the formants are extracted by some form of integration of a fundamental frequency peaks.

The 5th order PLP analysis of 18 synthetic cardinal vowels yields results which agree well with Bladon and Fant's [27] perceptual experiments: the second spectral peak approximates well the effective second formant F2' [1]. Moreover, the bandwidths of the PLP model preserve information about spread of the underlying formant clusters, thus alleviating a fundamental objection [28, 29] to the F2' concept (see [1] for evidence and discussion). The two peaks of the 5th order PLP model start merging when their distance approaches 3.5 Bark, thus being consistent with [13]

Hermansky and Broad [30] demonstrate a high correlation between positions of the second spectral peak of the 5th order PLP model and the resonance frequency of the uncoupled front cavity of the simulated vocal tract of front and mid vowels, used in articulatory synthesis of the vowel-like sounds. Table 1. is a summary of their results. The first row contains correlations of the tract legnth and the resonance frequency of the uncoupled front cavity with the second peak of PLP model, extracted from the synthesized speech. The second row shows averaged correlations with the first four formants. Note that the formant frequencies, which are strongly dependent on anatomy of the particular vocal tract, correlate highly with the tract length. The weak correlation of the second peak of the PLP model with the tract length implies its relative independence of the talker. Its strong correlation with the resonance frequency of the uncoupled front cavity supports Fant's proposal of its correspondence with the effective second formant F2' [26].

Table 1.

	Tract Length	Front cavity resonance
Second Peak of PLP Model	-0.18	0.9
Formants (Averaged)	-0.71	0.22

Later [31] they also show a high correlation of the PLP-estimated F2' with the front cavity resonance estimated from the x-ray microbeam data. Additional work is needed to get full support for their hypothesis.

RASTA and forward masking

If a loud sound is followed closely in time by weaker sound, the audibility of the weaker sound is diminished. This effect, called forward masking, reflects a significant nonlinearity since, independently of the masker amplitude, the effect seems to last for about 200 ms (see e.g. [32]).

As we noted earlier, the phenomenon of forward masking reflects aspects of temporal properties of the auditory system. Forward masking effect is typically measured by presenting, on each trial, a masker (tone or band-passed noise) for 200 milliseconds or longer. Human observers are asked to detect a brief probe presented after a variable delay following the offset of the masker. The masking effect is summarised by the sound level of the probe, above its threshold, required for fixed detection performance.

Typical data from such experiments exhibit features that implicate non-linear aspects of the auditory system. For short delays, the masking effect is determined by the masker level. However, the masking effect decays rapidly, and becomes negligible for delays greater than 200 milliseconds, independent of the masker level. The decaying dependence of the masking effect on the logarithmic delay is well approximated by a set of straight lines that intersect at a point corresponding to the delay of approximately 200 milliseconds. This is illustrated in Fig. 6 by the shaded triangle which was derived from extrapolated mean human data for 1kHz and 30-60 dB SPL maskers (experiment 1 in [34]).

Prior attempts to account for the data led researchers to models based on automatic gain control such as proposed by [32]. In his model, the effect of the masker was to reduce temporarily the system gain. Although this model could account for the temporal behaviour of forward masking data, it did not specify a plausible process for the temporal dependency of the gain.

A decade later, a scrutiny of the RASTA engineering model provided two interesting insights [33]. First, a reduction in gain in the AGC model is equivalent to a subtraction preceded by a logarithmic transformation. Second, exponential decay in the logarithmic domain with appropriate choices of time constant can produce data that closely approximate linear decay. Both such operations are implemented in the RASTA model.

To investigate the potential of RASTA processing for modeling the temporal masking effect, we duplicated a part of experiment 1 from [34]. Critical-band spectra were computed by PLP analysis using 1 kHz stimuli. The critical-band spectra were processed by our standard RASTA filter [20]. Probe detection was mediated by a comparison of a spectral distance measure of RASTA processed loudness profiles (critical-band spectra in cube-root power) of a masker alone and of the masker followed by a probe. The process is illustrated in Fig. 5.

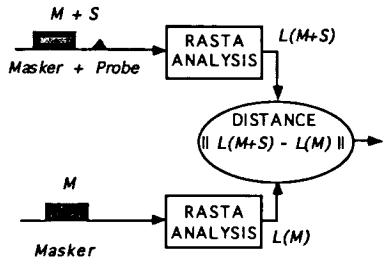


Fig. 5 A model of the experiment for investigation of temporal properties of RASTA processing.

Results, shown in Fig. 6, are qualitatively consistent with conclusions from human forward masking experiments [34] which implications are indicated in the figure by the shaded triangle overlaid over our data. To obtain the fit, we allowed for a linear

optimization of the distance measure, i.e. the actual Euclidean distance between loudness profiles was multiplied by a constant. (0.12) and another small constant (0.9) was added to the result.

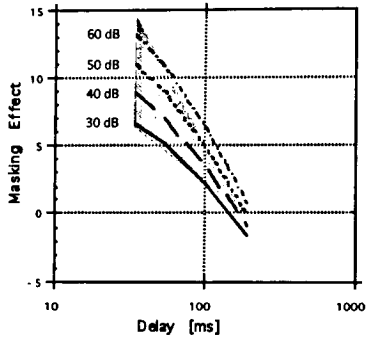


Fig. 6. Spectral distances betweena tonal masker and the masker with a probe. The parameter is the level of the masker. Extrapolated human performance [34] is shown by shaded area.

CONCLUSIONS

We have reviewed two successful engineering approaches, designed to alleviate sensitivity of speech processing to extra-linguistic factors and used widely in speech engineering. We noted the similarity in their behaviour to that of the human auditory system. Some of these consistencies were obtained because of an explicit motivation to model the human auditory system, but others were strictly the results of engineering optimisations.

We have also noted that analysis of engineering systems may lead to new insights into the processes underlying human auditory perception. There are instances where engineering technique, even though designed only as a practical solution to a particular engineering problem, turned out to be a good model of human auditory perception.

ACKNOWLEDGEMENTS

This work has been supported in part by NSF-ARPA Grant IRI-9314959 to Oregon Graduate Institute. The authors thank Steven Greenberg for pointing out the relation of RASTA processing to perception of modulation and for useful editorial suggestions.

REFERENCES

[1] Hermansky, H., *Perceptual linear predictive (PLP) analysis of speech.* J. Acoust. Soc. Am., 1990. **87**(4): p. 1738-1752.

[2] Stevens, S.S., *On the psychophysical law.* Psychol. Rev., 1957. **64**: p. 153-181

[3] Lim, J.S., *Spectral root homomorphic deconvolution system.* Proc. IEEE ASSP-27, 1979. **27**(3): p. 223-233.

[4] Hermansky, H., H. Fujisaki, and Y. Sato. *Analysis and synthesis of speech based on spectral transform linear predictive method.* in *Int. Conf. Acoust. Speech and Sig. Proc.* 1983.

[5] Porter, J.E. and S.F. Boll. *Optimal estimators for spectral restoration of noisy speech.* in *Int. Conf. Acoust. Speech and Sig. Proc. 84.* 1984.

[6] Hanson, B. and D. Wong. *The harmonic magnitude suppression (HMS) technique for intelligibility enhancement in the presence in interfering speech.* in *Int. Conf. Acoust. Speech and Sig. Proc.* 1984.

[7] Hermansky, H., E. Wan, and C. Avendano. *Speech enhancement based on temporal processing.* in *Int. Conf. Acoust. Speech and Sig. Proc.* 1995.

[8] Bridle, J.S. and M.D. Brown. *An experimental automatic word recognition system.* in *JSRU Report No. 1003.* 1974. Ruislip, England: Joint Speech Research Unit.

[9] Mermelstein, P., *Distance measures for speech recognition, psychologicaland instrumental,* in *Pattern Recognition and Artificial Intelligence,* R.C.H. Chen, Editor. 1976, Academic Press: New York. p. 374-388.

[10] Davis, S.B. and P. Mermelstein, *Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences.* Trans. IEEE ASSP-28, 1980(4): p. 357-366.

[11] Itahashi, S. and S. Yokoyama. *Automatic formant extraction utilizing auditory equal loudness contour,.* in *Fall Meeting Acoust. Soc. Japan.* 1974.

[12] Klatt, D.H., *Speech processing strategies based on auditory models,* in *The representation of speech in the peripheral auditory system,* R. Carlson and B. Granstrom, Editors. 1982, Elsevier - Biomedical Press: New York. p. 181-202.

[13] Chistovich, L.A., *Central auditory processing of peripheral vowel spectra.* J. Acoust. Soc. Am., 1985. **77**: p. 789-805.

[14] Pols, L.C.W., L.J.T.v.d. Kamp, and R. Plomp, *Perceptual and physical space of vowel sounds.* J. Acoust. Soc. Am., 1969. **46**: p. 458-467.

[15] Pols, L.C.W., *Real-time recognition of spoken words.* IEEE Trans. Computers, 1971. **20**: p. 972-978.

[16] Riesz, R.R., *Differential intensity sensitivity of the ear for pure tones.* Phys. Rev., 1928. **31**(Ser. 2): p. 867-875.

[17] Drullman, R., J.M. Festen, and R. Plomp, *Effect of temporal envelope smearing on speech reception.* J. Acoust. Soc. Am., 1994. **95**: p. 1053-1064.

[18] Drullman, R., J.M. Festen, and R. Plomp, *Effect of reducing slow temporal modulations on speech reception.* J. Acoust. Soc. Am., 1994b. **95**: p. 2670-2680.

[19] Furui, S., *Cepstral analysis technique for automatic speaker verification.* IEEE ASSP-29, 1981(2): p. 254-266.

[20] Hermansky, H. and N. Morgan, *RASTA processing of speech.* Proc. IEEE Speech and Audio Processing, 1994.

[21] Rosenberg, A.E., C. Lee, and F.K. Soong. *Cepstral channel normalization techniques for HMM-based speaker verification.* in *International Conference on Spoken Language Processing.* 1994. Yokohama, Japan.

[22] Hermansky, H., N. Morgan, and P. Kohn. *Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP).* in *Eurospeech '91.* 1991. Genova.

[23] Hirsh, H.G., P. Meyer, and H. Ruehl. *Improved speech recognition using high-pass filtering of subbband envelopes.* in *Eurospeech '91.* 1991. Genova.

[24] Morgan, N., *et al. Stochastic perceptual models of speech.* in *Int. Conf. Acoust. Speech and Sig. Proc. 95.* 1995.

[25] Fant, G. and A. Risberg. *Auditory matching of vowels with two formant synthetic sounds.* in *STL-QPRS 2-3.* 1962. Stockholm: Royal Institute of Technology.

[26] Fant, G., *Acoustic theory of speech production.* 2nd printing ed. 1970, The Hague: Mounton. p.123.

[27] Bladon, A. and G. Fant. *A two-formant model and the cardinal vowels.* in *STL-QPRS.* 1978. Stockholm: Royal Institute of Technology.

[28] Fujimura, O., *On the second spectral peak of front vowels: a perceptual study of the role of the second and third formants.* Language and Speech, 1967: p. 10181-10193.

[29] Bladon, A.W., *Two-formant models of vowel perception: shortcomings and enhancements.* Speech Communication, 1983. **2**: p. 305-313.

[30] Hermansky, H. and D. Broad. *The effective second formant F2' and the vocal tract front cavity.* in *Int. Conf. Acoust. Speech and Sig. Proc.* 1989.

[31] Broad, D. and H. Hermansky, *The front cavity/F2' hypothesis tested by data on tongue movements.* J. Acoust. Soc. Am., 1989. **86**(Suppl. 1): p. S13-S14.

[32] Pavel, M., *Homogeneity in complete and partial masking,* . 1980, New York University.

[33] Pavel, M. and H. Hermansky, *Temporal masking in automatic speech recognition.* J. Acoust. Soc. Am, 1994. **95**(5): p. 2876.

[34] Jestead, W., Sid P. Bacon, and James R. Lehman, *Forward masking as a function of frequency, masker level, and signal delay,* J. Acoust. Soc. Am, 1982,

A COMPUTATIONAL APPROACH TO RECOGNITION OF SPEECH FEATURES USING MODELS OF AUDITORY SIGNAL PROCESSING

Thomas Holton
School of Engineering, San Francisco State University, San Francisco, CA 94132 USA

ABSTRACT

We present a computational approach to the detection of important speech features, such as formants and pitch, based on a model of auditory signal processing. Algorithms have been designed to be computationally simple, physiologically reasonable and to emulate human psychophysical performance.

INTRODUCTION

Most current approaches to computer speech recognition are based on a spectrographic approach to feature extraction[1]. In this approach, the energy of speech is measured as a function of frequency, and parameters derived from the resulting spectrum are compared to a template or rule. Spectrographic techniques include computation of FFTs, extraction of LPC and cepstral coefficients and processing by filter banks.

Spectrographic approaches suffer from well-known problems. Because spectrograms are sensitive to anything that changes the relative magnitude of in-band energies, their performance is often severely degraded in situations of practical interest; for example, in conditions of reduced spectral bandwidth (over the phone) or in the presence of background or line noise.

In our approach, we have sought to understand the fundamental strategy used by the auditory system to process speech signals and apply this understanding to the design of improved algorithms for detection of speech features. We have:

- developed a comprehensive model of signal processing by the peripheral and early central auditory system,
- studied the response of this model to speech and other stimuli, and
- distilled what we believe are important signal processing techniques of the auditory system into practical algorithms for feature extraction that provide noise-immune, speech-specific detection of formants and pitch pulses in sonorant parts of speech.

RESULTS

A model of auditory signal processing

The model of auditory signal processing[2] includes components describing the external and middle ear, a detailed three-dimensional hydro-mechanical model of the cochlea, a biophysical model of mechano-electric transduction by the cochlear hair cells, a description of the time-dependent synaptic chemistry of hair cells and auditory-nerve fibers including models of the hair cell's calcium channel and synapse and a 'micro-neural-net' description of signal processing in the cochlear nucleus. A comparison of the predictions of this model with experimental physiological data in response to both simple stimuli (i.e. tones) and complex stimuli (i.e. speech) suggests that the model adequately describes essential features of auditory signal processing.

The response of the model to /a/

Figure 1 shows the response of the auditory model to a voiced utterance, /a/, spoken by a male speaker. The model response to this utterance comprises two distinct spatio-temporal patterns occurring in alternation. We term these patterns the *impulsive epoch* and the *synchronous epoch*. The impulsive epoch occurs in response to the glottal pulse. In this epoch, most fibers respond at a rate that

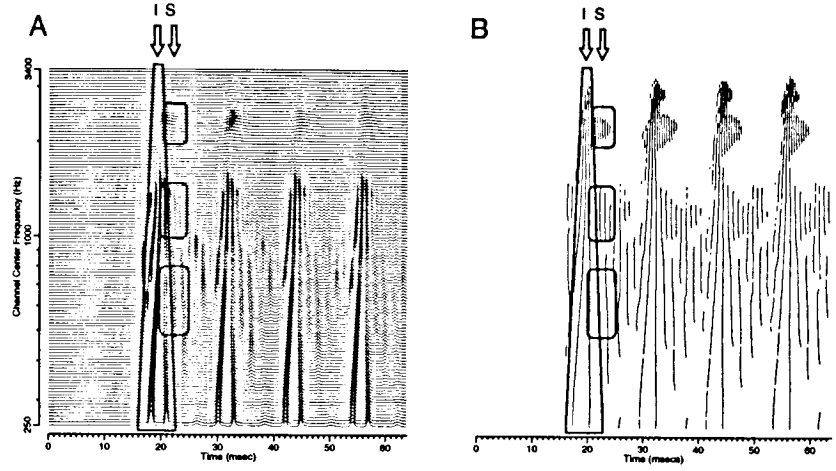


Figure 1. A. The response of 120 model fibers with characteristic frequencies (CFs) spanning the range from 250 Hz (bottom trace) to 3.4 kHz (top trace). Each waveform represents the probability density function of neural discharge for an auditory nerve fiber innervating one location along the cochlea which is maximally sensitive to a particular CF. The responses of fibers have been time aligned to remove the delay that results from the transit time of sound along the basilar membrane and the delay of neural response. The impulsive (I), *and* synchronous (S) *epochs are marked. B. The times at which the nerve fiber ensemble in A is most likely to fire. This plot results from processing the waveforms of A with a threshold-crossing algorithm that places a tick mark at the times at which each fiber is most likely to fire. The plot gives a stylized description of the pattern of timing information that this ensemble of fibers delivers to the brain in response to /a/.*

corresponds with their best or characteristic frequency (CF), giving the pattern of response of the ensemble of fibers a splayed appearance. In the synchronous epoch which follows, several groups of fibers respond distinctly at a rate that corresponds to the frequency of a proximal formant. We poetically term each group of fibers entrained to one formant an "island of synchrony". There appear to be at least three sharply delineated islands of synchrony: fibers with CFs between approximately 500 and 800 Hz are synchronized to F1; fibers with CFs between 1000 and 1400 Hz are synchronized to F2; fibers with CFs above 2000 Hz are synchronized to F3.

The alternation of an impulse-like pattern with a synchronous pattern is highly characteristic of the response to voiced speech. These observations suggest that the alternately impulsive and synchronous nature of the model's response could be used to locate and track linguistically interesting quantities such as the times of occurrence of pitch pulses and the frequencies of the formants. Our approach has been to build separate physiologically motivated "detectors" for the impulse-like first epoch and the synchronous second epoch and then use these detectors to identify formants and pitch pulses.

The response of the auditory model to an impulse

Figure 2 shows the response of the

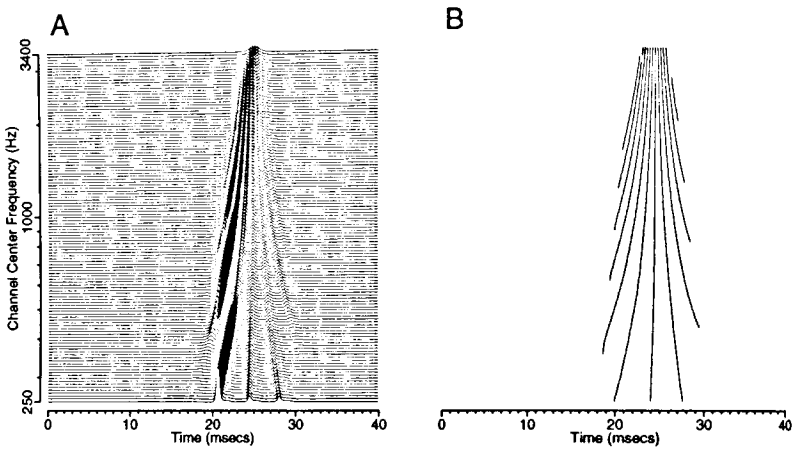


Figure 2. A. The response of the auditory model to an impulse. B. The result of processing the waveforms in A through a threshold-crossing algorithm that puts a tick mark at the times each fiber is most likely to fire.

auditory model to an impulse. Examined individually, every fiber tends to respond at a rate equivalent to its own CF. An engineering approach to designing a detector of this impulsive epoch might be to assert that an impulse is detected if, at any moment, enough fibers respond at a rate equivalent to their own CF. Algorithmically, one might implement this by computing an interval histogram of the time between firings for each fiber, taking the inverse to get a distribution of firing rate and extracting the dominant frequency component by a transform method[5]. However, there is no evidence that the brain has any processes analogous to those of forming or inverting histograms, or performing transforms to extract frequency components.

What the brain most likely can do is to detect patterns occurring in the response of a large number of simultaneously active parallel channels. We suggest that what is interesting about this picture is not an *individual* fiber's response, but the pattern of response of the *ensemble* of fibers. Specifically, the cochlea's response to an impulse is characterized by a "splayed" pattern of firing: before the peak of the impulse, fibers of lower CF respond before those of higher CF; after the peak of the impulse, fibers of lower CF respond after those of higher CF. In order to detect this pattern, we propose an array of cells, each of which correlates the response from a small number of adjacent channels and produces an output when this sequential, tonotopically organized pattern of firing is seen in the input for a period of time. The signal processing operations involved here are simple, physiologically reasonable time-correlation pattern detections; this approach does not require the computation of non-physical quantities like histograms and transforms.

While it is possible to build a detector that finds impulsive features in the stimulus using the approach just outlined, there are two practical problems with this idea: 1) *Computational complexity:* generating waveform plots such as those in Figure 1 requires the solution of a system of nonlinear, time-varying differential equations that specify the cochlear-mechanical, hair-cell and neural components of the model. The solution of these equations is highly computationally intensive; 2) *Temporal granularity:*

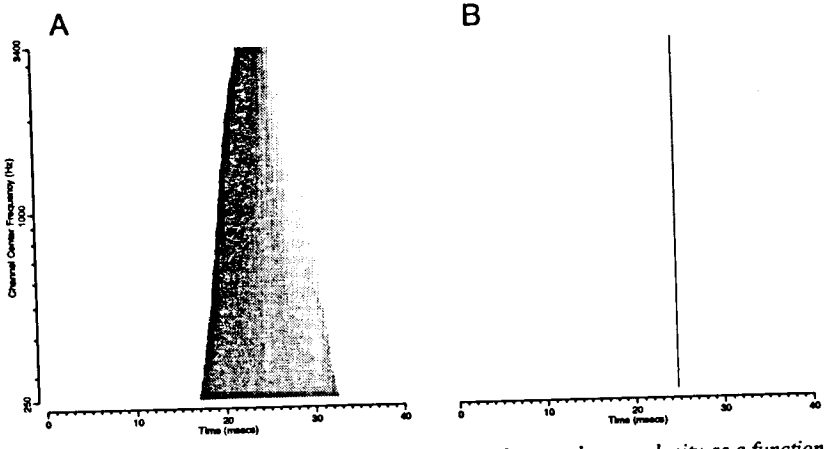


Figure 3. A. The spatial derivative of the phase of basilar-membrane velocity as a function of time in response to an impulse. The ordinate corresponds to the 120 locations along the basilar membrane with CFs logarithmically distributed between 250 Hz (bottom) and 3.4 kHz (top). Negative phase-velocity is plotted dark and positive phase velocity is light. Before the peak of the impulse the phase velocity is uniformly negative, and becomes uniformly positive after the peak of the impulse. B. The response of 119 local impulse detectors to an impulse. Each local impulse detector continuously examines the spatial phase velocity computed from the response of a pair of adjacent channels. An impulse is said to be detected when the spatial phase velocity becomes greater than zero after increasing monotonically for a period of at least one millisecond. This event corresponds to detecting the splayed pattern of nerve-fiber firings seen in Figure 2.

because model neural firings occur at discrete times, the estimate of the time of occurrence of the impulse has considerable temporal uncertainty or granularity.

To solve these problems we have used an important result derived from the study of the response of the cochlear model: patterns of neural firings correlate with patterns of basilar membrane motion; specifically, information about the sequential or simultaneous firings of groups of adjacent fibers reflects simple patterns in the spatial and temporal derivatives of the instantaneous phase of the basilar membrane's motion.

The local impulse detector

Figure 3A shows a plot of the spatial derivative of the phase of basilar-membrane velocity as a function of time in response to an impulse. At all points on the model cochlea, the spatial phase velocity is initially less than zero and increases monotonically over a period of time. This pattern of phase velocity is easy to detect. Figure 3B shows the response of an array of local impulse detectors. Each detector produces a response upon detecting the negative-to-positive pattern of spatial phase velocity.

Figure 4 shows the response of an array of local impulse detectors to */a/*. Each mark on the plot is derived by examining local spatial and temporal patterns of phase velocity over a small window of time (about 1.5 msec) and a small range of frequency (two adjacent channels, corresponding to about 0.3 critical bands). The wavy lines correspond to the times of occurrence of the pitch pulses.

The local formant detector

It is possible to use the same auditory model concepts to make phase-based local

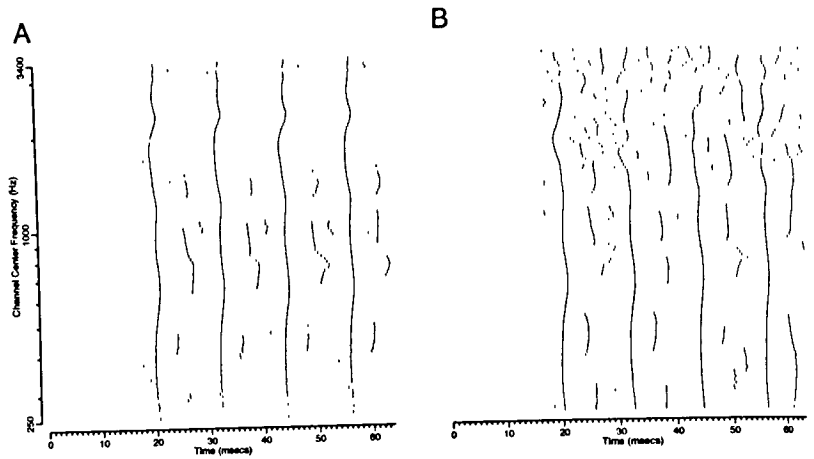


Figure 4. A. The response of the 119 local impulse detectors to /a/. B. The response of the impulse detectors to /a/ in the presence of white noise with S/N = 8 dB.

synchrony detectors which find the "islands of synchrony" discussed earlier. Specifically, it can be shown that the islands of synchrony correspond to spatio-temporal regions in which the spatial phase velocity of basilar-membrane motion is constant.

Voiced speech comprises impulsive epochs and synchronous epochs occurring in alternation. We have built a candidate formant detector that detects this pattern based on the detectors for the impulsive and synchronous epochs described above. The formant detector is an array of cells, each of which responds to an impulsive epoch in a given channel followed by a synchronous epoch.

Figure 5 shows the output of the local formant detector to /a/. Three formants (F_1, F_2, and F_3) plus a bit of F_4 are clearly represented. The representation of information in this plot is quite sparse; there is only information at frequencies corresponding to the formants and little elsewhere. None of the operations involved in generating this representation are either computationally complex or non-physiological, and none of the operations uses any of the conventional spectral techniques.

In natural speech, the frequencies of formants are not static, but change rapidly as a function of time depending on the consonantal context in which the vowel is embedded. Because all the stages of detection that generate this representation act on patterns which are temporally localized, the speech signal need not be periodic or quasi-periodic to determine the times of occurrence and frequencies of the formants. In this approach, formants are detected on a pitch-pulse-by-pitch-pulse basis with simultaneously high time and frequency resolution.

Human speech intelligibility, at least of vowels, is not very sensitive to additive background noise. Whereas spectrographic representations of speech are inherently sensitive to noise, the response of the local formant detector is relatively insensitive. Also, unlike spectrographic measures, it can be shown that the response of the formant detector is insensitive to pure tones and other non-speechlike stimuli.

A model of pitch

We have developed a theory for the detection and identification of pitch and

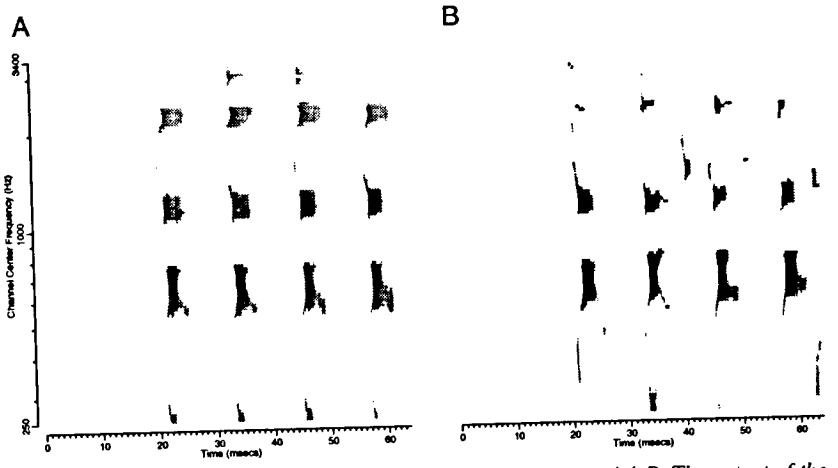


Figure 5. A. The output of an array of local formant detectors to /a/. B. The output of the formant detector to /a/ in the presence of additive noise, S/N=+8dB.

voicing based on the physiological model of auditory signal processing coupled with the idea of detecting spatially and temporally local patterns of response phase from a number of parallel channels[3].

Figure 6 shows the architecture of a physiologically motivated pitch detector. For each point on the cochlea, we postulate the existence of an array of *local pitch detector* cells. Each cell in the array detects in the time domain a different fixed time periodicity in the output of the underlying local impulse detector cell. These pitch-detector cells could be implemented physiologically by a series of neural delay correlators, as originally proposed by Licklider[4]. For each point on the cochlea, cells in the pitch detector array respond when a pair of impulses is received in the same channel with a given fixed time delay. Cells in the current model are selective for time delays spanning the range of 1 to 15 msec with a resolution of .25 msec. The sum of the response of local pitch detectors serving the whole cochlea gives a global measure of the periodicity of the entire ensemble of channels, which we term the *global pitch* detector.

Using this pitch detector method, it is possible to track rapidly varying pitch of natural speech. Figure 7 shows the output of the global pitch detector, a representation of the instantaneous average pitch frequency as a function of time, for an utterance that has relatively constant formant structure but rapidly varying pitch. The response of this pitch detector can be shown to be robust in noise. While the pitch detector is particularly sensitive to impulsive stimuli, such as voiced speech, it is highly insensitive to pure tones and other non-speech-like input. The pitch detector also reproduces effects seen in the psychophysics of pitch perception, such as the recovery of the missing fundamental of resolved and unresolved harmonics.

The auditory-model pitch detector is computationally straightforward and physiologically plausible. Calculations correspond to the correlation of simple neural events. No continuous-time autocorrelation functions are explicitly computed, nor does the input stimulus need to be periodic for pitch to be detected and tracked.

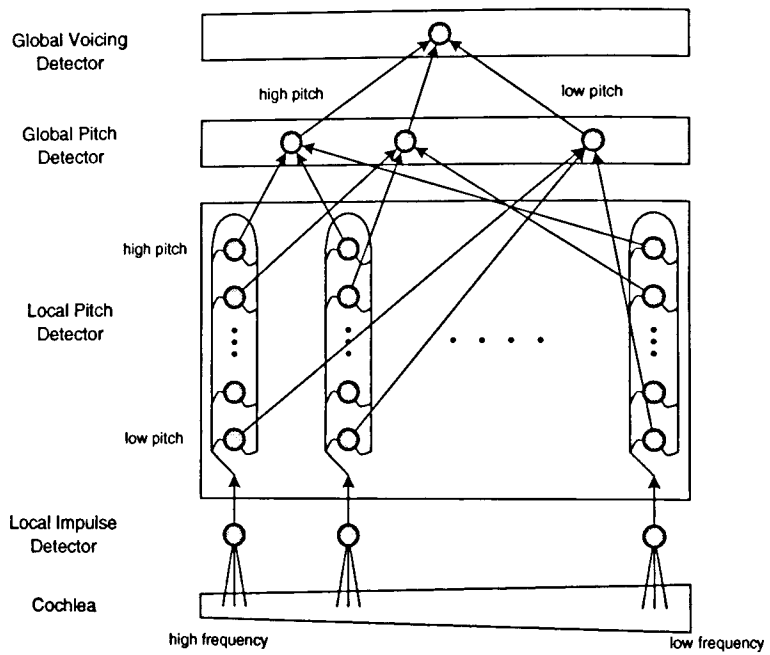


Figure 6. Architecture of the pitch detector

CONCLUSION

We have designed algorithms for the detection of important speech features based on an understanding of how the auditory system processes speech. Algorithms are computationally simple, physiologically reasonable and demonstrate performance that emulates that of humans.

Comparison with spectrographic approaches to feature extraction

Almost all current approaches to speech recognition are based on a spectrographic approach to feature extraction. These techniques include filter bank, fast Fourier transform (FFT), cepstral, power spectral density (PSD) and linear predictive coding (LPC) analysis. These spectrographic approaches are sensitive to anything that changes the magnitude of the input in a frequency band, for example by spectral shaping the input signal. Spectral approaches are sensitive to the frame size of analysis; a larger frame size may be used to average over pitch periods at the cost of coarser temporal and spectral resolution. Spectrographic approaches are also inherently noise sensitive, since they measure the energy in a frequency band, regardless of the source of that energy.

The auditory-model approach to detecting speech features differs in key respects from spectrographic methods. This approach, based on building detectors of spatially and temporally local patterns of response phase from a number of parallel channels, can be characterized as a *local time-domain phase-correlation* approach, in contrast with conventional spectrographic techniques, which can be characterized as examples of a *global frequency-domain energy* approach. Auditory-model algorithms for feature detection show noise insensitivity and amplitude independence, as well as

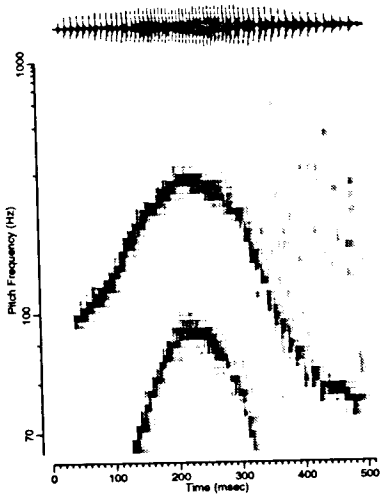


Figure 7. The response of the global pitch detector (lower plot) to an utterance /a/ spoken with rapidly increasing and then decreasing pitch (upper trace). The pitch plot shows bands at the fundamental pitch frequency, Fo, and at the first sub-harmonic, Fo/2.

selectivity for speech-like sounds. There are no inherent periodicity requirements for the stimulus, nor need the data be "framed" into arbitrary time segments as, for example, it must be prior to performing spectral analysis by Fourier transform or LPC coefficient extraction.

Comparison with other auditory model approaches

Several studies have used concepts of auditory physiology to motivate the design of algorithms for speech recognition. These approaches included the ensemble-interval histogram (EIH) method of Ghitza[5], the generalized synchrony detector (GSD) approach of Seneff[6] and the correlogram approach of Lyon[7]. All these methods are based on determination of the times of neural firings of a number of channels of a nonlinear auditory model. The response of each model fiber is then analyzed *individually*, for example by computing a period histogram of a fiber's response and then performing spectral or autocorrelation analysis of response. Global operations are then performed on the summed data from a number of individual channels to detect important features such formants. In addition to the drawbacks of temporal granularity and computational intractability discussed previously, the operations of accumulating histograms and performing spectral analysis are not likely to be physiological.

REFERENCES

[1] Deller, J.R. and Proakis, J.G. and Hansen, J.H.L. (1993). *Discrete-time processing of speech signals*, New York: Macmillan.

[2] Holton, T., Love, S.D. and Gill, S.P. (1991). "A fundamental approach to automatic speech recognition using models of auditory signal processing", *DARPA Technical Report DAAH01-91-C-R095.*

[3] Holton, T., Love, S.D. and Gill, S.P. (1994). "Robust pitch and voicing detection using a model of auditory signal processing", *Proc. ICSLP-94.*

[4] Licklider, J. (1951). "A duplex theory of pitch perception", *Experimentia*, vol. 7, pp. 128-131.

[5] Ghitza, O. (1986). "Auditory nerve representation as a front-end for speech recognition in a noisy environment", *Comput. Speech Lang.*, vol 1, pp. 109-130.

[6] Seneff, S. (1988). " A joint synchrony/mean-rate model of auditory speech processing", *J. Phonetics*, vol. 16, 55-76.

[7] Lyon, R.F. (1984). "Computational models of neural auditory processing", *Proc. IEEE ICASP*, pp. 1975-1978.

CORTICAL REPRESENTATION OF THE ACOUSTIC SPECTRUM

Shihab A. Shamma
Electrical Engineering Department and Institute for Systems Research
University of Maryland, College Park, MD 20742 USA

ABSTRACT

Acoustic signals are characterized by their timbre, pitch, loudness, forms of modulation, and onset/offset instants. These descriptions of sound quality have a close relationship to the instantaneous spectral properties of the sound waves. The auditory system has developed elegant mechanisms to extract and represent this spectro-temporal information through noise-robust perceptual features. At the level of the auditory cortex, these processes are manifested by an elaborate multidimensional representation of the shape of the dynamic acoustic spectrum. Specifically, at each frequency, the local shape of the spectrum is decomposed in terms of its bandwidth and asymmetry. Such a representation turns out roughly to correspond to a *local* cepstral-like representation of the spectrum, or more accurately, a wavelet transform of the acoustic spectral profile. Mathematical descriptions of this representation have become feasible and functionally relevant, and can be fruitfully used to derive the principles underlying time-frequency analysis in the auditory system. In turn, these principles can be applied in various contexts involving detection, analysis, synthesis, and recognition of sound.

INTRODUCTION

The spectral profile and its evolution in time play a key role in the perception of timbre of broad band sounds such as speech and music [1]. It is therefore important to understand how and which features of a spectral profile are extracted and encoded by the central auditory system. In this paper, we review first the fundamental response properties of neurons in the primary auditory cortex (AI), the last processing stage along the primary auditory pathway. Next, we discuss the implications of these findings to the representation of stationary and dynamic speech spectra such as those of a sustained vowel and the transitions in a CV syllable. Specifically, we shall demonstrate that the shape of the acoustic spectrum is represented along at least three different axes: the usual frequency axis, a local bandwidth (or scale) axis, and a local asymmetry axis. For dynamic spectra, the latter two axes additionally represent the speed and direction of formant transitions.

AI is strictly tonotopically organized because of the topographic order of neural projections from the cochlea through several stages of processing (Fig.1). Thus, when tested with single tones, AI neurons are selective to a range of frequencies around a best frequency (BF) [2]. Within this range, responses change from excitatory to inhibitory in a pattern that varies from one cell to another in its width and asymmetry around the BF (Fig.2); This response pattern is usually called the *response area or field* (RF) of the neuron [3]. When a broad band spectrum is used as a stimulus, the cell's response can be thought of as the net effect of all excitatory and inhibitory influences induced by the spectral region within its RF. However, despite the diversity of RF shapes and the complexity of their responses, two simple organizational principles underlie the way in which AI responses encode the shape of the acoustic spectrum. These are *linearity* and *selectivity* of AI responses.

LINEARITY OF AI RESPONSES

To first order, AI responses to broad band spectra are linear in the sense that they satisfy the *superposition* principle [4]. This is illustrated in Fig.3 as follows: Given the response patterns R_A and R_B evoked along the tonotopic axis by each of the stimulus spectra S_A and S_B, then the response pattern due to the sum of the two spectral profiles, $S_A + S_B$, is, to within a gain factor, the sum of the responses, i.e., $R_A + R_B$. This rather surprising finding is demonstrated experimentally in Fig.4, where single unit responses to different

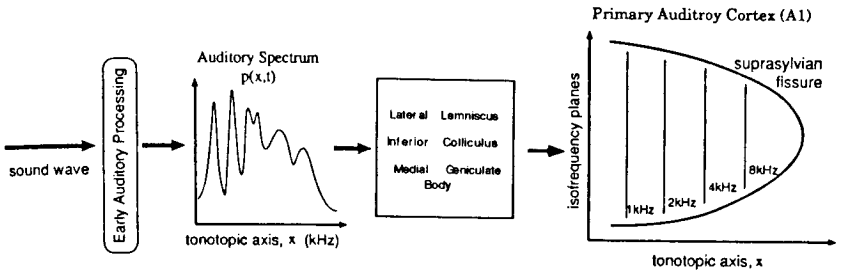


Figure 1. Schematic of the auditory pathway to the primary auditory cortex (AI) of the ferret. AI is tonotopically organizedwith units of similar BFs forming isofrequency planes as indicated.

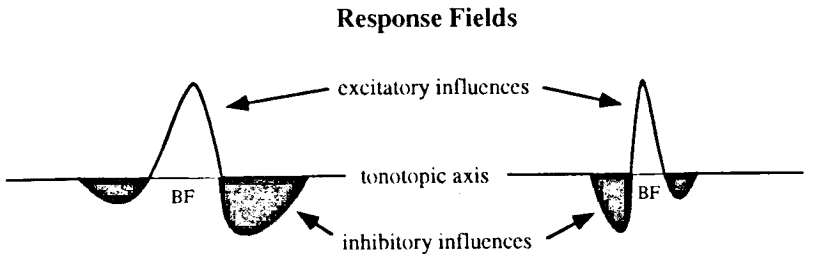


Figure 2. Schematics of two different tesponse fields (RF) measured in AI units. RFs may have different asymmetries with inhibition more prominant above the BF (left), below the BF (right), or simply symmetric. They also range in bandwidths from broad (left) to narrow (right).

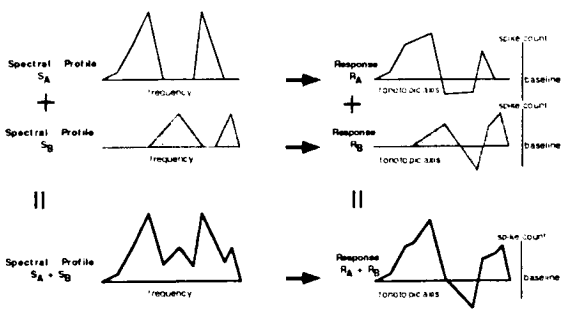


Figure 3. Linearity of AI responses imply the superposition principle. Acoustic spectral profiles S_A, S_B, and $S_A + S_B$ evoke schematic response patterns R_A, R_B, and R_A+R_B along the tonotopic axis. The response patterns are measured as spike counts relative to a "baseline" which is the response to a flat spectrum.

sinusoidal spectral profiles (also called *ripples*) are combined and compared with responses to superimposed ripple spectra. For instance, Fig.4A displays the responses to four spectra with different ripple densities (0.4, 0.8, 1.2, and 1.6 cycles/octave). In each case, the *dashed* response curve is constructed from the spike counts of the cell as the spectral profile is shifted to the left relative to the BF. As is evident, the responses track the sinusoidal shape of the input spectrum; They are largest when the ripple rate is 0.8 cycle/octave, and weakest at 1.6 cycles/octave. The *solid* curves are the best mean-square fits to the data.

When a complex spectrum is formed by the superposition of two ripples (Fig.4B), e.g., 0.4+0.8 cycles/octave (top) or 0.8+1.6 cycles/octave (bottom), linearity predicts that the response curves should resemble the superposition of the responses to the individual ripple spectra. This is confirmed by the similarity of the measured (*dashed*) and predicted (*solid*)) response curves in both cases. These results have been confirmed in a large number of tests involving spectral profiles composed of up to 10 superimposed spectra [5,6].

Linearity is a powerful simplifying principle that allows one to predict the responses to any arbitrary spectral profile. Specifically, if the responses to the basic set of rippled spectra are known, then it is possible to superimpose them uniquely to generate the responses to any arbitrary profile (this is the so-called *Fourier* decomposition) [4]. Therefore in the remainder of this paper, we shall examine in more detail the response properties of AI cells to various spectral ripple parameters.

SELECTIVITY OF AI RESPONSES

AI units are generally selective in that they respond only within a limited range of values of a given stimulus parameter. For instance, units are usually tuned along the tonotopic axis, i.e., they are driven by a relatively narrow range of frequencies around a BF as described earlier (Fig.2). AI responses are also selective to the parameters of a ripple spectrum, specifically the ripple frequency (or density, Ω) and ripple phase (Φ). For instance, in Fig.4A, the unit responds best around the ripple frequency 0.8 cycles/octave. Furthermore, the responses vary with the phase of the ripple, being excited in one-half cycle while suppressed in the other. This ripple selectivity can be efficiently displayed by a *transfer function* $T(\Omega)$ (Fig.5), where the amplitude and phase of the responses to different ripples are plotted as a function of ripple frequency. A complementary view of this information is contained in the unit RF which is (conceptually) formed by summing up the responses to the different ripples, or more accurately by *inverse Fourier* transforming $T(\Omega)$ [4].

Selectivity of an AI unit around a characteristic ripple frequency (Ωo) is intuitively inversely related to the width of its RF, or roughly to the bandwidth of the unit's frequency tuning curve [5]. Thus, the higher Ωo is, the more narrowly tuned the RF is. This suggests that a unit responds best (or is selective) to spectral patterns with a local bandwidth (or *scale*) that is comparable to that of its RF.

Similarly, selectivity to a particular ripple phase (characteristic phase, Φ_O) is directly reflected in the asymmetry of the RF. For a unit with Φ_O near zero, the RF exhibits a central excitatory region around the BF, flanked by symmetric inhibitory areas. If the Φ_O is positive (negative), the inhibition becomes asymmetrically strong below (above) the BF [3,5,7]. In this manner, an AI unit is selective to the local slope or asymmetry of the input spectral profile around the BF.

AI REPRESENTATION OF A VOWEL SPECTRUM

The combined selectivity of an AI unit to the asymmetry and scale around a local spectral region (BF) of the input profile means that it can encode explicitly the local shape of the spectrum. For example, the asymmetry of the RF in Fig.5 is directly responsible for the unit's selective responses (Fig.6) to the 2nd formant of the vowel spectrum /aa/, and not to the 1st formant. By having RFs with a range of BFs, bandwidths, and asymmetries, the AI can represent the shape of the entire input spectrum along three different axes. Such a representation is demonstrated in Fig.7 for the spectral profile of the vowel /aa/

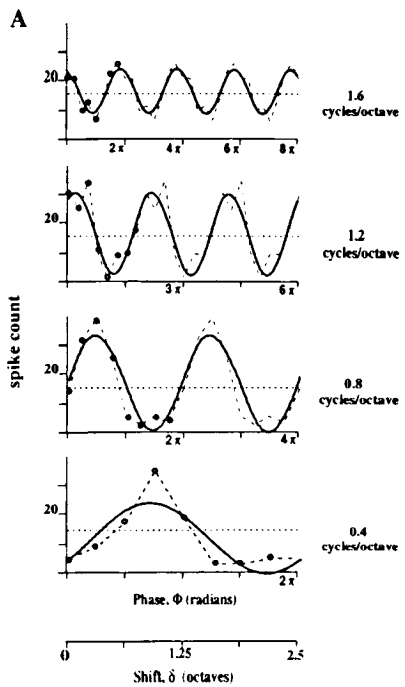


Figure 4. Superposition of responses to rippled spectra. (A) Responses of a unit to four input profiles at different ripple frequencies. In each case, the spectrum was shifted downwards relative to the BF of the unit to obtain the responses over the full cycle. (B) Comparison of recorded and predicted responses to spectra composed of ripple combinations 0.4+0.8 (top) and 0.8+1.6 (bottom) cycles/octave.

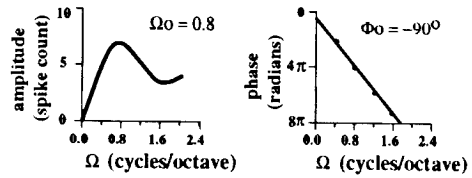
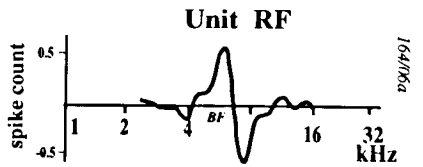


Figure 5. Ripple transfer function ($T(\Omega)$) of an AI unit. . The magnitude (left) and phase (right) of $T(\Omega)$ can be inverse Fourier transformed to obtain the RF (bottom).



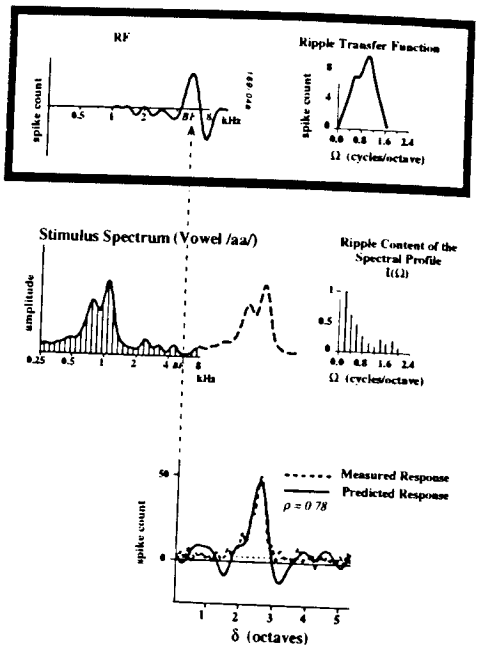


Figure 6. Measured and predicted single unit responses to the spectrum of a vowel /aa/. The unit's RF and T(Ω) are shown in the top box. The vowel spectrum (middle left) is synthesized by adding 10 ripples with amplitudes indicated by I(Ω) (middle right). Note, the unit responds well to the 2nd formant of the vowel (bottom); It does not respond to the 1st formant because of its asymmetric RF.

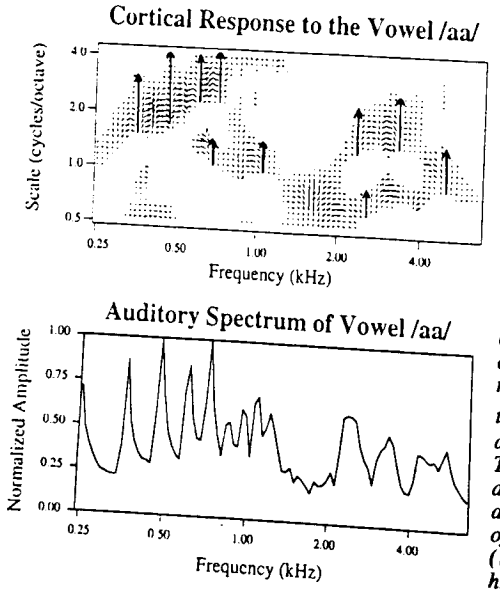


Figure 7. The spectrum of a naturally spoken vowel /aa/ (bottom), and the corresponding cortical representation (top). The scale axis is labeled by the characteristic ripple Ωo of the cortical cells. The characteristic phase is represented by the direction of the arrows (0 to 2π in clockwise direction, starting at up-arrow). The strength of the response is denoted by the length of the arrows. For clarity, trajectory of activated symmetric RFs (up-arrows) have been highlighted.

[7]. It is evident that the shape of the profile is dominated by the spectral peaks, i.e., the overall formant structure and the underlying harmonicity in the low frequency region (usually <1 kHz). These features are explicitly analyzed in the cortical representation in terms of their local asymmetry and bandwidth.

For example, the fine structure of the spectral harmonics is visible at the higher scales (usually > 1.5-2 cycles/octave). In contrast, the formants are relatively broad in bandwidth and thus are represented by the activity of units tuned to lower Ωo (< 2 cycles/octave). Sometimes, closely spaced formants are represented simultaneously at multiple scales as in the region of the 3rd formant (around 2.5 kHz), which is represented by activity near Ωo = 0.5 and 2 cycles/octave. The higher Ωo corresponds to the 3rd formant peak (approximately 0.25 octave in width). The lower Ωo captures the broad and skewed distribution of energy due to the combined 3rd and 4th formant peaks. A similar ``double-scale" representation occurs near 600-700 Hz, where the fine harmonic structure is represented at the higher scales, while the format structure (evident in the envelope of the harmonic peaks) is captured at lower Ωo.

The local asymmetry of the pattern in this representation is encoded by the direction of the arrow of the response. It provides a description of the local energy distribution in the spectrum. For example, the tonotopic locations at which the spectrum is locally symmetric (and hence represented by the up-arrows) closely reflect the positions of the peaks in the auditory spectrum; The left and right-arrows indicate whether the nearest spectral peak is at a higher or lower frequency. For instance, the spectral peak of vowel /aa/ at 3.25 kHz is not resolved at the broad scale, i.e., there is no up-arrow at Ωo =1 at this frequency. Instead, it is regarded as a trough (down-arrows) because it is flanked by two stronger peaks. However, the peak and its surrounding narrow valleys are resolved at a higher scale corresponding to twice the Ωo (around 2 cycles/octave).

AI RESPONSES TO DYNAMIC SPECTRA

Remarkably, AI units exhibit the same response properties of linearity and selectivity to *dynamic* spectral profiles. Thus, responses evoked by any combination of dynamic inputs can be roughly predicted from a linear sum of responses to the individual inputs. This is demonstrated in Fig.8 using a rippled spectrum (Ω = 0.8 cycles/octave) that moves to the left along the tonotopic axis with different *angular* velocities ω (4-24 cycles/sec). As is evident in Figs.8A and B, these stimuli evoke in this unit well synchronized responses to all ripple velocities. Combining ripples with different ω and Ω (Fig.8C) produce responses that are predictable by superposition of responses to the individual moving ripples. Again, the significance of linearity and of the basic set of moving ripple stimuli is seen through the *Fourier* decomposition theorem which allows us to generate and predict the responses to arbitrarily complex dynamic spectra, such as those of speech CV-syllables.

As with stationary ripples, responses to moving ripples are selective in that a given unit responds over a restricted range of velocities around a characteristic rate, ω_o [8,9]. Furthermore, there does not seem to be a relationship between ω_o and Ωo in a given unit, i.e., in a large population of AI units in the ferret, all combinations approximately within Ωo < 2 cycles/octave and ω_o < 20 cycles/sec may occur. It is likely that these ranges vary significantly across species reflecting their acoustic environment.

One possible implication of the selectivity to ω is the ability to encode the *rate* of spectral transitions. In addition, AI units are readily selective to the direction of a spectral transition by virtue of their RF asymmetries [7]. Combining those two features, together with those of bandwidth and BF creates a multidimensional cortical representation which explicitly extracts and maps out a variety of stationary and dynamic measures of the shape of the acoustic spectrum [7].

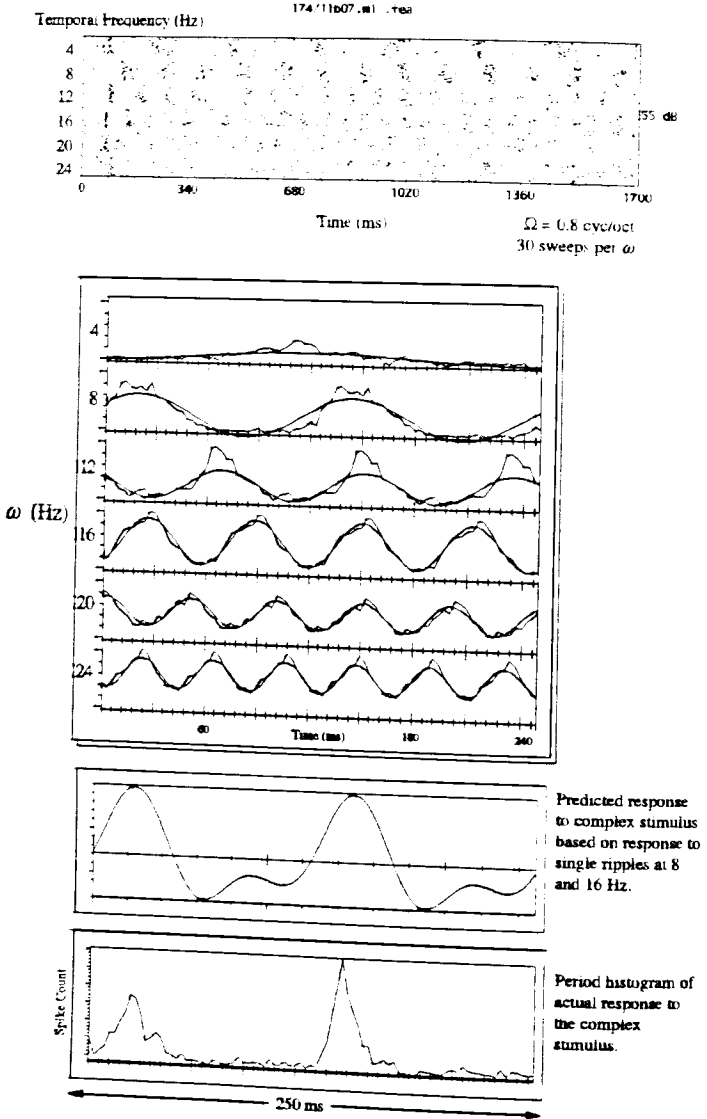


Figure 8. Responses to rippled spectra. (A) A 0.8 cycle/octave ripple moving at increasing angular velocities ω (4-24 cycles/sec) evokes synchronized responses. The ripple begins to move at time 0 ms, and is turned on at time 50 ms. Following an onset response, the unit synchronizes to the stimulus periodicity as seen in the period histograms below. (B) Period histograms of the responses and their mean-square error sinusoidal fits. (C) Measured (bottom) and predicted (top) responses for a two 0.8 cycle/oct ripples travelling at different velocities ω (8 and 16 cycles/sec). The top plot is constructed from the sum of the sinusoidal fits at 8 and 16 Hz shown in (B) above. The bottom plot is a 250 ms period histogram of the responses.

REPRESENTATION OF PITCH

So far, we have focused exclusively on the representation of spectral profiles. Another important percept that can be integrated in this representation in various ways is that of pitch. It is well established that the resolved harmonics of voiced sounds (such as those of the vowel /aa/ in Fig.7) contribute significantly to the perception of pitch [1]. On the logarithmic frequency tonotopic axis of the auditory system, resolved harmonics fit within a typical pattern which, except for a shift along this axis, remains unchanged regardless of pitch value. The harmonics in turn create a similarly stable cortical activation pattern at the higher scales from which pitch values and strength can be determined [7]. Other temporal mechanisms for the encoding of pitch can also be theoretically integrated in the cortical representation if the appropriate ranges of ω temporal selectivities are used, e.g., in the manner already suggested by [10] at lower auditory centers.

DISCUSSION

The auditory cortical representation of the acoustic spectrum is evidently quite redundant in that it expands the profile along several additional axes (e.g., ripple scale and phase, and temporal rate). This redundancy potentially serves many important functions. One is making explicit the spectral features responsible for the recognition of different phonemes, the evaluation of pitch, the perception of voice quality, and other auditory perceptual tasks. Another function is endowing the spectral representation with added stability and noise-robustness [7].

Another interesting area of speculation concerns the question of how the cortical representation can be gracefully mapped unto vocal tract parameters or models. This is an important issue both from a biological and an applications points of view since vocal tract models are heavily utilized in systems for data compression, vocoders, synthesizers, and speech recognizes.

ACKNOWLEDGMENT

Many colleagues participated in the experiments and analyses presented in this paper. They include Huib Versnel, Didier Depireux, Nina Kowalski, Kuansan Wang, Po-wen Ru, Tony Owens, and Preetham Gopalaswamy. This work is supported by the Air Force Office of Scientific Research, The Office of Naval Research, the National Science Foundation through NSF Grant (# NSFD CD 8803012), and the National Institutes of Health.

REFERENCES

[1] Plomp, R. (1976), *Aspects of Tone Sensation*, Academic Press

[2] Merzenich, M., Knight, P., Roth, G. (1975), "Representation of cochlea within primary auditory cortex", *J. Neurophysiology*, vol 28, pp.231-249.

[3] Shamma, S. , Fleshman, J., Wiser, P., Versnel, H. (1993) "Organization of response areas in ferret primary auditory cortex", *J. Neurophysiology*, vol.69(2), pp.367-383.

[4] Oppenheim, A., Willsky, A., and Young, I. (1983), *Signals and Systtems*, Prentice Hall.New Jersey.

[5] Shamma, S., Versnel, H., and Kowalski, N. (1995) "Ripple analysis in ferret primary auditory cortex. I. Response characteristics of single units to sinusoidally rippled spectra", *Auditory Neuroscience*, vol.1(2) *(in press)*.

[6] Shamma, S., and Versnel, H. (1995) "Ripple analysis in ferret primary auditory cortex. II. Prediction of single unit responses to arbitrary spectral profiles", *Auditory Neuroscience* , vol.1(2) *(in press)*.

[7] Wang, K., and Shamma, S. (1995) "Spectral shape analysis in the primary auditory cortex", IEEE *Trans. Speech and Audio (in press)*.

[8] Schreiner, C. and Urbas, J. (1988) "Representation of amplitude modulation in the auditory cortex of the cat. II. Comparison between cortical fields", *Hearing Res.*, vol.32, pp.49-64.

[9] Eggermont J. and Smith G. (1995) "Synchrony between single unit activity and local fiel potentials in relation to periodicity coding in primary auditory cortex", *J. Neurophysiology* , vol.73(1), pp.227-245.

[10] Schreiner, C., and Langner, G., (1988) "Periodicity coding in the inferior colliculuc of the cat", *J. Neurophysiology*, vol.60, pp.1823-1840.

CONSTRAINT-BASED APPROACHES TO PHONOLOGY

G. N. Clements, CNRS, UA 1027, Paris

ABSTRACT

Current research in phonology has placed increasing emphasis on the importance of constraints and their interactions in phonological systems, while decreasing or eliminating the role of generative rewrite rules. The present paper offers a brief review of constraint-based approaches to phonology, considering some of their advantages over traditional models.

1. THE EMERGENCE OF CONSTRAINTS IN RECENT PHONOLOGICAL THEORY

One of the fundamental hypotheses of generative phonology since its inception in the early 1960s has been that the phonological component of a grammar consists of a set of rewrite rules that apply in sequential order to generate surface forms from underlying representations. One of its main insights has been that regular alternations in the phonological shape of morphemes could be captured by assigning each such morpheme a single underlying representation, and generating its alternants by rules which often prove to be of considerable generality. A strong constraint on rules is that they cannot access any information other than that present in the input string. Thus, in particular, they are "blind" to the effects they produce in their individual and collective output.

While this view of the organization of a phonological system is the one that continues to be presented in textbooks, it has been undermined in recent years by the increasing role played by constraints as a central feature of phonological explanation. By "constraint" I mean any statement, universal or language-particular, which has the effect of defining the set of lawful phonological representations without directly specifying a change in structure. In various guises—structure conditions, phonotactics, filters, well-formedness conditions, etc.—constraints began to appear in the literature on a sporadic basis in the 1970s, at the margins of otherwise quite orthodox analyses. Toward the beginning of the 1980s, however, some researchers began to believe that constraints play a more central explanatory role in phonology than had previously been thought. Since that time, the notion of constraint has gathered considerable momentum, and today seems in a position to replace the notion of rewrite rule altogether.

This evolution in thinking has had a variety of causes. For one thing, a similar evolution had taken place in syntactic theory, where transformational rules have come to be largely eliminated in favor of a variety of types of constraints on representations; the successful elimination of derivational, rule-based approaches in syntax has no doubt inspired linguists to explore similar approaches to phonology.

However, there are other reasons for the emergence of constraints, having to do with the particular nature of phonological data. For one, many linguists have observed that phonological rules do not apply in a perfectly arbitrary fashion, but tend to favor certain types of outputs. For example, rules of epenthesis and deletion may apply in such a way as to produce open syllables, or clusters no longer than two consonants, depending on the language [1]. In tone languages, rules tend to assign tones to toneless syllables, and to disprefer contour tones [2]. The rules of stress systems apply in such a way as to create preferred types of stress patterns, avoiding adjacent stresses and favoring alternating stress, and placing main stresses at the extremities of words [3]. Segmental rule systems tend to avoid or eliminate adjacent identical segments [4]. The apparently goal-oriented character of such subsystems cannot be readily reconciled with the output-blind nature of rewrite rules. A further observation, which stimulated much discussion in the 1970s but no widely-agreed upon solutions [5], was that the effect of phonological rules is often replicated by constraints holding over phoneme sequences within morphemes. For example, languages which assimilate obstruents to the voicing of a following obstruent across a morpheme boundary usually require all members of an obstruent cluster to agree in voicing within a morpheme. This duplication of the effects of morpheme structure conditions and rewrite rules is purely accidental in the standard SPE framework.

However, perhaps the single most important factor leading to the emergence of constraints has been the development of nonlinear phonology in its various forms—autosegmental, metrical, syllabic, prosodic, and so forth. What these frameworks have in common is the complexity of their representational systems compared to the simple, linear representations of standard generative phonology. Given sufficiently rich representations, many properties of surface representations that had formerly been accounted for as the effect of ordered rules can be shown to follow from purely structural features of representations. To take a simple example, the recognition of the syllable as a phonological unit allows a significant reduction in the amount of rules needed to account for alternations that are (from our current standpoint) best viewed as syllable-conditioned [6]. Perhaps most significantly, the increasing richness of representational systems imposes a new need for severely constraining the ways the various parts of a representation can fit together. In autosegmental phonology, for example, it has proven desirable to eliminate certain types of cross-tier association patterns (notably, those in which assocation lines cross) in terms of a universal Well-formedness Condition, which functions both to eliminate ill-formed underlying representations and to "police" the operation of rules so that violations are not produced in rule outputs [2, 7]]. In metrical phonology, it has been found that stress systems obey rather strict constraints that do not follow directly from properties of metrical representations themselves, and much work has been directed toward the goal of constraining the theory by proposing a small number of representational parameters along which only a reduced number of choices are available [8, 9]. In syllable theory, an important set of constraints on syllable types can be stated in terms of the Sonority Sequencing Generalization, originally proposed in the 19th century and rediscovered in the context of the recent renaissance of syllable theory (see [10] for a review).

Alongside general system constraints of these types, phonologists have recognized more parochial constraints, specific to certain languages, that further restrict the variety of representational structures available to a language.

The notion of constraint is not unique to current linguistic frameworks. In pre-generative theory, constraints often played an important role in phonological description in the guise of "laws of euphony", "phonotactics", and other types of statements which specified what phoneme combinations could and could not occur in phonemic representations. Some such statements were framed in terms of a hierarchy of constituents in the modern sense; thus, Hockett [11] proposed that all languages contain sequences of syllables, and that syllables consist of ordered sequences of smaller constituents such as onset, peaks, and codas, etc. In his view, the specification of sequential constraints on phonemes in a language involves, in part, a specification of which phonemes may occur in which type of syllabic constituent.

What distinguishes current constraint-based frameworks from earlier work of this type is its retention of the generativist goal of accounting not only for static phoneme distributions, but also for phonologically-conditioned morpheme alternations. Thus, to take an example, we not only need to account for the fact that a language like LuGanda does not allow adjacent vowels in its morphemes and words (*ai, *iu, etc.), we also want to account for the related generalization that when two vowels abut as a result of morpheme combination, the first one is eliminated via glide formation if it is high (1a), and via deletion if it is low (1b).

(1) a. /li+ato/ lyaato 'boat'
/mu+iko/ mwiiko 'trowel'
b. /ma+ato/ maato 'boat' (dim.)
/ka+ezi/ keezi 'moon' (dim.)

The resulting vowel is long. Note that the prefix vowels are retained before consonant-initial stems such as /-mpi/ 'short': cf. [li-mpi], [mu-mpi], [ma-mpi], and [ka-mpi]. (Also, all vowels are lengthened before NC clusters by a subsequent rule, whose effect is not shown here; see [12] for fuller discussion.) To account for the surface forms in (1), it is not enough simply to

state the constraint against vowel sequences; we must also provide specific principles stating how violations of the constraint are lawfully resolved. This is not a straightforward matter, as we can see by considering the various ways that an anti-hiatus constraint can be resolved in principle: (i) by deleting the first vowel, (ii) by deleting the second vowel, (iii) by gliding the first high vowel, (iv) by gliding the second high vowel, (v) by assimilating one vowel to the other, (vi) by fusing the two vowels into a different one (coalescence), (vii) by epenthesizing a consonant between them, etc. Early attempts to incorporate constraints into phonological descriptions often neglected this problem, and so failed to provide satisfactory solutions to the treatment of alternations. Many of the specific features of current constraint-based frameworks can be understood in terms of the need to resolve the problem of alternations in a principled way.

2. SOME CURRENT CONSTRAINT-BASED APPROACHES

Most current constraint-based theories maintain a double commitment to the goals of accounting for static regularities of distribution and for genuine phonologically-conditioned morpheme alternations. Other than this common core, they differ in often substantial ways. One can currently count nearly a score of well-defined and distinguishable constraint-based theories. Here we will briefly review three frameworks that have received particular attention: constraint- and repair theory, declarative theory, and optimality theory. General overviews of these theories, containing illuminating comparisons among these (and other) approaches and further references, are given in [13, 14, 15]. Our discussion must necessarily be cursory, and we refer the reader to the fuller presentations available in these sources.

One useful basis of comparison is that between theories which are based on inviolable constraints, and those which allow constraints to be violated. In the earliest discussions, constraints were usually considered inviolable, the principal argument for this view being that the use of violable constraints would greatly weaken the predictive power of the model, particularly when used in conjunction with (violable) rewrite rules. However, some subsequent work has relaxed this condition, entertaining constraint violations on either a temporary or permanent basis.

We will first consider two constraint-based approaches in which constraints cannot be violated at the surface level. They differ in that the first allows constraint violations in the course of derivations, but not in surface representations, while the second, a non-derivational framework, allows no violations in any representations. We then consider a third (also nonderivational) approach which allows constraint violations in surface representations.

2.1. CONSTRAINT-AND-REPAIR APPROACHES

The family of constraint-and-repair approaches was one of the first developments of standard generative phonology in which constraints on representations have a well-defined (and in some versions, exclusive) role in monitoring derivations [16, 17, 18, 19, 20, 21, 22]. To see the relation between constraints and rules, it may be helpful to consider the logical structure of a standard rewrite rule making use of the format A → B / C__D. In rules of this type, the structural description is defined as the input string CAD, and the structural change as the output string CBD. Note that the structural description of an obligatory rule consists, in effect, of a description of a sequence which is ill-formed at the point in the derivation at which the rule applies, while the expression "A→ B" specifies the way in which this violation is eliminated. In other words, a rewrite rule pairs an input constraint with an operation which has the effect of producing a locally well-formed output. Once we perform this disassociation, we find that an SPE-type rule can be factored into what can be viewed as a local constraint and a local repair operation.

The particular insight of constraint- and repair theories is not, then, to introduce the notions of constraint and repair as such, but to *delink* the connection between these notions which had been inseparably paired up to that time. Once delinked from a specific repair, a constraint can operate pervasively, defining ill-formed sequences both in underlying representations and at subsequent levels, where such sequences may result from morpheme concatenation and from the operation of "output-blind" rules.

Moreover—and here is a crucial advantage—more than one way of repairing a given ill-formed sequence can be specified. Let us consider again the LuGanda forms presented in (1). In a traditional rewrite rule framework, the surface forms can be accounted for in terms of two rules, one turning an initial high vowel into a glide, and the other deleting an initial non-high vowel. In a constraint-and repair framework, these two rules can be replaced by a single constraint prohibiting vowel sequences—let us call it the *VV constraint— and two repair operations. The constraint both accounts for the absence of vowel sequences in the underlying representation of morphemes, and serves to trigger appropriate repair operations when vowel sequences are created by morpheme concatenation. The repair operations required in this case are [V, +high]→ G and [V, -high]→ Ø. (Compensatory lengthening of the second vowel must be assured by independent means.) As "repair strategies", these operations are kept in reserve, applying only when they are needed to eliminate constraint violations.

The *VV constraint, once extracted from conventional rule statements, can be recognized as expressing the familiar cross-linguistic dispreference for vowels standing in hiatus. We can consider it a member of the set of universal principles defining preferred or unmarked representations, one which is invoked in the grammar of LuGanda and in many other (but not all) languages. The repairs themselves can be assigned to a small pool of universal elementary operations, including linking, delinking, and deletion.

This treatment has several clear advantages over a standard rewrite-rule approach. First, it extracts a single anti-hiatus constraint from a set of rules which was forced, in the standard theory, to state it twice. Second, it accounts for underlying constraints, surface regularities and alternations by the same set of principles. Third, it reinterprets the structural description and structural change of two arbitrary rules in terms of a set of phonetically plausible universal constraints and repair operations.

There are a number of fairly obvious questions that a constraint-and-repair approach must address if it is to be internally coherent and descriptively adequate. First, it is commonly assumed that constraints fall into two types: those that have a blocking function, preventing rules from applying if their outputs would violate the constraint, and those that do not block rules, but rather trigger repairs *after* the rule has applied. Constraint-and-repair theory must provide a principled way of predicting which constraints are of which type, unless we are willing to allow each constraint to be annotated for this information on a case-by-case basis. Second, given that repairs are formally dissociated from constraints, it is no longer a straightforward matter to determine how a given constraint violation will be repaired. Two or more repairs may applicable to a given constraint violation, and if repairs are not extrinsically ordered as most current work assumes, then principles must be offered that will predict which of a set of competing repair operations will apply in any given situation. These questions can be subsumed under the general observation that constraint-and-repair theory, as a derivational approach, must provide a sufficient core of system-level principles to administer the rich sets of interactions predicted by its logical structure (see [23] for relevant recent discussion).

2.2. DECLARATIVE APPROACHES

A second family of constraint-based approaches is founded on the principle that phonological grammars consist exclusively of a pool of unordered constraints or well-formedness conditions which, taken together, associate each lexical entry with a well-formed surface representation. Such constraints are "declarative" in the sense that they specify conditions that must be satisfied by surface representations, rather than operations or procedures that must be applied to derive one from the other, as in standard generative phonology. Taken collectively, the constraints are generative in the sense that they completely specify

the surface form of each lexical entry, including phonologically conditioned morpheme alternations. Approaches of this type include Categorial Phonology, an extension of categorial grammar to the phonological level [24, 25], and Declarative Phonology, which similarly extends unification-based grammar to phonology [26, 27, 28].

Declarative approaches do not employ rules or other types of procedural statements, and do not impose extrinsic ordering on their constraints. An important consequence of these properties is that such approaches are necessarily non-derivational, in the sense that they associate full representations to lexical entries without passing through a series of derivational steps. Another consequence is that structure-changing operations are prohibited, including deletion; lexical entries must be properly or entirely contained in their surface representations (monotonicity). In contrast to constraint-and-repair approaches, the constraints of this family of theories are absolutely inviolable; this means that they must be formulated with enough precision to assure that only one of two or more potentially conflicting constraints can be satisfied by any given surface form.

To continue discussion of the LuGanda example, since constraints are inviolable in surface representations, a declarative analysis cannot allow any surface violations of the *VV constraint. But in the absence of deletion rules, how can we relate a surface form like [keezi] to its underlying representation / ka+ezi /, containing two vowels? In the case of alternating segments like the prefix vowel, declarative approaches typically underspecify, or declare as optional, any information that does not appear in all alternants. For example, since [a] does not appear in all the alternants of / ka-/, we may parenthesize it in lexical entries, as follows: /k(a)/. The parentheses indicate that the vowel is present only if it is not excluded by the constraint system. The *VV constraint requires the parenthesized vowel to be absent in /k(a)+ezi/, but correctly does not exclude this vowel in /k(a)+mpi/ 'short', where it is retained in the surface form [kampi]. The analysis of forms like [lyaato], in which the prefix vowel is realized as a glide, proceeds in principle along similar lines, which we will not attempt to work out here. (Again, compensatory lengthening of the second vowel must be assured by independent constraints.)

It will be noted that unlike the constraint-and-repair approach, the declarative account of LuGanda is non-derivational, in the sense that the surface form is not built up, step by step, by applying a series of rules or repair operations. Rather, the constraint system defines the full set of surface alternants that corresponds to each lexical representation.

This brief discussion, though incomplete, is sufficient to show that some of the problems that potentially face constraint-and-repair approaches do not appear in declarative approaches. Since declarative approaches do not make use of rules and repairs, and do not admit constraint violations, the problem of predicting which constraints have a blocking and which a repair-triggering function, or of determining which of several applicable repairs takes precedence in a given constraint violation, simply does not arise. On the other hand, several new questions must be addressed.

For example, declarative approaches resemble the traditional morpheme-alternant models of pre-generative linguistic theory in certain potentially problematical respects. Such theories do not derive the alternants of a morpheme from a single base form, but instead state distributional rules which predict which member of the set will be selected in any given context. There are well-known analytical problems confronting such theories, which have been discussed, for instance, by Kenstowicz and Kisseberth [30, pp. 180-96], and these must be resolved if declarative approaches are to capture the same range of linguistic generalizations that traditional rule-based (and constraint-and-repair) theories have succeeded in accounting for.

A further potential problem concerns the formulation of constraints. In case a lexical entry may potentially satisfy several conflicting constraints, principles must be provided to determine which takes precedence. One solution [31] is to require that constraints be stated in sufficient detail that no two constraints will ever compete for the same form, except for the special case in which their interaction can be predicted by the Kiparsky's Elsewhere Condition [32]. However, if constraints are stated in enough detail to eliminate conflicts in a given grammar, they quickly become complex, highly language-particular, and phonetically arbitrary. A result is that constraints in declarative systems cannot in general be related to universal constraints in any straightforward way, lending themselves subject to much the same sort of objections that were earlier raised against the arbitrariness of SPE-type rewrite rules.

2.3. OPTIMALITY THEORY

The leading idea of Optimality Theory as proposed by Prince, Smolensky, and McCarthy [33, 34] is that Universal Grammar consists in part of a set of constraints on representational well-formedness which are contained in all grammars. These constraints are highly conflicting and make sharply contradictory claims about the relative well-formedness of most representations. Unlike the approaches discussed up to now, the constraints posited by Optimality Theory are typically violated in the surface forms of any language. To resolve conflicting claims, each grammar ranks the constraints in a strict dominance hierarchy, such that each constraint has absolute priority over all those it dominates in the hierarchy. It is the relative ranking of the constraints on the hierarchy that determines which candidates, among possible alternatives, are selected as actual surface representations. The preferred candidate is the one that satisfies the conflicting constraint set not absolutely, but relatively better than all others. In other words, although all candidates will typically violate some constraints, the optimal (and hence selected) candidate is the one which violates the lowest-ranked constraints.

Optimality Theory provides two general mechanisms to implement this approach. One is a principle GEN which associates each unprosodified lexical entry with a typically infinite set of fully prosodified candidate output forms. This principle is subject to a principle of *containment* requiring that each lexical entry is properly contained in each output candidate; additional structure may also be postulated. A second mechanism is the principle EVAL, which selects the optimal candidate from among the set created by GEN. It proceeds by assessing the constraint violations presented by each candidate, eliminating candidates on a worst-first basis until only one is left.

Like the declarative approaches with which it shares a number of assumptions, Optimality Theory is a nonderivational theory of phonological form; it posits no rules or repairs that map one form into another in a step-by-step, deterministic fashion. In distinction to declarative theories, however, the optimal candidate is selected from the candidate pool created by GEN with no further reference to the structure of the original lexical entry; that is, even though some constraints are conditional in form, the precondition of any such constraint is not defined on the (lexical) input but upon the (surface) output. It is therefore only the principle of containment which links output forms to specific inputs.

Let us see how these principles might be applied to our LuGanda example /ka+ezi/. On the basis of the unprosodified lexical representation, GEN creates a set of candidate forms, of which we consider three for purposes of illustration: one which contains the violation of the hiatus constraint, and two which eliminate it. Notice that the form that contains the violation is not necessarily eliminated from consideration; it will in fact be selected as the optimal candidate if the other forms from the candidate pool violate higher-ranked constraints. Therefore, in order to insure that EVAL selects the correct output [keezi], we must determine how LuGanda ranks the members of the universal constraint set. Let us assume for purposes of illustration that this set contains, in addition to a *VV constraint, the following additional constraints:

FILL^X: every skeletal position must dominate segmental material

PARSE: every segment must be incorporated into syllable structure

FILL^X has the effect of ruling out epenthesis, viewed as introducing empty consonant positions into the CV- or mora-skeleton (such positions, if present in surface representations, are viewed as

filled at a different level of representation, perhaps in the phonetics). PARSE requires all segments to be syllabified; it is assumed that unsyllabified segments are not phonetically realized (but not deleted; it will be recalled that by the containment principle, no material can be deleted). In order to select the correct output, in which the prefix vowel is unparsed (in violation of PARSE), the constraints *VV and $FILL^X$ must outrank PARSE. These rankings are sufficient to select the correct output (c) over its two competitors in the tableau shown in (2), showing a selection of candidate surface representations for / ka+ezi /.

(2)

	*VV	$FILL^X$	PARSE
a. kaezi	*		
b. kaCezi		*	
→c. k<a>ezi			*

Asterisks in any row indicate constraint violations, brackets indicate an unparsed segment, and C represents an unfilled consonant position. Since PARSE is the lowest-ranked constraint, candidate (c) (the least bad) is elected, as shown by the arrow.

This simplified example is intended, as before, only to give an idea of the strategic approach of optimality theory, and any actual analysis will necessarily be more complex. In the present case, we have not discussed the treatment of glide formation, or of compensatory lengthening, which may require some further enrichment of the representational system of this framework [35]. Grammars of different languages are viewed as differing not in their selection of constraints (since all members of the constraint set are present in the grammars of all languages), but in terms of the rankings they impose on them. For example, a language that ranked PARSE above $FILL^X$ and *VV would reject output (c) and select (a) or (b), depending on which of the remaining constraints is the higher-ranked.

Optimality theory has attracted much attention and is still undergoing development. We can see that it addresses most of the potential problems raised in regard to preceding frameworks, of which it provides, to a certain extent, a synthesis. Prince and Smolensky have pointed out: "Among the principles of Universal Grammar [i.e., the universal constraint set—GNC] are cognates of those formerly thought to be no more than loose typological and markedness generalizations. Formally sharpened, these principles now provide the very material from which grammars are built" ([33], 219). While the incorporation of substantive constraints into the marrow of the grammar is a desirable goal, it is apparent that the set of universal constraints required to account for the full range of phonological diversity will prove to be quite large, and will necessarily contain a sizable number of arbitrary constraints having limited cross-linguistic generality. And it is difficult to see in what sense a proposed constraint such as *P/a ("[a] does not form a syllable peak"), essential to the Prince and Smolensky system, can be regarded as universal, in view of the fact that [a] is an optimal syllable peak in all known languages. It can be expected that such questions, and others, will be addressed as research proceeds.

3. CONCLUSION

It is now apparent that not only can constraint-based systems of phonology account for many (or most) of the phenomena that theories based on rewrite rules could account for, they can do so in many respects in a much more principled way. This fact by itself justifies the considerable attention being devoted to constraint-based phonologies at the present time. On the other hand, the diversity of current ideas suggests that theoretical models are still in evolution, and should be regarded as still tentative in many respects. For this very reason, however, added to its preliminary achievements, this direction of research must be regarded as an especially dynamic and highly promising one at the present time.

REFERENCES

[1] Kisseberth, C.W. (1970) "On the Functional Unity of Phonological Rules," *LI* 1, 291-306.

[2] Goldsmith, J. (1976) Autosegmental Phonology. Ph.D. dissertation, MIT.

[3] Prince, A. (1983) "Relating to the Grid", *LI* 14, 19-100.

[4] McCarthy, J. (1986) "OCP Effects: Gemination and Antigemination," *LI* 17, 207-64.

[5] Kenstowicz, M. and C.W. Kisseberth. 1977. *Topics in Phonological Theory.* Academic Press, N.Y.

[6] Itô, J. (1986) Syllabic Theory in Prosodic Phonology. Ph.D. dissertation, Univ. of Massachusetts, Amherst, Ma.

[7] Goldsmith, J. (1990) *Autosegmental and Metrical Phonology.* Oxford: Basil Blackwell.

[8] Halle, M. and J.-R. Vergnaud (1979) "Metrical Structures in Phonology," ms., MIT.

[9] Hayes, B. (1980) A Metrical Theory of Stress Rules. Ph.D. dissertation, MIT.

[10] Clements, G.N. 1990. "The Role of the Sonority Cycle in Core Syllabification." In J. Kingston and M. Beckman (eds.), *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech.* Cambridge University Press, Cambridge, 283-333.

[11] Hockett, C.F. (1955) *A Manual of Phonology.* Memoir 11 of *IJAL* 21.4, Part 1. Baltimore: Waverley Press. Reprinted by the University of Chicago press, 1974.

[12] Clements, G.N. (1986) "Compensatory Lengthening and Consonant Gemination in Luganda," in L. Wetzels and E. Sezer, eds., *Studies in Compensatory Lengthening,* Foris Publications, Dordrecht, pp. 37-77.

[13] Scobbie, J.M. (1991) "Towards Declarative Phonology", *Edinburgh Working Papers in Cognitive Science* 7, 1-26.

[14] Paradis, C. and D. Lacharité, eds. (1993): *Constraint-based Theories in Multilinear Phonology* (Special Issue of the *Canadian Journal of Linguistics*, vol. 38.2)

[15] Prince, A. and P. Smolensky (1993) Optimality Theory: Constraint Interaction in Generative Grammar. Ms. (in press)

[16] Odden, D. (1982) "A Nonlinear Approach to Vowel Length in Kimatuumbi," unpublished ms., Ohio State University.

[17] Stewart, J.M. (1983) "Akan Vowel Harmony: the Word Structure Conditions and the Floating Vowels," *Studies in African Linguistics* 14, 111-39.

[18] Kaye, J., J. Lowenstamm, and J.-R. Vergnaud (1985) "The Internal Structure of Phonological Elements: a Theory of Charm and Government," *Phonology Yearbook* 2, 305-328.

[19] Singh, R. (1987) "Well-formedness Conditions and Phonological Theory." In W.U. Dressler et al., eds., *Phonologica 1984.* Cambridge: CUP, pp. 273-85.

[20] Calabrese, A. (1988) Towards a Theory of Phonological Alphabets. Doctoral dissertation, MIT, Cambridge, Ma.

[21] Paradis, C. (1988) "On Constraints and Repair Strategies," *The Linguistic Review* 6, 71-94

[22] Myers, S. (1991) "Persistent Rules," *LI* 22, 315-344.

[23] Lacharité, D. and C. Paradis (1993) "Introduction: the Emergence of Constraints in Generative Phonology and a Comparison of Three Recent Constraint-based Models," in [14], pp. 127-53.

[24] Wheeler, D. (1981) Aspects of a Categorial Theory of Phonology. Ph.D. dissertation, University of Massachusetts, Amherst.

[25] Wheeler, D. (1988) "Consequences of Some Categorially Motivated Phonological Assumptions," in R. Oehrle, et al. (eds.)*Categorial Grammars and Natural Language Structures,* Dordrecht: Reidel.

[26] Bird, S. (1990) Constraint-based Phonology. Ph.D. dissertation, University of Edinburgh.

[27] Scobbie, J.M. (1991) Attribute Valued Phonology. Ph.D. dissertation, University of Edinburgh

[29] Coleman, J. (1992) "The Phonetic Interpretation of Headed Phonological Structures Containing Overlapping Constituents," *Phonology* 9.1, 1-44.

[30] Kenstowicz, M. and C.W. Kisseberth (1979) *Generative Phonology: Description and Theory.* Academic Press, N.Y.

[31] Scobbie, J.M. (1993) "Constraint Violation and Conflict from the Perspective of Declarative Phonology," in Paradis and Lacharité, pp. 155-67.

[32] Kiparsky, P. (1973) "Elsewhere in Phonology," in S. Anderson and P. Kiparsky, eds., *A Festschrift for Morris Halle.* New York: Holt, Rinehart and Winston. pp. 93-106.

[33] Prince, A. and P. Smolensky (1993) Optimality Theory: Constraint Interaction in Generative Grammar. Ms.

[34] McCarthy, J. and A. Prince (1993) Prosodic Morphology I: Constraint Interaction and Satisfaction. Ms.

[35] Zec, D. (1994) "Coda Constraints and Conditions on Syllable Weight," ms., Cornell University

NATIVE AND LOANWORD PHONOLOGY AS ONE: CONSTRAINTS VS. RULES*

Carole Paradis
Laval University, Quebec, Canada

ABSTRACT

Constraint-based theories have gained increasing recognition over the past four years. This paper aims to show the superiority of one of these theories, the Theory of Constraint and Repair Strategies, over a rule-based approach with respect to loanword behavior. While the latter requires phonology to be split into two sets of language-specific processes — one for loanwords and another one for native words — the former proposes a unique set of universal processes.

1. INTRODUCTION

Constraint-based theories — which are characterized by the rejection of arbitrary "rules" — certainly constitute the liveliest area in current linguistics. Among these theories the best known in phonology are Optimality Theory (Prince & Smolensky 1993), Harmony Phonology (Goldsmith & Larson 1992; Goldsmith 1993), the Theory of Constraints and Repair Strategies (TCRS) (Paradis 1988a,b; Paradis & LaCharité 1993) and Declarative Phonology (Scobbie 1991; Bird et al. 1993). Except perhaps for proponents of the latter, whose primary purpose is the computerization of phonology (cf. Paradis & LaCharité 1993), proponents of constraint-based theories claim that constraint-based accounts are more explanatory than those couched in rule-based theories — the first being that of Chomsky & Halle (1968) — in the sense that the former capture more generalizations, remove a great deal of arbitrariness and redundancy from analyses, and make more predictions on language-specific and universal bases. This paper defends this view — which is already strongly supported by internal (native) evidence — within the framework of TCRS, and on the grounds of loanword adaptation.

Constraints are often detected when violated because a violation normally yields a deviation from what would otherwise be expected. TCRS identifies morphological operations as the main source of constraint violations. For instance, consider the case of the mid vowel [ɔ] in French, which never occurs word-finally. Its absence can be interpreted in two ways: as an accidental gap or as evidence for a constraint against [ɔ] in word-final position. The second option is selected because there is what I will call «dynamic» phonological evidence provided by the morphology of French supporting it. Two pieces of evidence come from the vocalic alternation found in adjectival inflection such as *sot* [sɔ̲] (*[sɔ])/*sotte* [sɔ̲t] 'stupid (masc./fem.)' and in verbal derivation such as *complot* [kɔ̃plɔ̲] (*[kɔ̃plɔ̲]) 'plot'/*comploter* [kɔ̃plɔ̲te] 'to plot'. Note that the existence of adjectives such as *chaud* [ʃɔ̲] / *chaude* [ʃɔ̲d] (*[ʃɔ̲d]) 'hot (masc./fem.)' and verbal derivations such as *endos* [ɑ̃dɔ̲] 'endorsement'/ *endosser* [ɑ̃dɔ̲se] (*[ɑ̃dɔ̲se]) 'to endorse' in French — where the vowel [o] is realized in word-final and non-final position — clearly indicates that the prohibition bears on the vowel [ɔ] in word-final position, not the vowel [o] in non-final position. Abbreviation constitutes another source of evidence for the constraint: *professionnel* [prɔ̲fɛsjɔnɛl] → *pro* [prɔ̲] (*[prɔ]) 'professional', *Carole* [karɔ̲l] → *Caro* [karɔ̲] (*[karɔ]) 'Carol', *police* [pɔ̲lis] → *popo* [popɔ̲] (*[pɔpɔ]) 'police', etc. In all cases, the underlying vowel /ɔ/, which surfaces in non-final position, yields [o] at the end of abbreviations, thus showing that the process is too general, i.e. it occurs in too many different morphological contexts, and too systematic (there is no exception) not to be the result of a phonological constraint against ɔ# in French.

* I am deeply indebted to Heather Goad and Jean-François Prunet for precious comments. Thanks also to Yvan Rose for his help with the editing of the paper. The author remains solely responsible for any remaining errors or omissions. I acknowledge SSHRCC grant # 410-94-1296 and FCAR grant # 95-ER-2305.

However, it is common for linguists to be left only with «static» evidence, i.e. the absence of an element or structure *x* in a given language, to suspect the existence of a constraint. For instance, it can be observed that #CC sequences do not exist in Fula, a West-African language. It is tempting for a linguist to resort to a constraint to express this fact, but one does not know with certainty whether the lack of such a structure is due to an accidental gap in the language, a diachronic constraint or a synchronic one (cf. Paradis & Prunet 1993). Derivation and inflection of native words do not provide any insight here since there is no morphological operation in Fula which would generate such a sequence, i.e. there is no mono-consonantal prefix which would attach to a consonant-initial word, and thus yield a #CC cluster.

This is where borrowings play a crucial role: they often contain elements or structures that are absent from the native vocabulary. Depending on how these foreign elements and structures are treated by the borrowing language — is *x* accepted or systematically modified (adapted)? — the linguist may know whether the absence of such elements or structures in the studied language is due to a constraint or an accidental gap. For instance, Fula has borrowed extensively from French, a language with branching onsets. Adaptation of French borrowings with such onsets provides dynamic evidence for the existence of a constraint against #CC clusters in Fula since all such French clusters are automatically modified in Fula. They usually yield #CVC sequences, i.e. sequences with a vowel inserted in between the two consonants (e.g. Fr(ench) *tracteur* [tʀaktœʀ] 'tractor' → F(ula) [taraktɔr] and Fr. *place* [plas] 'place' → F. [palas]). From the perspective of TCRS, borrowings constitute an invaluable source of constraint violations, which allow the linguist to observe how a language "reacts" to unfamiliar elements or structures.

Paradoxically, however, the fact that these phonological "reactions" are sometimes restricted to loanwords — for the reasons we have just seen in Fula, i.e. there is sometimes no context in the language from which a constraint violation might stem, and thus no possible "reaction" to violations — has led some linguists to conclude that there were two separate sets of phonological processes, one for loanwords and one for native words. Silverman (1992) is among the ones who maintain this view the most explicitly. Such a position, which stems from a rule-based perspective, is at best useless in a constraint-based view (cf. also Yip 1993: 262). I will show that what I call the Two Process-Set Hypothesis, in (1), entails non-desirable effects such as duplicating identical processes in the same language and, above all, missing important links among facts, on language-specific and universal grounds.

(1) *Two Process-Set Hypothesis*:
Loanwords and native words each have their own set of processes (rules).

To this effect, we will examine three constraints (*CC#, *CC, *#V) each in a different language (Fula, Kinyarwanda and Moroccan Arabic, respectively), and observe how the processes triggered by these constraints would have to be handled in a rule-based approach. The paper will be organized as follows. Section 2 presents my assumptions regarding borrowings (2.1 and 2.3), and the relevant tenets of TCRS (2.2). Section 3 addresses the three constraints mentioned above, while section 4 offers a brief conclusion.

2. ASSUMPTIONS

2.1 Borrowings

Two opposite views are debated in loanword studies: the "phonetic approximation stance" (e.g. Haugen 1950 and Silverman 1992), where a borrowed word is analyzed as a non-linguistic acoustic signal, and the "phonological stance" (e.g. Hyman 1970 and Prunet 1990) where a borrowed word is instantaneously assigned a mental representation in the recipient language (L1). Strong arguments based on sociolinguistic, psycholinguistic and phonological studies have been recently brought forward by Paradis et al. (1995a,b) in favor of the phonological stance. For instance, sociolinguistic studies (e.g. Haugen 1950 and Poplack et al. 1988) clearly indicate that borrowings are introduced by bilinguals (not monolinguals), who have ac-

cess to the phonology of the source language (L2). Loanwords are introduced by bilinguals through what sociolinguists call "code-switches", "nonces" and "idiosyncrasies". Sociolinguistic studies also clearly show that phonological patterns of adaptation are imposed by bilinguals; they are community-wide, especially in mid and high community bilingualism stages. This indicates that borrowing integrators and adapters have access to word representations in L2.[1] Otherwise adaptations could not display the strong consistency observed by Haugen (1950) in the mid and high community bilingualism stages, and by us in our own corpora of loanwords (cf. Paradis et al. 1993, 1995a,b for a thorough argumentation in favor of the phonological stance).

2.2 Framework: TCRS

In TCRS, a language's phonology consists of both universal and non-universal constraints which, when violated, trigger the application of a repair strategy (e.g. *ɔ# → *o* in section 1), defined in (2).

(2) *Repair strategy*: A universal, context-free phonological operation that is triggered by the violation of a phonological constraint, and which inserts or deletes content or structure to ensure conformity to the violated constraint.

As mentioned in section 1, TCRS claims that constraint violations originate mainly from morphological operations (e.g. the constraint *ɔ# discussed in section 1, which is violated because of an abbreviation operation, etc.). Other internal sources include constraint conflicts and underlying ill-formedness (Paradis 1988a, b). Loanwords (Paradis et al. 1993) and paraphasias (Béland et al. 1993) constitute external sources. However, while violated constraints press for repair, the Preservation Principle, (3), protects the input, i.e. resists segmental loss.

(3) *Preservation Principle*: Segmental information is maximally preserved within the limits of the Threshold Principle ((4)).

I maintain that the Preservation Principle is responsible for the low rate of segment deletion observed in the four corpora of loanwords that we have built (4,031 borrowings from French into Kinyarwanda, Moroccan Arabic and Fula, and English into Quebec French), which contain altogether 12,630 malformations. The Preservation Principle works in the following way. Repair is accomplished by the insertion or deletion of content (e.g. features, timing units, etc.) or structure (links between features, various levels of structure, etc.). At its most basic, repair by insertion occurs when a constraint violation is due to a lack of content or structure whereas deletion applies when a constraint is offended by an excess of content or structure. Whether a problem is due to a lack of something or an excess of something is often a matter of perspective. For example, in a language with a constraint against consonant clusters (CC) such as Kinyarwanda, a CC (loan) input can be regarded as an excess of consonants, leading to deletion (of a consonant), or as the lack of a vowel, leading to insertion (of a vowel). All else being equal, the Preservation Principle, which resists the loss of phonological information, favors viewing a problematic structure as a lack of content or structure, giving preference to insertion over deletion.

TCRS nevertheless posits limits to preservation, i.e. to the price languages are ready to pay to conserve segmental information. This is expressed by the Threshold Principle in (4).

(4) *Threshold Principle*:
a) All languages have a tolerance threshold to segment preservation.
b) This threshold is the same for all languages: two steps (or two repairs) within a given constraint domain.[2]

[1] What is the exact nature (lexical or phonetic) of these representations is a question which has not been totally settled yet. The evidence gathered by Paradis et al. (1995a,b) tend to show that this representation is lexical, not phonetic.

[2] This limit has been found to hold for Fula (Paradis & Lebel 1994) and for Kinyarwanda (Rose 1994). We therefore hypothesize that it is a universal ceiling on the cost of adapting, as opposed to deleting a problematic structure. Should the threshold be set differently in other languages, the second part of the principle, (4b), would have to be parametrized.

The Threshold Principle stipulates that a problematic segment requiring more than two steps to be adapted within a constraint domain — a constraint domain being simply the scope of a constraint violation — is not protected by the Preservation Principle.

Repair, be it by deletion or insertion, must nevertheless apply economically. Economy is expressed first and foremost by the Minimality Principle in (5).

(5) *Minimality Principle*:
a) A repair strategy must apply at the lowest phonological level to which the violated constraint refers.
b) Repair must involve as few strategies (steps) as possible.

The "lowest phonological level" referred to in (5a) is determined by the Phonological Level Hierarchy (PLH), in (6), which simply reflects the phonological organization required independently of TCRS.

(6) *Phonological Level Hierarchy*:
Metrical level > syllabic level > skeletal level > root node > feature with a dependent > feature without a dependent.

The Preservation Principle in (3) is served by (5a) which minimizes alteration of the input, for example disallowing the loss of a syllable, if the loss of a segment will correct the problem. In other words, it ensures that a constraint violation is solved with as little loss of phonological information as possible. (5b), for its part, requires that, given more than one possible way of repairing an ill-formed structure, priority be given to the repair involving the fewest steps.

TCRS maintains that the phonological structure of a language results from principles (universal constraints) and parameter settings. Principles describe what is common to all languages, whereas parameter settings handle differences (contrasts) among languages (cf. Chomsky 1986). In TCRS, parameters are marked options offered by Universal Grammar. The default reply for a language is to say "no" to such an option, which results in the rejection of a given type of complexity, and thus a negative constraint in the language in question. In this perspective, the segmental inventory of a language is viewed as the direct result of positive and negative language-specific answers (settings) to segmental options offered by Universal Grammar (parameters). In the case of borrowings, one can thus hypothesize that the reason why French *coupon* [kupɔ̃] is realized as [ku:pɑn] in English, i.e. with a (partly) denasalized vowel followed by a nasal consonant is because English says "no" to the following parameter.

(7) *Phonemic nasal vowels?*
French: *yes*
English: *no* (default = constraint)

The negative parameter setting in (7) explains why nasal vowels introduced into English through loanwords are adapted. In the view of TCRS, the recasting of ṽ into a VN shape is not the result of a rule specific to loanwords — as would be the case with the Two Process-Set Hypothesis — but of a constraint active throughout the phonology of English, whose only source of violation is loanwords. This position, that I call the One Process-Set Hypothesis, is formalized in (8).

(8) *One Process-Set Hypothesis*:
Phonology has access to a *single set* of two universal processe: insert *x* and delete *x*. These processes are repair strategies, whose sole purpose is to yield constraint satisfaction. If there is no constraint violation, they do not apply.

2.3 Core and Periphery

The One Process-Set Hypothesis does not imply, however, that the phonological behavior of loanwords and native words is identical in all respects. If we consider again the case of nasal vowels introduced into English, we realize that while nasal vowels are totally absent from native English words, they are sometimes tolerated in borrowings (e.g. Fr. *entrée* [ɑ̃tre] → English [ɑntɹɪ] or [ɑ̃tɹɪ])). In a study of loanword adaptation, it is crucial to distinguish between "prohibited" segments, i.e. segments that are systematically and immediately adapted or

eliminated as soon as they are introduced into a language (e.g. the French front round vowels *y* and *ø* in English), and "tolerated" segments, which are sometimes adapted and sometimes not (such as the French nasal vowels in English) at least in some speech registers. The latter are called "imports" in the literature (cf., e.g., Haugen 1950). To account for the distinction between prohibited and tolerated segments, the TCRS loanword model proposed by Paradis et al. (1995b) views the phonology of a language as being organized into domains. Essentially, a distinction is drawn between the "core" and the "periphery". The core contains all of a language's constraints; by and large, the core defines the phonology of a language and governs its vocabulary. However, not all items in a language are part of the core; some, such as interjections, onomatopoeia, proper names and learned words, along with (partly) unassimilated borrowings, may lie in the periphery, temporarily or even indefinitely. The periphery contains a subset of a language's constraints, which means that items in the periphery are not subject to all the constraints that govern the core. That is to say, the parameter settings for some Universal Grammar options may be set to "yes" rather than "no" in the periphery or some subdomains of the periphery, which effectively deactivates those particular constraints, and accounts for imports (unassimilated foreign sounds). The distinction between core and periphery is not particular to TCRS. It was suggested by Chomsky (1986:147), and further developed by Itô & Mester (1993). However, the core and the periphery are not different in nature. The periphery is not governed by "new" constraints, i.e. constraints different from those of the core. It contains only "fewer" constraints than the core. In this view, a "borrowing" can be defined as in (9).

(9) *Borrowing*: An individual word, or compound functioning as a single word, from L2 that a) phonologically conforms to (at least) the outermost peripheral phonological constraints of L1, b) has a mental representation in L1, and c) is incorporated into the discourse of L1 (cf. Paradis et al. 1995a,b for more details).

3. CONSTRAINTS VS. RULES

3.1 Language-Specific Issues

Words in Fula never surface with a final CC cluster or an internal CCC one. We know that this is due to a constraint against branching codas because when such a cluster is present underlyingly or arises in the course of a morphological derivation, it is immediately split into different syllables as in (10).

(10) *Native words in Fula*

fɔɔft-rɛ	→ fɔɔft-ɛ-rɛ	'breath'
talk-ru	→ talk-u-ru	'amulet'
lacc-ri	→ lacc-i-ri	'couscous'

The constraint is formally expressed by the negative parameter setting in (11) (recall from section 1 that Fula does not allow branching onsets either).

(11) Parameter:
Branching non-nuclear constituents?
French: *yes* Fula: *no* (constraint)

As shown in (12), the constraint also applies to loanwords since CC# clusters in those words undergo vowel insertion too.

(12) *French Loanwords in Fula*
Fr. carde [kard]
→ F. karda 'card (comb)'
Fr. force [fɔrs]
→ F. fɔrsɔ Fula 'force'
Fr. gendarme [ʒɑ̃darm]
→ F. san^{n}darma 'gendarme'

With the Two Process-Set Hypothesis, one would have to posit two separate rules, as in (13), even though both rules would be identical.

(13) *Two Process-Set Hypothesis*:
a) native words: Ø → V/CC__{#, C}
b) loanwords: Ø → V/CC__{#, C}

This reduplication of identical rules is seriously flawed in two ways. First, it complicates the grammar. Second, it does not formally capture the fact that both rules are actually the same process (vowel insertion) which is triggered by the same context ({#, C}) in loanwords as in native words. These disadvantages are eliminated with the One Process-Set Hypothesis. As shown in (14), the data in (10) and (12) necessitate only one context-free universal process, i.e. insertion of *x*.

(14) *One Process-Set Hypothesis*:
native and borrowed words: Ø → x

While the Preservation Principle in (3) ensures that insertion will have priority over deletion, the Minimality Principle guarantees that the material inserted will pertain to the level to which constraint (11) refers (cf. (5a)). Since (11) refers to the syllabic level, the repair will apply at that level. Insertion of a nucleus is selected because this is the only repair which fully satisfies both principles, the Preservation and Minimality Principles. The empty nucleus is subsequently filled by vowel spreading.

The rule-based approach, in which the Two Process-Set Hypothesis is couched, is problematic in other respects. Consider the French borrowings in (15), where vowel insertion occurs in between the two consonants of a CC# cluster, not at the end of it as in (12).

(15) Fr. contre [kɔ̃tr]
→ F. kɔntɔr 'against'
Fr. filtre [filtr]
→ F. filtir 'filter'
Fr. table [tabl]
→ F. taabal 'table'

Not only would the Two Process-Set Hypothesis require the reduplication of the same rule as in (13), the rule-based approach in which it lies, more generally, would require positing a third rule — shown in (16) — to account for the facts in (15).

(16) Ø → V/C__C#

This new rule would be needed because the context of rule (13b) is not identical to that of rule (16). Again, the fact that the trigger is a CC# cluster would be missed. This generalization is straightforwardly captured by constraint (11), however. CC# clusters are prohibited because they would form an illicit branching coda. The insertion locus of the vowel depends entirely on the sonority of the cluster. It is determined by universal markedness, which disfavors syllabic contacts where an onset is more sonorous than the preceding coda, even though such clusters are found in some Fula native words (e.g. *faabru* 'toad'). In other words, in the absence of opposite morphologically-induced specifications, default settings, provided by Universal Grammar, apply. From this perspective the phonological behavior of loanwords tells us significantly more about universal default settings than that of native words, which is often morphologized or heavily influenced by diachrony. Once distorting factors such as orthography, analogy, etc. are clearly identified and discarded, one can easily state that loanword phonology *is* the "emergence of the unmarked" in phonology (cf. McCarthy & Prince 1994 on this notion).

In a rule-based approach, a fourth rule would even have to be posited. As explained above, the Preservation Principle gives precedence to vowel insertion over consonant deletion. However, consonant deletion does occur in a few cases such as those in (17), where *v* is lost.[3]

(17) Fr. pieuvre [pjœvr]
→ F. pijuri 'octopus'
Fr. cuivre [kɥivr]
→ F. kiri 'copper'

However, as shown in Paradis et al. (1993, 1995a,b), consonant deletion is not random. It is always caused by the presence of an ill-formed segment — here the voiced labial fricative **v* — contained within an unsyllabifiable cluster. Preservation of the two cluster consonants would be too costly in these cases: it would necessitate too many steps (repairs). It would require nucleus insertion and filling as in (10), (12) and (15). But it would also require a third step, i.e. the adaptation of the ill-formed segment **v* itself (**v* normally yields *w* in Fula; e.g. Fr. *verre* [vɛr] → F. [wɛːr]), since it is encompassed within the scope of constraint (11). This would clearly violate the Threshold Principle in (4), which establishes that the limit to segmental preservation is two repairs, within a given constraint domain. Thus not only does TCRS account for the variation in the insertion point of the epenthetic vowel in (12) and (15) without any extra language-specific device, but it also handles straightforwardly the variation in the processes themselves, i.e. insertion of a vowel ((10), (12) and (15)) vs. deletion

[3] More exactly, phonologically-induced deletions in the Fula corpus occur with 32 malformations out of 858 (3.7% of cases).

of a consonant in (17) (cf. Paradis et al. 1995a,b for more examples and a thorough discussion of these cases). In contrast, a rule-based approach is unable to economically capture this variation, as well as being unfit to perceive the link between the numerous rules it would require to handle the data presented in this section.

3.2 Universal Issues

A rule-based approach would be undesirable on universal grounds also. It would treat the processes observed in the previous section as idiosyncrasies of Fula, despite the fact that restrictions on branching codas are common among languages. Such restrictions are found in Tigrinya and Classical Arabic, for instance. This fact is predicted by TCRS since constraints in TCRS' view stem from negative parameter-settings. Since parameters are options offered by Universal Grammar, it is predicted that a number of languages will share the same parameter setting, be it positive or negative. Recall from 2.2 that negative parameter settings are default (unmarked) options: they consists in a language's refusal of a given type of complexity. Negative settings are thus expected to be relatively frequent.

The same is true of the *CC constraint in Kinyarwanda, a Bantu language, and the *#V constraint in Moroccan Arabic, which respectively prohibit codas and empty onsets. Both constraints, which are formalized in (18a) and (18b) respectively, are common across languages. For instance, the former is found in Luganda as well as in most Bantu languages, while the latter is found in Tigrinya, Biblical Hebrew and many other Semitic languages.

(18) a) Parameter: *codas?*
French, English: *yes*
Kinyarwanda, Luganda: *no*
b) Parameter: *empty onsets?*
French, English: *yes*
Moroccan Arabic, Tigrinya: *no*

The constraints in (18) are supported internally, and also externally by the behavior of borrowings like those in (19a) and (19b).

(19) a) *French → Kinyarwanda:*
client [klijã] → [umu-cirija]
citron [sitrɔ̃] → [sitoro]
b) *French → Moroccan Arabic:*
i. arbitre [arbitr] → [larbiṭ]
ii. ascenseur [asɑ̃sœr] → [sensur]

In (19a), we can see that French CC sequences are automatically separated by an epenthetic empty nucleus in Kinyarwanda, to which the following vowel spreads. In Moroccan Arabic, a violation of (18b) triggers more diversified repairs, i.e. either insertion of a consonant, as in (19bi), or deletion of the initial vowel as in (19bii). Selection of one repair over another here is conditioned by the length of the output (cf. Paradis et al. 1995b). The longer the output in L1, the more likely vowel deletion is. Nonetheless, both strategies fully preserve (18b) in preventing a vowel from surfacing word-initially. Again, this principled diversity of repairs could not be captured in an explanatory way in a rule-based approach. In such a framework, two completely unrelated rules would have to be posited, thus failing to express the fact that the trigger (*#V) is identical in both cases.

4. CONCLUSION

This paper has attempted to show the superiority of constraints over rules in general. More specifically, TCRS and the traditional rule-based approach of Chomsky & Halle (1968) — which continued to be used under different forms in multilinear phonology and pre-constraint-based frameworks — were compared in their capability to deal with loanwords. The former has proved markedly more economical and explanatory. In particular, it has rendered the Two Process-Set Hypothesis — where loan words and native ones are considered to be each governed by a distinct set of processes — vacuous. On more universal grounds, it was shown that the processes applying to borrowings and native words are not language-specific idiosyncrasies but the result of the language's replies to options offered by Universal Grammar, i.e. parameters. The phonological behavior of borrowings, which seems ad hoc in a rule-based view, proves very regular and predictable in TCRS. On the one hand, TCRS provides linguists with a formal framework which handles straightforwardly one of the richest sources of dynamic evidence for constraints: borrowings. On the other hand, the study of borrowings opens a large window on the general functioning of constraints, and ultimately the organization of the language in the human brain, by allowing us to observe how languages react to foreign elements.

REFERENCES

[1] Béland, R., C. Paradis & M. Bois (1993), "Constraints and Repairs in Aphasic Speech: A Group Study", in C. Paradis & D. LaCharité, eds., pp. 279-302.
[2] Bird, S., J. Coleman, J. Pierrehumbert & J. Scobbie (1993), "Declarative Phonology", *Proceedings of the XV^th International Congress of Linguists*, A. Crochetière et al., eds., Québec: Presses de l'Université Laval.
[3] Chomsky, N. (1986), *Knowledge of Language: Its Nature, Origin and Use*, New York: Praeger.
[4] Chomsky, N. & M. Halle (1968), *The Sound Pattern of English*, New York: Harper & Row.
[5] Goldsmith, J. (1993), *The Last Phonological Rule: Reflections on Constraints and Derivations*, Chicago: University of Chicago Press.
[6] Goldsmith, J. & G. Larson (1992), "Using Neural Networks in a Harmonic Phonology", *Proceedings of the Chicago Linguistic Society*, C. Canakis et al., eds., Chicago: The University of Chicago Press, pp. 94-125.
[7] Haugen, E. (1950), "The Analysis of Linguistic Borrowings", *Language* 26: 210-231.
[8] Hyman, L. (1970), "The Role of Borrowings in the Justification of Phonological Grammars", *Studies in African Linguistics*: 1: 1-48.
[9] Itô, J. & A. Mester (1993), "Japanese Phonology: Constraint Domains and Structure Preservation", *A Handbook of Phonological Theory*, J. Goldsmith, ed., Cambridge, Mass.: Blackwell.
[10] McCarthy, J. & A. Prince (1994), "The Emergence of the Unmarked: Optimality in Prosodic Morphology", *Proceedings of the North Eastern Linguistic Society*, M. Gonzàlez, ed., Amherst: University of Massachusetts, pp. 333-379.
[11] Paradis, C. (1988a), "On Constraints and Repair Strategies", *The Linguistic Review* 6: 71-97.
[12] __ (1988b), "Towards a Theory of Constraint Violations", *McGill Working Papers in Linguistics* 5: 1-43.
[13] Paradis, C. & D. LaCharité, eds. (1993), *Constraint-Based Theories in Multilinear Phonology*, Canadian Journal of Linguistics 38: 127-303.
[14] Paradis, C., D. LaCharité & C. Lebel (1995a), "Savings and Cost in French Loanword Adaptation in Fula: TCRS Predictions", *McGill Working Papers in Linguistics*. [in press]
__ (1995b), "Preservation and Minimality in Loanword Adaptation", ms.
[15] Paradis, C. & C. Lebel (1994), "Contrasts from Segmental Parameter Settings in Loanwords: Core and Periphery in Quebec French", *Proceedings of the MOT Conference on Contrasts in Phonology*, C. Dyck, ed., Toronto Working Papers in Linguistics 13: 75-94.
[16] Paradis, C., C. Lebel & D. LaCharité (1993), "Adaptation d'emprunts: les conditions de la préservation segmentale", *Proceedings of the 1993 Canadian Association Meeting*, C. Dyck, ed., Toronto Working Papers in Linguistics, pp. 461-476.
[17] Paradis, C. & J.-F. Prunet (1993), "On the Validiy of Morpheme Structures Constraints", in C. Paradis & D. LaCharité, eds., pp. 235-256.
[18] Poplack, S., D. Sankoff & C. Miller (1988), "The Social Correlates and Linguistic Processes of Lexical Borrowing and Assimilation", *Linguistics* 26: 47-104.
[19] Prince, A. & P. Smolensky (1993), "Optimality Theory: Constraint Interaction in Generative Grammar", ms.
[20] Prunet, J.-F. (1990), "The Origin and Interpretation of French Loans in Carrier", *International Journal of American Linguistics* 56: 484-502.
[21] Rose, Y. (1994), "L'adaptation des voyelles nasales en kinyarwanda: l'effet du principe de préservation", *Actes des VIII^e Journées de linguistique*, F. Belyazid et al., eds., pp. 161-165.
[22] Scobbie, J. (1991), "Attribute Value Phonology", Ph.D. thesis, University of Edinburgh.
[23] Silverman, D. (1992), "Multiple Scansions in Loanword Phonology: Evidence from Cantonese", *Phonology* 9: 289-328.
[24] Yip, M. (1993), "Cantonese Loanword Phonology and Optimality Theory", *Journal of East Asian Linguistics* 2: 261-291.

THE VOICE SOURCE. MODELS AND PERFORMANCE

Gunnar Fant
Dept of Speech Communication and Music Acoustics, KTH, Stockholm, Sweden

ABSTRACT

This is a summary of research into functional models of the human voice source with considerations to production theory, experimental techniques and individual and contextual variations in connected speech. The emphasis is on work carried out in our department, including the development of a transformed LF-model, and studies of source-tract interaction. The voice source as a prosodic parameter is discussed. Of special interest is the covariation of source parameters, F_0 and inferred contours of lung pressure variations found in focal accentuation.

INTRODUCTION

A major tool for the study of the human voice source is inverse filtering. Over the years a substantial amount of work in this area has been carried out at KTH, see the review in [1].

Inverse filtering is a processing of undressing the vocal tract filter function of the speech wave thus regenerating a replica of the underlying source. This process provides us with some insight in the production mechanism and also a physical substance to be quantified and described within a suitable parameter system.

Early parameter systems concentrated on main shape aspects of glottal flow pulses such as rise time, decay time and open quotient. The importance of the flow discontinuity at closure as an excitation function was early discovered in connection with inverse filtering and was included in a Laplace transform production modeling in 1979 [2]. Five years later the importance of the return phase in the flow derivative was fully acknowledged [3] and became a major constituent of the LF-model [4]. The effective duration of the return phase, T_a, was proved to be inversely proportional to a frequency $F_a=1/2\pi T_a$ where the source spectrum attains an extra -6dB oct slope. Increasing T_a thus implies a low pass filter effect, a relative attenuation of formants located above Fa. This parameter is usually of greater significance than the main pulse shape parameters.

The ability to capture wave shape essentials has promoted a wide use of the LF model. However, human data from inverse filtering may deviate substantially from model data, and mainly in terms of a superimposed fine structure which displays both typical recurrent patterns and a seemingly randomness. The underlying mechanisms for this structure has been extensively studied in several publications from KTH [1, 5-8].

There exist systematic covariations in the LF parameters which have been exploited in a transformed version [9] of the model. It operates with a fewer number of parameters retaining wave shape essentials, combined with a more detailed specification in terms of deviations of the original LF-parameters from default values. This new system also has advantages from an experimental point of view and as a basis for rule oriented speech analysis and synthesis.

The covariation of source and filter functions, in more general terms phonatory and articulatory processes, is of particular interest. It is the combined gesture rather than the source function alone which has a communicative function. Supraglottal constrictions impede the voice source [10] and glottal abduction introduces additional

bandwidths and F_1 increase, subglottal coupling and aspiration noise adding to the source features [7-8, 11-12]]

A specific topic of interest in prosody is the coordination of glottal adjustments, adduction/abduction gestures and F_0-control, and lung pressure. There are apparent differences between singing and speech that need to be studied in greater detail, e.g. vowel consonant contrast, relative emphasis and accentuation.

BASIC SOURCE-FILTER MODEL

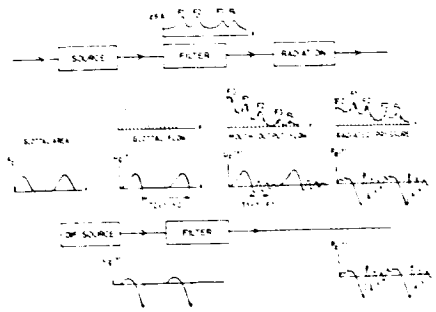


Figure 1. Frequency- and timedomain view of the production of voiced sounds.

The basic concept of source-filter decomposition of voiced sounds in the frequency domain and in the time domain is illustrated in Figure 1. It conveys the traditional view of the source as a raw material of spectral harmonics which is shaped by a filter function. The latter, imposing the formant structure is made up of two parts, the vocal tract transfer function relating the volume velocity flow at the lips to the glottal flow, and a radiation transfer from flow at the lips to the radiated sound pressure wave at some distance from the lips. The radiation transfer is usually approximated by a simple differentiation, in the frequency domain a +6dB octave spectral rise.

In the time domain representation a glottal flow pulse is a skewed version of the glottal areafunction. Glottal parameters are often defined with respect to the time derivative of glottal flow. One advantage of the differentiated source, see the bottom part of the figure, is that it accounts for the radiation transfer component.

Production theory [2,7] states a proportionality between the amplitude of glottal flow derivative at its negative discontinuity, which usually is identical to the negative peak, and formant amplitudes. With an abrupt return to the zeroline and assuming a single formant filter function there is a continuity between the negative peak amplitude E_e and the initial amplitude of the corresponding damped oscillation in the radiated wave This is indicated in the figure. However, the mouth output volume velocity flow, which is the integral of the radiated wave, shows a relative reduction of oscillatory energy but retains the pulse shape of the initial (non-differentiated) glottal flow.

As a matter of fact, integrating the speech wave provides an approximation to the maximum amplitude of glottal flow U_o, constant leakage omitted, while the E_e amplitude information is approximately retained in the envelope contour of the negative side of the radiated speech wave [1, 9, 13] Since U_o and E_e are the main constituents of glottal waveshape as proposed in the transformed LF model [9], important information about the temporal variation of voice source parameters can be derived without proper inverse filtering.

SELECTIVE INVERSE FILTRING

Inverse filtering experiments confirm these general statements. Figure 2 illustrates regenerated glottal flow and so called selective inverse filtering [1] with cancellation of all formants but one, in this case F1, which appears as a damped oscillation following each glottal flow derivative pulse. The pattern for the [ae] is typical of a sonorous male voice.

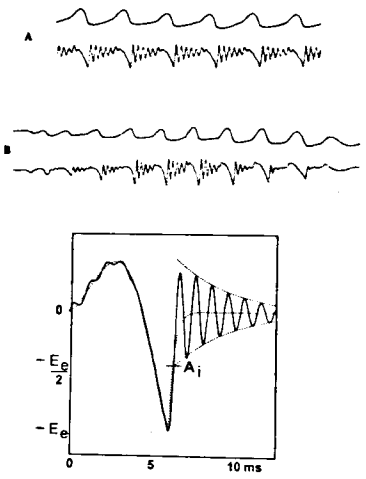


Figure 2. Selective inverse filtering retaining the F1 oscillation. A; vowel [ae]. B; vowel [a] preceding an unvoiced stop. The lower graph is model generated with F_1/F_a=2.

The lower pair of curves pertain to a vowel [a] preceding the occlusion of an unvoiced aspirated stop. Within the series of the three final pulses the F1 oscillation decays more rapidly than the supporting flow derivative E_e residue. This is a matter both of increasing damping, i.e. of a formant bandwidth increase following glottal abduction and of the F1 initial amplitude becoming progressively smaller than the negative E_e peak. The latter is the time domain equivalent to an increasing low pass filtering associated with the decreasing F_a which is also a consequence of the abduction gesture. Figure 3 illustrates the principal relation of the F1 initial amplitude being reduced by the same amount as implied by the spectral tilt [1]. Similar effects also appear when a formant is under influence of a neighboring zero, e.g. in a nasalized vowel.

VOCAL TRACT-SOURCE INTERACTION.

More marked instances of pre-occlusive aspiration is treated in [8, 25] In addition to increased spectral tilt and first formant bandwidth, glottal pulse modulated noise appears in the final part of the vowel, in extreme cases combined with pole-zero spectral modifications and extra formants from the subglottal system Noise components are also consistently found in breathy voices [11-12, 16].

A number of other interaction effects complicate the source-filter interpretation of inverse filtering data. One obvious aspect is that a constant leak during the maximally closed glottal interval will pose a problem of how to tune F1 bandwidth and frequency. If these are set for maximal cancellation the inverse filtering will not provide a picture of the true glottal flow. Instead, an ideal regeneration of the true glottal flow would require a setting of the inverse filter to cancel the supraglottal transfer function alone which differs from that of the complete system and can not be derived from the speechwave. The true flow, which has the theoretical burden of conveying the difference between the coupled and the uncoupled system, may have a more complex fine structure than what is seen in ordinary inverse filtering. An example is the appearance of formant oscillations in the maximally closed phase.

A prominent interaction effects is the nonlinearity of the glottal impedance, i.e. the second power dependency of pressure drop on flow, in combination with the presence in the transglottal pressure drop of oscillations evoked from previous excitations [6-7, 15). A typical feature is the double peak appearance of the positive part of the glottal flow derivative and a corresponding spectral dip around $2F_1$ in the source spectrum. [5, 8]

Other aspects of nonlinearities is that a constant glottal chink may counteract the F_a induced spectral fall in the mid and high frequency range of the source spectrum [7, 17, 18] . It is also found [18] that the T_a of the glottal flow derivative becomes larger than an equivalent T_a of the underlying glottal areafunction.

A consequence of glottal impedance nonlinearity is that the superposition imposed by an integer relation between formant frequency and F_0, i.e. when a harmonic hits the formant peak, also effects the driving source function as well as the vocal folds vibratory pattern.. It has indeed been found that the amplitude of F2 and F3 seem to follow the F_1/F_0 ratio rather than the F_2/F_0 and F_3/F_0 ratios [19] An extreme aspect of the nonlinear superposition is that the air consumption is minimized when F_1 hits F_0 but is maximal when F_1 is in the region of 1.5 F_0 which has consequences for soprano singers [6].

A major aspect of vocal tract-source interaction is that a supraglottal narrowing anywhere in the vocal tract or at the lips will be associated with a pressure drop which reduces the transglottal pressure [9-10] and thereby the excitation amplitude E_e and changes the waveshape of glottal flow, increasing the open quotient and the return time T_a. This effect is maximal in voiced plosives and in voiced fricatives but is also noticeable in narrow vowels and in nasals specially in Swedish [1, 13, 16]

VOICE SOURCE MODELING

We shall now return to the more pragmatic aspects of quantifying voice production and source characteristics. In general, irrespective of the particular parameterisation, we may note the close correspondence between the peak value U_o of glottal flow and the amplitude H_1 of the voice fundamental in a harmonic representation of the source component of the speech wave at a distance a cm from the speaker [8].

$$H_1 = U_o k\pi F_0(\rho/4\pi a) \quad (1)$$

where k is close to 1 for opening qoutients of the order of 0.5-0.7. Adopting the notation $F_a=1/(2\pi T_a)$ where T_a is the effective duration of the return phase we may write the following expression for the amplitude H_m of any harmonic of frequency f_m well above F_0 in the glottal flow derivative spectrum submitted to an extra +6 dB/octave rise with respect to F_0.

$$H_m = (E_e/\pi)(\rho/4\pi a)(1+f_m^2/F_a^2)^{-0.5} \quad (2)$$

The relative levels of the fundamental and the next two harmonics have to be treated separately by an analysis of the specific glottal pulse shape as in (1). The result is an additional reinforcement, a "glottal formant" located at a mean frequency of $F_g=1/2T_p$ and providing a few dB larger gain than implied by (2)

A consistent mapping of time domain features into the frequency domain allows us to perform an inversion and predict glottal flow shape and magnitudes from absolute calibrated spectral data [7].

An alternative to the Fourier analysis is to decompose the glottal pulse into a sequence of discrete excitation functions [2]. This is necessary for the understanding of the details of observed waveforms and of interaction phenomena. Assuming a single bell shaped glottal pulse with a rising branch of $(U_o/2)(1-\cos 2\pi f_g t)$ and a symmetrical falling branch the flow derivative becomes $U_o \pi f_g \sin 2\pi f_g t$ which is similar to that of the LF-model. The derivative discontinuity at the onset of the rising branch thus contributes with a -12 dB/oct spectrum slope, i.e. -18 dB/oct in the flow domain. Providing the falling branch does not include an additional

discontinuity prior to its end it will provide the same excitation function as the rising branch but with opposite phase, and if $T_0=1/F_0=2T_p$, i.e. OQ=1, the net effect in the source domain is a sinewave. This is one extreme condition to be preserved in a parametric scheme. In general, however, formants exited at the onset will be damped out quicker than those at the offset. The major excitation will thus be at the offset even if it does not contain an additional discontinuity. The limiting value of the source spectral tilt is accordingly -12 dB/oct (in the flow derivative) as with an extremely low F_a. On the other hand an abrupt and instantaneous return of the flow derivative at the excitation point T_e provides a spectrum slope of -6dB/oct.

An additional high frequency gain in the source spectrum can be attained only if the duration of the falling branch of the flow is very short, i.e. with an extreme asymmetry and a very small opening quotient, in which case the E_e spike becomes very narrow. This extreme is generally not encountered but can be approached within reasonable limits in simulations.

Occasionally there is to be seen an abrupt step in the flow derivative at the opening phase which adds an excitation of the same type as at closure. This has been taken into account in a modification of the LF model proposed in [23].

THE EXTENDED LF-MODEL

The LF-model [4] is illustrated in Figure 3. We have already discussed the significance of the return phase which accounts for the degree of spectral tilt through the $F_a=1/(2\pi T_a)$ parameter which is frequently used as an alternative to $R_a=T_a/T_0$.

The $R_k=(T_e-T_p)/T_p$ parameter specifies the relative duration of the falling branch from the peak at time T_p to to the discontinuity point T_e.

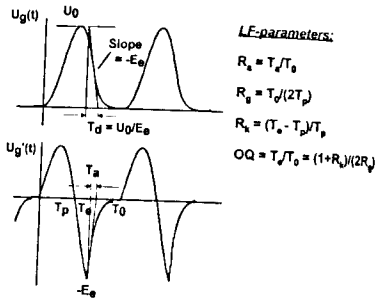


Figure 3. The LF-model. Glottal flow and flow derivative.

The $R_g=T_0/2T_p$ parameter increases with a shortening of the rise time T_p. A large R_g and a small R_k thus produce a small opening quotient, $OQ=(1+R_k)/2R_g$. An alternative common definition is $OQ=(1+R_k)/(2R_g)+R_a$.

Typical values for male vowels are F_a=700 Hz, R_k=0.35, R_g=1.20 and for female vowels F_a=500 Hz, R_k=0.45 and R_g=1. An increase of R_k or a decrease of F_a as in breathy phonation will produce an increase of U_o and thus of the voice fundamental H_1 at constant E_e. An increase of R_g at constant E_e will increase the relative level of the second harmonic of the source spectrum at the expense of a lowered U_o and produces a decrease of OQ which is typical of pressed phonation. A sonorous voice has a relative high F_a of the order of 2000 Hz.

The R_d-parameter

A statistical and functional analysis of covariation of LF-parameters ranging from an extreme tight adducted phonation with low OQ and high F_a to a very breathy abducted phonation with high OQ and low F_a brings out characteristic trends. These can be quantified along a single shape parameter R_d which is closely related to the effective pulse decay time $T_d=U_o/E_e$ (in ms) of the falling branch, see Figure 3.

$$R_d=(U_o/E_e)(F_0/110) \quad (3)$$

T_d is of the order of 0.5-1 ms for both male and female vowels.

Within a population of vowels and voiced consonants we find a statistical relation:[9]

$$R_a=(-1+4.8R_d)/100 \quad (4)$$

and

$$R_k=(22.4+11.8R_d)/100 \quad (5)$$

An important additional finding is that R_d can be estimated from the geometrical constraints of the LF model given the set of R_a,R_k,R_g

$$R_d=(1/0.11)(0.5+1.2R_k)(R_k/4R_g+R_a) \quad (6)$$

R_g can be derived statistically in the same way as R_a and R_k [9], but a better approach is to calculate R_g from R_d given R_a and R_k. This ensures a conformity with the LF model.

An interesting finding [9] based on female vowel data supplied by Karlsson [20] is that given her full specification of the R_a, R_k and R_g values of a set of nine Swedish vowels these can be predicted with considerable accuracy from R_d alone. This involves the process of first condensing R_a, R_k and R_g values into a single R_d parameter (6) and then applying (4-6).

A conclusion is thus that essentials of the glottal source wave shape may be contained into a single default parameter, $R_d=(U_o/E_e)(F_0/110)$ which is relatively easily accessible from a primitive inverse filtering which has special merits for tracking temporal variations in connected speech. However, for more detailed analysis we need the full set of LF parameter from a proper inverse filtering. Deviations of these from default predicted values can be specified in terms of coefficients $k_a=R_a/R_{ap}$, $k_g=R_g/R_{gp}$ and $k_k=R_k/R_{kp}$ for extra aspiration, press, or flow respectively.

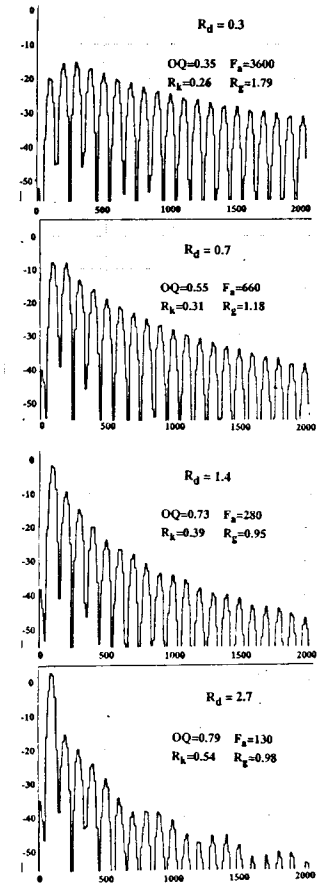


Figure 4. Glottal flow derivative spectra in the frame of R_d values with default LF-parameters included

Glottal flow derivative spectra assuming a constant E_e and F_0=100Hz are shown in Figure 4. The four samples of R_d =0.3, R_d=0.7, R_d=1.4 and R_d=2.7 illustrate the variation from a medially compressed phonation with a small open quotient and a high F_a to a highly abducted, phonation with a large open quotient and a low F_a. Observe the large variation in the ratio H_1/H_2 of voice fundamental to second harmonic amplitude. Females show higher R_d and

H_1/H_2 values than males, and voiced consonants, and aspirated vowels show higher R_d and H_1/H_2 values than regular vowels which is in agreement with earlier findings [11-12].

VOICE SOURCE DYNAMICS

Studies of voice source dynamics, i.e. of temporal variations in connected speech is a developing area which has not yet received the same attention as stationary voice qualities. There remains much to learn about the coordination of glottal adjustments with intonation and lung pressure within a phonetic-linguistic frame.

From our recent work we find systematic covariations of E_e and U_o with F_0. These occur both in glissando sustained phonations of a vowel [15] and in connected speech [9, 13] Both U_o and E_e increase with F_0 up to a maximum or a plateau which is located in the speaker's mid frequency F_0 range, somewhat higher for E_e than for U_o, and E_e increasing more steeply than U_o. Statistical data sampled from prose reading have shown a location of the E_e maximum at around F_0= 100-130 Hz for two males and at F_0=215 Hz for a female voice [13]. In a neutral intonation without focal accents we see clear tendencies of the E_e contours following the general pattern of the F_0 contour. An exception is when in focal accentuation F_0 overshoots a critical value of maximum E_e in which case the temporal contour of E_e and intensity may show a minimum at the F_0 peak with local maxima on both sides. The minimum is not always present. It will be flattened out under the influence of a subglottal pressure rise. Lung pressure is known to increase with F_0 in singing but in speech F_0 operates largely independent of pressure.

INTENSITY VARIATIONS

An important physiological parameter in voice production is the time varying glottal area. $A_g(t)$. At one and the same lung pressure the E_e and the intensity (SPL) increases with A_{gmax} [7, 15] which is capitalized by trained singers [27]. For a more complete understanding of prosodic phenomena we need more data on how A_{gmax} [28] and subglotttal pressure [29] covary with supraglottal and glottal articulations, F_0, SPL, E_e and source spectral shape parameters. Increased lung pressure, and thus subglottal pressure, is found in contrastive and higher degrees of stress but is probably not a necessary component of focal accentuation.

It has long been known that increasing voice effort is associated with a relative emphasis at higher frequencies In an early study [21-22] it was found that a 10 dB increase in the F_1 region was accompanied by about 4 dB increase in the voice fundamental and 14-18 dB in the F_2-F_3 region. This spectral nonlinearity can be interpreted as R_d and R_a decreasing (F_a increasing) with voice effort. Local increments of this magnitude are seldom encountered in speech [26]. The average intensity difference between stressed and unstressed syllables is about 2 dB only and 3 dB with high frequency preemphasis. Twice these values are normally encountered in contrastive stress marking. The intensity parameter has a greater importance as a boundary marker and shows temporal variations similar to those of an accompanying F_0 declination within a phrase.

ACKNOWLEDGEMENTS

This research has been supported by a grant from The Bank of Sweden Tercentenary Foundation and by Telia Promotor Infovox AB.

REFERENCES

[1] Fant, G., (1993), "Some problems in voice source analysis," *Speech Communication* 13, pp. 7-22, 1993.

[2] Fant, G. (1979), "Glottal source and excitation analysis". *STL-QPSR* 1/1979: 85-107

[3] Ananthapadmanabha, T.V. (1984), "Acoustic ananlysis of voice source dynamics", *STL-QPSR* 2-3/1984, pp.1-24.

[4] Fant, G., Liljencrants, J. & Lin, Q. (1985), "A four-parameter model of glottal flow", *STL-QPSR* 4/1985, pp. 1-13.

[5]. Fant, G., (1982), "Preliminaries to the analysis of the human voice source," *STL-QPSR* 4/1982, pp. 1-27.

[6] Fant, G. (1986), "Glottal flow: models and interaction," *J of Phonetics*, 14 Nos (3/4), pp.393-399.

[7] Lin, Q. (1990), "*Speech Production Theory and Articulatory Speech Synthesis*", Ph.D. Thesis, Dept. Speech Com. and Music Acoust., KTH, Stockholm.

8] Fant, G., & Lin,Q. (1988), "Frequency domain interpretation and derivation of glottal flow parameters", *STL-QPSR* 2-3/1988, pp. 1-21.

[9] Fant, G., Kruckenberg, A., Liljencrants J. & Båvegård, M. (1994), "Voice source parameters in continuous speech. Transformation of LF-parameters", *ICSLP-94*, Yokohama.

[10] 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.

[11] Stevens, K.N. & Hanson, M. (1994), "Classification of Glottal Vibration from Acoustic Measurements", in Eds. Osamu Fujimura and Minoru Hirano, *Vocal Fold Physiology 1994*, Singular Publ. Group. pp.147-170.

[12] Klatt, D., & Klatt, L. (1990), "Analysis, synthesis and perception of voice quality variations among female and male talkers", *J. Acoust. Soc. Am.* 87, pp. 820-857.

[13] 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.

[14] Fant, G., & Kruckenberg, A. (1995): "The Voice Source in Prosody", ICPhS 95

[15] Fant, G. (1982) "The voice source. Acoustical modelling",STL-QPSR 4/1982, pp.28-48.

[16] Karlsson, I. & Neovius, L., (1993), "Speech synthesis experimens wit the GLOVE synthesiser, *Eurospeech 93*, pp.925-928.

[17] Cranen, B. and Schroeter, J. (1993), "Modelling a leaky glottis". *Proc. Dept. of Language and Speech* 16/17, University of Nijmegen. pp 56-64.

[18] Båvegård M., Fant G. (1995), "Interactive voice source modelling". ICPhS-95.

[19] Fant, G., Fintoft, K., Liljencrants, J., Lindblom, B. & Martony, J. (1963), "Formant amplitude measuremets", *J. Acoust. Soc. Am.* 35, pp.1753-1761.

[20] Karlsson, I. (1990) "Voice source dynamics for female speakers,"*Proc. I.C.S.L.P. Kobe*, pp. 69-72, 1990.

[21] Fant, G. (1959), "Acoustic Analysis and Synthesis of Speech with Applications to Swedish", Ericsson Technics No. 1 /1959)

[22] Fant, G. (1980), "Voice source dynamics", *STL-QPSR* 2-3/1980, pp.17-37.[

[23] Schoentgen, J.(1995), "Dynamic models of the glottal pulse", *Levels of Speech Communication: Relations and Interactions*, C. Sorin et al. (Eds) Elsevier Science, B.V. pp.249-266.

[24] Gobl. C. (1988), "Voice source dynamics in connected speech," *STL-QPSR* 1/1988, pp. 123-159.

[25] Gobl, C. "& Ní Chasaide, A. (1988), "The effects of adjacent voiced/voiceless consonants on the vowel voice source: a cross language study", *STL-QPSR* 2-3 1988, pp 23-59.

[26]. Stevens, K.N (1994), "Prosodic influences on glottal waveform: Preliminary data", *Int. Symp. on Prosody*, Sept. 18 1994, Yokohama, pp. 53-63.

[27] Sundberg, J., Titze, I., & Sherer, R. (1993), "Phonatary Control in Male Singing: A Study of the Effects of Subglottal Pressure, Fundamental Frequency, and Mode of Phonation on the Voice Source", *Journal of Voice*, Vol. 7, No. 1, pp. 15-29.

[28] Sundberg, J. (1994) "Vocal fold vibration patterns and phonatory modes", *STL-QPSR* 2-3/1994, pp.69-80.

[29] Sundberg, J., Elliot, N., Gramming,, P., & Nord, L. (1993): "Short-Term Variation of Subglottal Pressure for Expressive Purposes in Singing and Stage Speech: A Preliminary Investigation", *Journal of Voice*, Vol.7, No. 3, pp. 227-234.

HOW EMOTION IS EXPRESSED IN SPEECH AND SINGING

Klaus R. Scherer
University of Geneva

ABSTRACT

This contribution focusses on some of the major issues in the study of emotional expression in the speaking and the singing voice. Adopting a sociopsychobiological approach, it is claimed that affect vocalizations have multiple determinants and serve multiple functions. Based on research examples, it is demonstrated how these determinants can be empirically distinguished. Furthermore, recent data on emotion differentiation via acoustical profiles is presented. Brief allusions are made concerning the appeal and the symbol functions of affect vocalization. Finally, an approach to study emotional expression in the singing voice is presented.

INTRODUCTION

In his influential manual of rhetorics, the *Orator*, Cicero remarks: "There are as many movements of the voice as there are movements of the soul, and the soul is strongly affected by the voice". While the terminology is no longer fashionable, this brief statement sums up much of what current research on the vocal expression of emotion in speech and music is empirically documenting. In this contribution, I will present some of the work of our research group, both with respect to theory and data.

MULTIPLE FUNCTIONS AND MULTIPLE DETERMINANTS

While it has been customary to consider vocal expression in animals primarily as indicative of underlying affective or motivational states [1], recent research indicates that the situation is more complex. Marler and his colleagues, studying the alarm calls of vervet monkeys, found that calls are not only indicative of the emitter's fear state but are also specific to certain types of predators [2]. Alarm calls produced for leopards, eagles or snakes, for example, have different sounds, energy levels and frequency ranges. Therefore, Marler and his colleagues reject the notion that animal communication is limited to indicating the animal's emotional or motivational state, arguing that most animal calls have a very strong referential symbolic component. Supporting this notion is the observation that alarm calls seem to be partially learned. Therefore, the alarm call system does not simply "push out" the underlying affect but reflects the outcome of predator classification, which reflects rudimentary cognitive processes.

An exclusive emphasis on either a motivation-affect expression or a symbolic function, neglects the fact that most vocal signals are multifunctional. The Organon model, developed by Bühler [3], can be used to analyse the functions of vocal affect signals. In this model, a sign has three functions: as a *symbol* in representing the object, event or fact it stands for, as a *symptom* of the state of the sign user and as an *appeal*, or signal, trying to elicit a response from the receiver. The vervet monkey alarm call, for example, serves all three: a) as a symbol of different predators, b) as a symptom of the fear state of the animal, and c) as an appeal to others to run away. Furthermore, these functions are mutually interdependent: if a call refers to an air predator, both the emotional reaction and the appeal may be very different from one made in reference to a ground predator - in the first case, both emitter and receiver might seek shelter under a bush and freeze; in the second, they might become highly activated and run up a tree.

Apart from multiple functions we also need to consider multiple determinants. We have suggested to distinguish between *push effects* in which physiological processes such as muscle tone push vocalisations in a certain direction and *pull effects* where external factors such as the expectations of the listener pull the affect vocalisation toward a particular acoustic model [1]. In the push effect, given that muscle tone is likely to be higher in sympathetic arousal, the fundamental frequency of the voice (F0) will also be higher. Pull effects, on the other hand, are governed by social conventions such as display rules. These cultural conventions influence the production of signs required in social situations by specifying a particular acoustic target pattern, as opposed to mental concepts or internal physiological processes which push out expression. This distinction is important in understanding the differences between vocal productions.

Thus, push factors are defined as producing changes in subsystem states in the organism which have a direct effect on vocalisation parameters. They work largely involuntarily; the effects on vocal organs and the resulting acoustic parameters are almost exclusively determined by the nature and force of physiological mechanisms. Pull factors on the other hand, although they are mediated through internal systems, are externally based - they operate toward the production of specific acoustic patterns or models, as in the case of detailed optimum signal transmission features or socially defined signal values.

Clearly, the push/pull concept of two major types of determinants of vocal signals is directly linked to the Bühler model of multiple functions. One can argue that the symptom aspect, i.e. the expression of an internal state, represents push, whereas the symbol and appeal aspects represent pull. Different factors might determine the nature of the expression in each case. And, it might be the antagonism between push and pull, e.g. high physiological arousal pushing voice fundamental frequency up and the conscious attempt to show "control" pulling it down, which can produce mixed or even contradictory messages. If this were so, it would be important to empirically isolate the two determinants. Future research and theorizing in this area will need to more clearly differentiate between these multiple determinants and multiple functions in order to avoid futile controversies about the "true nature" of affect vocalizations..

EMPIRICAL ASSESSMENT OF MULTIPLE DETERMINANTS

We argue that the type of determinant will have a major effect on the *coding*, i.e. the relationship between the underlying referent and the sign features. If, in a push condition, muscle tension goes up under stress, producing an increase in the fundamental frequency of the voice, we would expect direct *covariation* between the amount of muscle tension increase measured by electromyography and the increase in fundamental frequency as measured by digital voice analysis. In this convariation model, we would expect a continuous (and probably linear) covariation between the two variable classes. The alternative is what we call the *configuration* model [4]. The configuration model is more "linguistic" than the covariation model, itself more psychological in nature. The configuration model argues that to achieve a certain effect in the listener, one uses a particular combination of intonation, accent, word and/or syntactic structure, e.g. a rising intonation contour in a WH-question, a falling one for in a Y/N question. There are no variable dimensions, no continua, in configuration effects: certain classes of phenomena have to co-occur to produce an effect. In terms of push and pull, push is likely to follow covariance rules, while pull, if anything, would follow configuration rules. In trying to understand how communication processes follow either a covariation or a configuration model, we need to gather insight into the determinants.

Scherer, Ladd and Silverman conducted two studies to distinguish covariation and configuration. The first [4] used a corpus of questions (from a large scale study on interactions between civil servants and other citizens) that were homogenous in structure, but that varied in terms of pragmatic force. Some questions were clearly reproaches, though phrased as WH questions, while others were factual information questions. We used three filtering or degrading techniques to systematically isolate particular acoustic cues: a) low pass filtering, b) random splicing, and c) reversing. (see [5] for a comparative list of the acoustic cues retained by the respective techniques).

The results show that even when the text of the questions used is rendered unintelligible, much of the affective meaning remains in the acoustive signal. This confirms the covariance model in its claim that nonverbal vocal cues convey affect in a direct and context-independent way. However, this is true

only for those masking conditions in which voice quality cues are audible, i.e. in the random splicing and reversing conditions. In both of these conditions, the intonation contour of the sentence is lost or destroyed. This would seem to imply that this feature of speech utterances plays no major role in the communication of affective meaning. Obviously, this is rather counter-intuitive and contradicts empirical evidence showing that this information is relevant for the communication of affect. Could it be that intonation follows configuration rather than covariance rules? In order to investigate this question with the speech material in this study, we divided the questions into WH and Yes/No questions and classified the intonation contours into final fall and final rise [4]. The results show that intonation contour obviously has a strong effect on the impression of speaker affect but it seems to be mediated by context effects, verbal or syntactic.

One might interpret these results as reflecting the traditional descriptions of "normal" or "unmarked" intonation for the two different question types. The supposedly "normal" combinations of intonation type (i.e. falling WH questions and rising yes/no questions) were judged as more polite and agreeable. "Marked" combinations on the other hand were rated rather more negatively. This clearly points to strong configuration effects. It is possible then, to presume that some acoustic cues, such as voice quality, operate according to covariance rules, whereas others, such as intonation contours, are used in accordance with configuration rules. This would make sense in terms of a psychobiological approach to communication. One could argue that those cues that show a remarkable degree of phylogenetic continuity - such as the differential nature of phonation which yields different voice qualities - are closer to direct covariance with physiological states. In contrast, cues that have been domesticated within a language system, such as intonation, should follow a configuration model.

In order to further test these notions, we used digital resynthesis of speech in order to be able to experimentally vary different acoustic cues in a factorial design. Such an approach obviously avoids the disadvantage of using a natural corpus, since it allows greater experimental control of the variables under study. In a series of studies [6,7], we used this technique to systematically vary intonation contour, F0 range, intensity, timing, accent, structure, and other parameters. The advantage of this procedure, as pointed out above, is that all of these acoustic features under study can be manipulated independently of each other in a factorial design while leaving all of the remaining acoustic cues constant. Three major types of findings will be highlighted. First, we did not find any interaction effects in the analysis of variance, suggesting that the acoustic variables we studied function largely independently of each other. Secondly, in those studies where we used several speakers and several utterances, we found virtually no interaction between these factors and the acoustic variables manipulated. This encourages one to think that the effects can be generalised over a wide range of speakers and utterances. Thirdly, out of the variables studied, F0 range had the most powerful effect by far on the judgment of the raters, particularly on the attributions of arousal. Furthermore, we were able to show that these effects seemed to be a continous function of changes in F0 range since arousal related ratings go up in a linear fashion with increasing range.

Results for intonation contours and voice quality were complex and seem to require further study. In the case of intonation contours, this may well be due to the important role of the configuration model for this variable. In consequence, we feel that the distinction between configuration and covariance rules may be very useful in understanding the communication of affect in vocal utterances and it would seem useful to continue this type of research with the aid of modern digital signal manipulation techniques.

ACOUSTIC EMOTION PROFILES

The research reported above dealt with affective states of relatively low intensity as one is likely to encounter in normal social interactions. Full-blown, intensive emotions are difficult, if not impossible, to study in an experimental fashion. Therefore, much of the work on the acoustic concomitants of emotion has used actor portrayals of different emotional states to obtain vocal expression samples that could then be analyzed acoustically. Pittam & Scherer [8] have summarized the state of the literature to date as follows:

Anger: Anger generally seems to be characterized by an increase in mean F0 and mean energy. Some studies, which may have been measuring "hot" anger (most studies do not explicitly define whether they studied hot or cold anger), also show increases in F0 variability and in the range of F0 across the utterances encoded. Studies in which these characteristics were not found may have been measuring cold anger. Further anger effects include increases in high frequency energy and downward directed F0 contours. The rate of articulation usually increases.

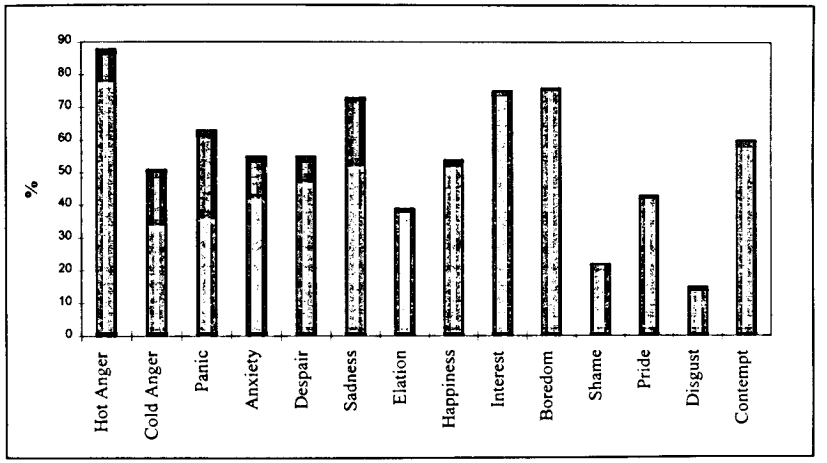


Figure 2. Accuracy of emotion recognition in vocal actor portrayals (based on data from[10]). The shaded parts of the bars represent confusions between members of the same emotion family.

Fear: There is considerable agreement on the acoustic cues associated with fear. High arousal levels would be expected with this emotion, and this is supported by evidence showing increases in mean F0, in F0 range, and high frequency energy. Rate of articulation is reported to be higher. An increase in mean F0 has also been found for milder forms of the emotion such as worry or anxiety.

Sadness: As with fear, the findings converge across the studies that have included this emotion. A decrease in mean F0, F0 range, and mean energy is usually found, as are downward directed F0 contours. There is evidence that high frequency energy and rate of articulation decrease. Most studies have investigated the quieter, subdued forms of this emotion rather than the more highly aroused forms such as desperation. The latter variant might be characterized by an increase of F0 and energy.

Joy: This is one of the few positive emotions studied, most often in the form of elation rather than more subdued forms such as enjoyment or happiness. Consistent with the high arousal level that one might expect, we find a strong convergence of findings on increases in mean F0, F0 range, F0 variability and mean energy. There is some evidence for an increase in high frequency energy and rate of articulation.

Disgust: The results for disgust tend to be inconsistent across studies. The few that have included this emotion vary in their encoding procedures from measuring disgust (or possibly displeasure) at unpleasant films to actor simulation of the emotion. The studies using

the former found an increase in mean F0, whereas those using the latter found the reverse - a lowering of mean F0. This inconsistency is echoed in the decoding literature.

Even though these results seem to indicate a rather clear acoustic differentiation of the major basic emotions, it cannot be excluded that many of the differences are due to a simple arousal factor - high sympathetic arousal, which is typical for several emotions, driving up F0, energy, and high-frequency spectral energy. The issue of whether vocal expression only indexes sympathetic arousal, rather than qualitative emotion differences (as found in prototypical facial expressions) has been one of the major concerns in this area [9].

A study recently conducted by our research group allows an advance with respect to this issue. 12 professional actors were asked to portray 14 emotions varying in intensity and valence or quality [10]. A total of 224 different portrayals, 16 per emotion category, were presented to judges who were asked to decode or infer the emotion category intended by the sender or encoder. The results on decoding replicate and extend earlier findings demonstrating the ability of judges to infer vocally expressed emotions with much better than chance accuracy for a large number of emotions. Figure 1 presents the differences in recognition accuracy across the 14 emotions. Consistently found differences in the recognizability of different emotions are also replicated.

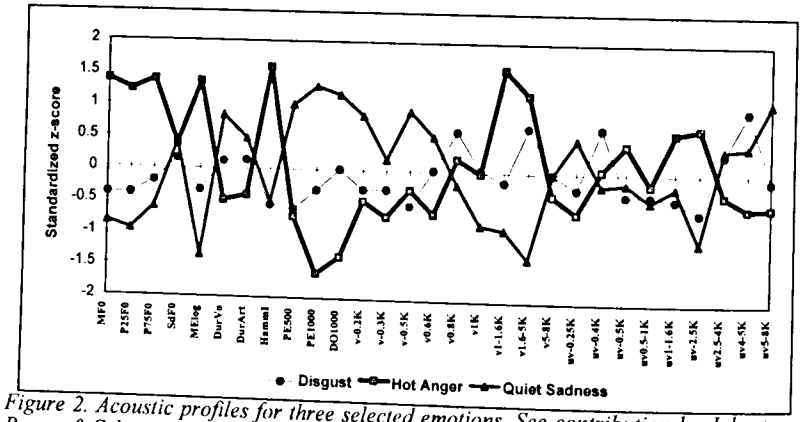


Figure 2. Acoustic profiles for three selected emotions. See contribution by Johnstone, Banse, & Scherer, this volume, for the legend of acoustic variables.

All 224 portrayals were subjected to digital acoustic analysis to obtain profiles of vocal parameters for different emotions, using a large set of acoustic variables. Figure 2 presents the acoustic profiles for some of the most interesting emotions. The data provide first indications that vocal parameters not only index the degree of intensity typical for different emotions but also differentiate valence or quality aspects. They also suggest that with further refinement in acoustic measurement it might be possible to determine stable acoustic profiles for most emotions - provided that appropriately differentiated emotional states are used [9]. Discriminant analysis and jack-knifing were used to determine how well the 14 emotions can be differentiated on the basis of the vocal parameters measured. The results show remarkably high hit rates and patterns of confusion that closely mirror those found for listener-judges. It could also be shown that much of the variance in the judges' inferences could be predicted on the basis of the acoustic measurements, allowing a more detailed assessment of the acoustic cues that listeners use in inferring emotion from the voice.

APPEAL FUNCTIONS OF AFFECT VOCALIZATION

Research on the signalling or appeal aspect of vocal affect expression is particularly underdeveloped. However, it is possible to make a case for an important appeal function of vocal emotion expression in social influence settings, particularly persuasion. For example, it can be argued that appropriate emotional expression by a persuader will tend to increase the effectiveness of the persuasive message because of a) the attribution of greater credibility and trustworthiness to the sender, and b) the production of appropriate emotions in the audience which may induce the desired attitudes or behaviors or make the cognitive processing more amenable to accepting the message emitted by the persuader [11].

SYMBOLIC, REFERENTIAL FUNCTIONS OF VOCAL AFFECT EXPRESSION

I ventured a rather speculative proposal on how one might conceive of the symbolic function of vocal affect signs by arguing that the acoustic charateristics of an emotional vocalisation reflect the complete pattern of the cognitive appraisal process that produced the emotional state in the sender. This information about the criteria used in the emotion-antecedent evaluation should allow the listener to reconstruct the major features of the emotion producing event and its effect on the speaker [12]. In order to explain this postulate I have to expose some recent theorising on emotion. Many theorists in the field of psychology of emotion seem convinced that most human emotions are preceded by cognitive evaluation of events and situations (although the type of cognitive process can be relatively low level, automatic and unconscious). If this is the case, then knowing an organism's emotional state should allow us to infer the emotion eliciting cognitive processes, and thus, the approximate nature of the emotion eliciting event. If listeners are able to identify a particular emotional state of the sender from the acoustic features of the vocalisation, thus inferring the nature of the emotion producing event, then one might claim a symbolic function for emotional vocalisations. One could go even further. We are not only able to identify emotional states on the basis of acoustic cues, we may even have direct access to the results of the cognitive appraisals that have produced a particular emotional state. It is possible to elaborate predictions on how we would expect the major phonation characteristics to vary as a result of the major emotion antecedent evaluation criteria [1,9]. (The data from the actor portrayal study reported above were used to test these theoretical predictions on vocal patterning based on the component process model of emotion. While most hypotheses are supported, some need to be revised on the basis of the empirical evidence.)

If this line of reasoning is correct, one might conclude that by appropriate inferences from particular acoustic cues, receivers should be able to judge not only the nature of the emotional state of the speaker but also, and maybe even more directly, the outcomes of the pattern of cognitive appraisals which have produced the respective emotional states. In consequence, listeners should also be able to infer the approximate nature of the emotion producing event or situation as well as information about the speaker's ego involvement and coping potential. If this were the case, and if this effect were to be powerful enough to transcend individual ideosyncracies and the influence of contextual clues, then one would be justified in claiming a symbolic representational function for nonverbal vocal affect expression.

EMOTIONAL EXPRESSION IN SINGING

One can argue that emotion vocalizations might be at the root of all of human speech and singing [13]. It is not surprising, then, that much of what has been said above about multiple functions and multiple determinants is also true for singing. The acoustic signal produced by a singer reflects his or her emotional state, produces affect in the listeners, and often symbolizes abstract notions about emotionality (as shown.for example, in the *Affectenlehre* of Baroque opera). Reviews of the literature on all three of these aspects can be found in [14,15].

Unfortunately, empirical work is scarce in this area. I will conclude this contribution with an illustration of a recent study of our group on emotional expression in operatic singing [16]. Two

excerpts from the cadenza in Ardi gli incensi from Donizetti's opera Lucia di Lammermoor were acoustically analyzed for five recorded versions of the air by Toti dal Monte, Maria Callas, Renata Scotti, Joan Sutherland, and Edita Gruberova. The measured acoustic parameters of the singing voices were correlated with preference and emotional expression judgments, based on pairwise comparisons, made by a group of experienced listener-judges. In addition to showing major differences in the voice quality of the five dive studied, the acoustic parameters permit one to determine which vocal cues affect listener judgments. Furthermore, two component scores, based on a factorial-dimensional analysis of the acoustic parameters, allow the prediction of 84% of the variance in the preference ratings. Thus, we were able to show 1) that the different interpretations elicited significantly different listener ratings of emotional expressiveness, 2) that the voice samples of the five singers differ quite substantially with respect to objective acoustic variables, and 3) that we can quite successfully predict listener attributions on the basis of the objective acoustic characteristics.

REFERENCES

[1] Scherer, K. R. (1985), Vocal affect signalling: A comparative approach. In J. Rosenblatt, C. Beer, M. Busnel, & P. J. B. Slater (Eds.), *Advances in the study of behavior* (pp. 189-244), New York: Academic Press.

[2] Marler, P. (1984), Animal communication: Affect or cognition? In K.R. Scherer & P. Ekman (Eds.), *Approaches to emotion* (pp. 345-368), Hillsdale, N.J.: Erlbaum.

[3] Bühler, K. (1934), *Sprachtheorie*, Jena: Fischer (new edition 1984).

[4] Scherer, K. R., Ladd, D. R., & Silverman, K. E. A. (1984), Vocal cues to speaker affect: Testing two models, *Journal of the Acoustical Society of America*, vol 76, pp. 1346-1356.

[5] Scherer, K.R., Feldstein, S., Bond, R.N., & Rosenthal, R. (1985). Vocal cues to deception: A comparative channel approach. *Journal of Psycholinguistic Research*, vol. 14, pp. 409-425.

[6]Ladd, D. R., Silverman, K. E. A., Tolkmitt, F., Bergmann, G., & Scherer, K. R. (1985), Evidence for the independent function of intonation contour type, voice quality, and F0 range in signaling speaker affect, *Journal of the Acoustical Society of America*, vol. 78, pp. 435-444.

[7] Bergmann, G., Goldbeck, T., & Scherer, K.R. (1988), Emotionale Eindruckswirkung von prosodischen Sprechmerkmalen, *Zeitschrift für experimentelle und angewandte Psychologie*, vol. 35, pp. 167-200.

[8] Pittam, J., & Scherer, K. R. (1993), Vocal expression and communication of emotion. In M. Lewis & J. M. Haviland (Eds.), *Handbook of emotions* (pp. 185-198),. New York: Guilford Press.

[9] Scherer, K. R. (1986), Vocal affect expression: A review and a model for future research, *Psychological Bulletin*, vol. 99, pp. 143-165.

[10] Banse, R., & Scherer, K.R. (in press), Acoustic profiles in vocal emotion expression, *Journal of Personality and Social Psychology*.

[11]Scherer, K. R. (1993), Interpersonal expectations, social influence, and emotion transfer. In P. D. Blanck (Eds.), *Interpersonal expectations: Theory, research, and application* (pp. 316-336), Cambridge and New York: Cambridge University Press.

[12] Scherer, K. R. (1988), On the symbolic functions of vocal affect expression. *Journal of Language and Social Psychology*, vol. 7, pp. 79-100.

[13] Scherer, K.R. (1991), Emotion expression in speech and music. In J. Sundberg, L. Nord, & R. Carlson (Eds.). *Music, Language, Speech, and Brain.* Wenner-Gren International Symposium Series: vol. 59. (pp. 146-157), London: Macmillan.

[14] Scherer, K.R. (in press), Expression of emotion in voice and music. *Journal of Voice*.

[15] Sundberg, J. (1987), *The science of the singing voice*, DeKalb, IL: Northern Illinois University Press.

[16] Siegwart, H. & Scherer, K.R. (in press), Acoustic concomitants of emotional expression in operatic singing: The case of Lucia in Ardi gli incensi, *Journal of Voice*.

SUBGLOTTAL PRESSURE BEHAVIOR IN SINGING AND SPEECH

Johan Sundberg, Dept. of Speech Communication and Music Acoustics, Royal Institute of Technology, Box 70014, S-10044 Stockholm, Sweden

ABSTRACT

Subglottal pressure is determined by muscular forces, elasticity forces, and gravitation and represents the major control parameter for vocal loudness. In neutral speech subglottal pressure is generally rather constant, while in emotive speech it is quite variable. In singing it is varied also with F0. As subglottal pressure affects pitch, singers need to learn a virtuosic breath control to stay in tune.

INTRODUCTION

Lately, several advances have been made in our understanding of subglottal pressures in singing. After the pioneering investigations by Proctor, Mead, and associates [1], summarized in Proctor [2], important contributions on breathing kinematics have been made by Hixon and associates [3]. During the eighties, the author had the privilege of carrying out a series of investigations of singers' breathing together with the neurologist Curt von Euler and the late phoniatrician Rolf Leanderson [4], [5], [6]. Here, the anatomy and physiology of the breathing apparatus will first be reviewed, and then certain characteristics of subglottal pressures in singing and speech will be described.

THE BREATHING APPARATUS

Overpressures of the air below the glottis, henceforth *subglottal pressure*, is produced by decreasing the volume of the rib cage. There are three different forces that influence this pressure: muscular forces, elasticity forces and gravitation. The main muscular forces are exerted by the *intercostal muscles*, the *diaphragm*, and the *abdominal wall muscles*. The intercostals join the ribs. The external intercostals widen the rib cage by lifting the ribs, and so provide an inspiratory muscle force. The internal intercostal muscles decrease the rib cage volume. The diaphragm is an inhalatory muscle inserting into the lower contour of the rib cage. When contracting, it is flattened so that the floor in the rib cage is lowered, and lung volume is increased. With the body in an upright position, the diaphragm muscle can be restored to its upward-bulging shape by means of the abdominal wall muscles. By contracting, these muscles press the abdominal content upward, into the rib cage, so that the diaphragm, the floor in the rib cage, moves upward and the lung volume is decreased. Consequently, the abdominal wall muscles are exhalatory.

The external and internal intercostals represent a paired muscle group producing inspiratory and expiratory forces. The diaphragm and abdominal wall the represent a similar paired muscle group for inhalation and exhalation. In costal breathing, the intercostals are used for respiration, and in ventricular breathing the diaphragm and abdomen are used as respiratory muscles. Mostly a combination of costal and abdominal breathing is used.

The volume of the abdominal content cannot be altered appreciably. Therefore, when the diaphragm contracts, it presses the abdominal content downward which, in turn, presses the abdominal wall outward. If the abdominal wall remains flat during inspiration, this means that only the intercostal muscles were used. An expansion of the abdominal wall during phonation is not necessarily a sign of diaphragmatic activation. It may equally well result from the increased

lung pressure that is required for phonation, because an overpressure in the lungs is transmitted downward through the diaphragm so that the subglottic pressure exerts a pressure on the abdominal wall. By contracting the abdominal wall muscles, this expansion can be avoided.

Apart from these muscular forces, there are also *elasticity* forces. The magnitude of these forces depends on the amount of air contained in the lungs, or the lung volume. The lungs always attempt to shrink, somewhat as rubber balloons, when hanging inside the rib cage. They are prevented from doing so by the fact that they are surrounded by a vacuum. The lungs therefore exert an entirely passive expiratory force which increases with lung volume. This force corresponds to a pressure that may amount to around 20 cm H_2O after a maximum inhalation and after a deep exhalation, it is only a few cm H_2O.

If the rib cage is forced to deviate from its rest volume, e.g., because of a contraction of the intercostal muscles, it strives to return to a smaller volume. Therefore, also the rib cage produces elastic forces. At high lung volume a passive expiratory force is generated that may produce an overpressure of about 10 cm H_2O. Conversely, if the rib cage is squeezed by the expiratory intercostal muscles, it strives to expand again. After a deep costal exhalation, the resulting passive expiratory force may produce an underpressure of about -20 cm H_2O.

Subglottal pressure is affected also by *gravitation*. In an upright position, the abdominal content is pulling the diaphragm downward and hence produces an inhalatory force. In supine position, gravitation strives to move the abdominal content into the rib cage and so produces an exhalatory force.

As the elasticity forces are both exhalatory and inhalatory, depending on lung volume, there is a particular lung volume, at which these passive forces are equal. This lung volume value is called the *functional residual capacity* (FRC). As soon as the lungs are forced to depart from FRC by expanding or contracting, passive forces try to restore the FRC volume.

REGULATION OF SUBGLOTTAL PRESSURE

Above we have seen that subglottic pressure is dependent on the activity in different respiratory muscles plus the lung volume dependent passive elasticity forces, plus the posture dependent influence of gravitation. The muscular activity required for maintaining a constant subglottic pressure is dependent on the lung volume because the elasticity forces of the lungs and the rib cage strive to raise or to lower the pressure inside the lungs, depending on whether the lung volume is greater or smaller than the functional residual capacity, FRC. When the lungs are filled with a large quantity of air, the passive exhalation force is great, and it generates a high pressure. If this pressure is too high for the intended phonation, it can be reduced by a contraction of inhalatory muscles. The

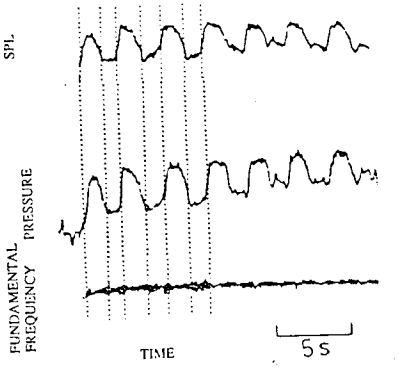


Figure 1. Variation of subglottal pressure during variation of vocal loudness. The top curve shows sound level, the middle curve esophageal pressure, and the bottom curve F0.

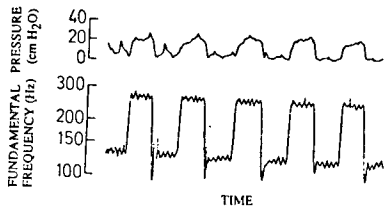


Figure 2. Variation of subglottal pressure with pitch. The graph shows a registration from a professional baritone singer performing a sequence of alternately ascending and descending octave intervals. The top curve shows esophageal pressure captured by a GELTEC pressure transducer, and the bottom curve F0. The wiggles in the esophageal pressure reflect pulse beats.

need for this activity then gradually decreases as the lung volume decreases, reaching zero at the lung volume where the elasticity forces provide the target pressure. Beyond this point the muscles of exhalation must take over more and more, thus compensating for the increasing inhalatory recoil force of the increasingly compressed rib cage.

When we speak, we generally use rather small lung volumes, typically just above FRC [7], [8]. Under these conditions, the elasticity forces are not very strong. In singing, larger portions of the vital capacity are frequently required. Thus long phrases may be initiated at very high lung volumes and end when the lungs nearly depleted [7]. Under these conditions, the elasticity forces are considerable.

Subglottal pressure in singing

Ideally, subglottal pressure is measured by inserting a fine needle into the trachea, obviously a rather intrusive method. However, it can be measured also as the mouth pressure during [p]-occlusion [9], [10], [11].

As mentioned, subglottal pressure is the main physiological parameter for variation of vocal loudness. Figure 1 illustrates this. It shows the sound level and the underlying subglottal pressure in a singer who alternates between *subito forte* and *subito piano* at constant pitch. Sound level and subglottal pressure change quickly and in synchrony between two rather stationary values such that square-wave-like patterns emerge.

In singing, variation of subglottal pressure is required not only for loudness variation but also with pitch [12]. When we increase pitch, we stretch the vocal folds. It seems that stretched vocal folds require a higher driving pressure than more lax vocal folds [13]. Figure 2 illustrates this pitch dependence in terms of a recording of a singer performing a series of alternating rising and falling octave intervals. It can be observed that the higher pitch was produced with a much higher pressure than the lower pitch. The wrinkles in the pressure curve represent the singer's heart beats and the undulations in the F0 curve correspond to the vibrato. Figure 3 illustrates the combined dependence of subglottal pressure on loudness and pitch in a singer. As subglottal pressure affects

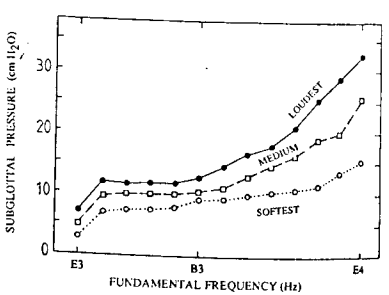


Figure 3. Illustration of the dependence of subglottal pressure on pitch and loudness. Pressures were captured as the oral pressure during [p]-occlusion for the tones in ascending chromatic scales sung at low, middle and high vocal loudness by a professional tenor. From Cleveland & Sundberg, [12].

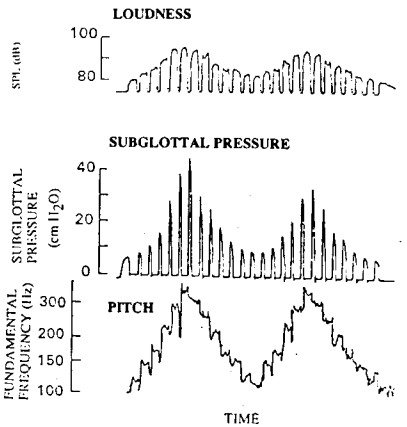


Figure 4. Pitch and loudness dependent variation of subglottal pressure in singing. The top, middle and bottom curves represent sound level, oral pressure during [p]-occlusion, and F0 in a professional baritone singer performing an exercise with an ascending triad on the pitches of a major tonic chord up to the duodecime followed by a descending dominant seventh triad.

pitch, an error in the subglottal pressure is manifested not only as an error in loudness, but also as an error in pitch. Therefore singers must tune their subglottal pressure quite accurately. Accordingly, one finds very well-formed subglottal pressure patterns in proficient singers. Figure 4 illustrates this. It shows the pressures produced by a baritone singing an ascending triad on the tonic chord and a descending triad on the dominant seventh chord. Note that the singer did not give the top pitch the highest pressure. Instead, the peak pressure is given to the first note after the top note. At this note the new dominant chord appears which would represent the musical peak of this phrase. Consequently, the singer gives this note the main stress [14].

The skill required for an accurate reproduction of this exercise is obviously very high, and it is even greater if the tones are sung staccato rather than legato. In staccato, the vocal folds must open the glottis during the silent segments. For this to be possible without wasting air, subglottal pressure must be reduced to zero during the silent intervals between the tones. As a consequence, the singer has to switch from the target value, that was required for the pitch, down to zero during the silent interval, and then up to the new target value which is different from the previous one. A failure to reach the target pressures is manifested as a pitch error. This pitch error becomes quite substantial in loud singing, particularly at high pitches. From the point of view of breath and pitch control, this exercise is clearly virtuosic.

Subglottal pressure in speech

Subglottal pressure during speech has been studied in several investigations (for an excellent overview, see Ohala, 1990 [15]). Earlier it was believed that in speech each syllable was produced with a subglottal pressure peak. However, this was not confirmed in later investigations. Rather, subglottal pressure has been found to be rather smooth and constant, at least in neutral speech. Occasional peaks occur but are presumably caused by downstream variations in flow resistance, e. g., during consonant production [15]. In emphatic or emotive speech, on the other hand, subglottal pressure peaks are frequently observed. An example comparing the same subject's neutral and emphatic speech is shown in Figure 5. In neutral speech, void of emphatic stress, it seems sufficient to signal stress by F0 gestures and syllable duration while in emphatic and emotional speech also subglottal pressure is recruited.

In normal speech, changes in overall vocal loudness are generally associated with shifts in overall F0; the louder the speech, the higher the mean F0. This

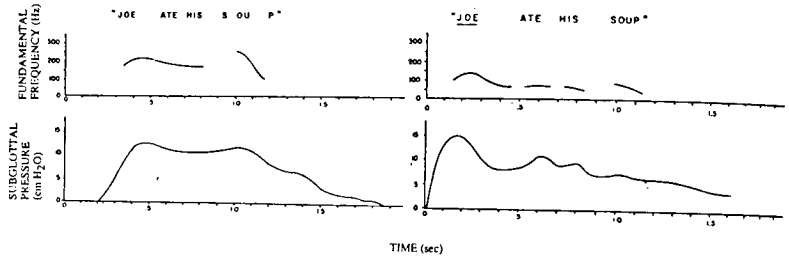


Figure 5. F0 and subglottal pressure (upper and lower curves) during neutral and emphatic speech. The underlined words were emphasized. Emphasis is realized by increase of subglottal pressure. After Lieberman [16].

covariation has been analyzed, revealing high degrees of correlation [17]. The average growth was found to be about 0.4 semitones per dB increase of equivalent sound level. From a measured sound level change, it is possible to roughly estimate the underlying increase in subglottal pressure; a doubling of subglottal pressure leads to an increase in sound level of approximately 9 dB [18], [13]. The effect of a subglottal pressure increase on pitch can also be estimated: on the average, a 1 cm H_2O rise in subglottal pressure results in a F0 increase of about 4 Hz [19], [20]. According to Gramming & al. [17] the postulated pressure increase could indeed explain all of the increase in F0. Thus, the

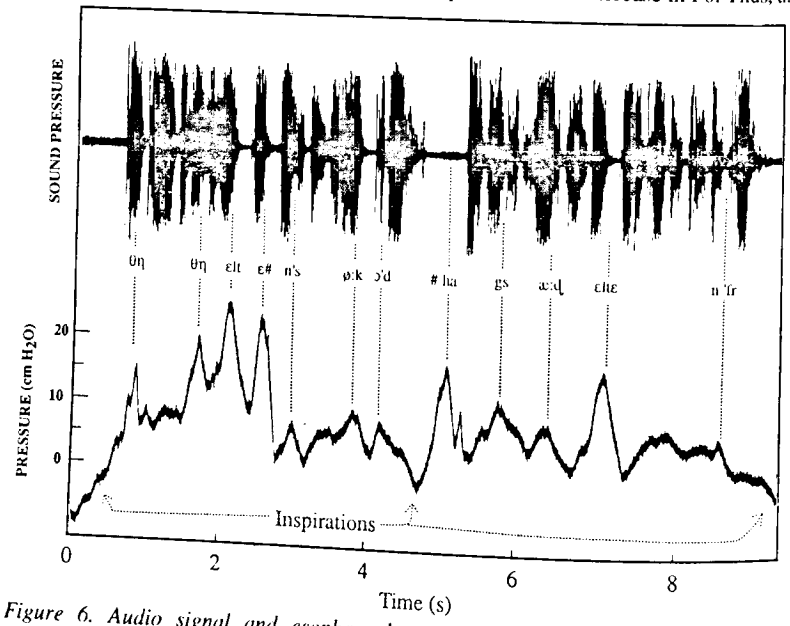


Figure 6. Audio signal and esophageal pressure (upper and lower curve) of a professional actor reading a poem in Swedish as during a stage performance with an audience. The sounds produced during the pressure peaks are shown between the curves. From Sundberg & al [21].

subjects did not seem to bother to use the pitch-raising musculature in order to raise their voice pitch in loud reading. Instead, the pitch just increased passively because subglottal pressure was raised.

Figure 6 shows a graph from an investigation of a professional actor [21]. The esophageal pressure was measured while the subject read a poem as during a theatre play. Several short subglottal pressure peaks can be seen, unexpectedly appearing during the production of voiced consonants. It seems difficult to explain these high pressure peaks as the results of downstream variations of airflow resistance. A more plausible explanation is that they were deliberately produced with the purpose to increase the audibility of the consonants. Unfortunately, only one actor was studied.

CONCLUSIONS

There are great differences in the subglottal pressure behavior in speech and singing. In neutral speech, subglottal pressure is used mainly for control of the overall vocal loudness and is thus basically constant. In singing, subglottal pressure is tailored with regard to both pitch and loudness and must therefore be varied within wide limits. As pitch may change at intervals of 200 ms or shorter in singing, subglottal pressure must be changed quickly. Furthermore, as a change in subglottal pressure affects F0, singers also need to match the target subglottal pressures quite accurately. In speech, loudness and pitch are typically interdependent, so that a rise in loudness is associated with a rise in mean F0. In neutral speech a narrow range of lung volumes just above FRC is used and hence the elasticity forces contributing to subglottal pressure are moderate. In singing these forces represent an important factor, since wide ranges of lung volumes are used. Thus, the demands raised on the breathing apparatus are much higher in singing than in speech.

REFERENCES

[1] Bouhuys, A., Proctor, D., & Mead, J. (1966): "Kinetic aspects of singing," *J.Appl.Physiol.* **21**, pp. 483-496.

[2] Proctor, D. (1980): *Breathing, Speech and Song*, Springer Verlag, New York.

[3] Hixon, T. (1987): *Respiratory Function in Speech and Song*, Taylor & Francis Ltd., London.

[4] Leanderson, R., Sundberg, J., & von Euler, C. (1987): "Role of the diaphragmatic activity during singing: a study of transdiaphragmatic pressures," *J.Appl.Physiol.* **62**, pp. 259-270.

[5] Sundberg, J., Leanderson, R., & von Euler, C. (1989): "Activity relationship between diaphragm and cricothyroid muscles," *J. Voice* **3**, pp. 225-232.

[6] Sundberg, J. (1987): *The Science of the Singing Voice*, N. Ill. Univ. Press, DeKalb, IL.

[7] Watson, P. & Hixon, T. (1985): "Respiratory kinematics in classical (opera) singers," *J.Speech & Hear.Res.* **28**, pp. 104-122.

[8] Winkworth A, Davis P, Adams R & Ellis, E (in press) "Breathing patterns during spontaneous speech", to be published.

[9] van den Berg, Jw. (1962): "Modern research in experimental phoniatrics," *Fol.Phoniat.* **14**, pp. 81-149.

[10] Rothenberg, M. (1968): "The breath-stream dynamics of simple-reased-plosive production," *Bibliotheca Phonetica* No. 6.

[11] Smitheran, J. & Hixon, T. (1981): "A clinical method for estimating laryngeal airway resistance during vowel production," *J. Speech & Hear. Disorders* **46**, pp. 138-146.

[12] Cleveland, T. & Sundberg, J. (1985): "Acoustic analysis of three male voices of different quality," pp. 143-156 in A. Askenfelt, S. Felicetti, E. Jansson,

& J. Sundberg, eds.: *Proc. Stockholm Music Acoustics Conference (SMAC 83): Vol. I*, Royal Swedish Academy of Music, Publ. No. 46:1, Stockholm.

[13] Titze, I. (1989): "On the relations between subglottal pressure and F0 in phonation," *J.Acoust.Soc.Am.* **85**, pp. 901-906.

[14] Sundberg, J. (1989): "Synthesis of singing by rule," pp. 45-55 and 401-403 in (M. Mathews & J. Pierce (eds.) *Current Directions in Computer Music Research*, System Development Foundation Benchmark Series, The MIT Press, Cambridge, MA:

[15] Ohala J (1990) "Respiratory activity in speech", in W Hardcastle & A Marchal, ed.s: Speech Production and Speech Modelling, Dordrecht, NL: Kluwer Academic Publishers, 23-53.

[16] Lieberman P. (1967) *Intonation, perception, and language.* Cambridge, MA: The MIT Press, Research Monograph No 38.

[17] Gramming, P., Sundberg, J., Ternström, S., Leanderson, R., & Perkins, W.H. (1988): "Relationship between changes in voice pitch and loudness", *J. Voice* **2,** pp. 118-126

[18] Fant G. (1982): "Preliminaries to analysis of the human voice source", *Speech Transmission Laboratory, Quarterly Progress and Status Report.* No. 4, pp. 1-27. Royal Institute of Technology, Stockholm.

[19] Baer, T. (1979): "Reflex activation of laryngeal muscles by sudden induced subglottal pressure changes," *J.Acoust.Soc.Am.* **65**, pp. 1271-1275.

[20] Titze, I. (1991): "Mechanisms underlying the control of F0," pp. 129-138 in (B. Hammarberg & J. Gauffin, eds.) *Vocal Fold Phyusiology. Acoustic, Perceptual, and Physiological Aspects of Voice Mechanisms,* Singular Press, San Diego, CA.

[21] Sundberg J, Elliot N, Gramming P & Nord L (1993) "Short-term variation of subglottal pressure for expressive purposes in singing and stage speech. A preliminary investigation", *Journal of Voice 2*, 227-234.

MODELING THE PERCEPTION OF BIMODAL SPEECH

Dominic W. Massaro and Michael M. Cohen
Program in Experimental Psychology, University of California-Santa Cruz

ABSTRACT

How does visible speech contribute to speech perception? Extant theories and their methodological implementations are evaluated and the process of multimodal integration is discussed. Separate operations of the Fuzzy Logical Model of Perception (FLMP) are clarified and found to be consistent with empirical phenomena. An important property of the FLMP is that multiple representations can be held in parallel. We also discuss appropriate methods for model testing.

THEORIES OF BIMODAL SPEECH PERCEPTION

The occurrence of this symposium attests to the powerful impact that visible speech has been shown to have in face-to-face communication, and the recent interest scientists have shown in the process of multimodal integration. The natural integration of several sources of information from several modalities provides a new challenge for theoretical accounts of speech perception. Although it is potentially dangerous to interpret how extant theories are impacted by the positive role of visible speech, we see a negative impact for several of them.

One class of theory seems to be either contradicted or at least placed outside the domain of bimodal speech perception. Psychoacoustic accounts [1] of speech perception are grounded in the idea that speech is nothing more than a complex auditory signal, and its processing can be understood by the psychophysics of complex sounds, without any reference to language specific processes. The strict form of this view is no longer sufficient because speech perception is not a simple function of the auditory information. In addition to the convincing findings of the influence of higher-order linguistic context in speech perception, there is the overwhelming evidence on the important influence of visible speech from a speaker's (or even an animated character's) face. It turns out that the psychoacoustic account even fails in the arena of auditory speech perception [2].

Three other theories have survived or even basked in the findings of audible-visible speech perception. The Gibsonian theory [3] states that persons perceive the cause of the sensory input directly. In spoken language, the cause of audible-visible speech is the vocal-tract activity of the talker. Accordingly, it is reasoned that visible speech should influence speech perception because it also represents the vocal-tract activity of the talker. Furthermore, by this account, vocal tract activity can be picked up directly from touching the speaker's mouth [3] which was found to influence the perceiver's interpretation of the auditory speech presented at the same time. Fowler and Dekle [3] interpret their results as evidence against the FLMP because there would be no haptic information available in the prototype descriptions. Normal perceivers supposedly have not experienced directly haptic information, nor have they experienced acoustic and haptic information about speech occurring together. Accordingly, there would be experience that would allow the development of the appropriate prototype descriptions. However, it is only natural to relate experience along one modality to experience along others.

As speakers and perceivers of language, we can easily describe what global haptic differences would be between /ba/ and /da/, for example. It should not be surprising if perceivers are influenced by haptic information when haptic and acoustic information are presented jointly in a speech identification experiment. As noted by Brunswik and demonstrated in many experiments, perceivers have difficulty selectively attending to just a single dimension of the stimulus input. Independently of the intentions of the observer, he or she tends to integrate multiple sources of information. Given the compatibility of the results with the FLMP, the results do not unambiguously support the idea that it is the events of

the articulatory tract that are perceived. If an uppercase letter is drawn on a person's back, it can be recognized even though the person has never experienced this event previously. Its accurate recognition does not mean that reading letters involves direct perception of the handwriting movements that produced the letter. Rosenblum and Fowler (1991) also state that the FLMP cannot predict a contribution of visual effort on perceived loudness. They state that the model does not have loudness prototypes. However, they interpret the use of prototypes in the model much too rigidly. As stated in several venues, "prototypes are generated for the task at hand" [4, p. 17]. Our experience as perceivers of speech in face to face communication includes the positive correlation between loudness and perceived vocal effort by the talker.

The Motor theory assumes that the perceiver uses the sensory input to best determine the set of articulatory gestures that produced this input [5, 6]. One consistent theme for this theory has been the lack of a one-to-one correspondence between the auditory information and a phonetic segment. The inadequate auditory input is assessed in terms of the articulation, and it is only natural that visible speech could contribute to this process. The motor theory has not been formulated, however, to account for the vast set of empirical findings on the integration of audible and visible speech. Traditionally, the motor theory assumes that listeners analyze the acoustic signal to generate hypotheses about the articulatory gestures that were responsible for it. The outcome of the hypothesis testing is derived from the listener's speech motor system. Although the motor theory is consistent with a contribution of visible speech, it has difficulty in accounting for the strong effect of higher-order linguistic context in speech perception [4]. That is, there is nothing in this theory, grounded in modularity, that would allow context to penetrate the the "innate vocal tract synthesizer."

Remez and his colleagues [7] use the perception of sine-wave speech to argue for a view of speech perception very similar to our own.. They assume that the distal objects of perception are phonetic objects. We agree, but would replace phonetic with linguistic so as not to limit ourselves to a particular type of object, or to preclude higher-order recognition at the word or sentence level without recognition at the phonetic level. People might easily perceive a word without being aware of the phonetic segments that make it up. Remez et al assume that there is an unlimited set of cues that can be used to perceive a message. These cues have no prior grouping relationship to one another: the meaningfulness of the input binds them together. Neither the Gestalt laws of organization nor a schema-based grouping [8] can account for the perceptual grouping of these cues. Finally, somehow the appropriate sensory convergence takes place without reference to prototypes or standards in memory.

The major difference between the Remez et al. view and our view probably has to do with the role of prototypes or standards in memory. We have shown that perception occurs in the framework of one's native language [9, 10]. The same speech signal has very different consequences for speakers of different languages. It is difficult to comprehend how this could occur without a central role of memory.

INTEGRATING AUDIBLE AND VISIBLE SPEECH

More generally, many of us have grappled with the appropriate metaphor for audible-visible speech perception. A simple metaphor comes from the use of visible information in speech recognition by machine [11]. The auditory information is the workhorse of the machine, and the visible information is used in a relatively posthoc manner to decide among the best alternatives determined on the basis of the auditory information. At a psychological level, this model is similar to but differs from an auditory dominance model in which the perception is controlled by the auditory input unless it is ambiguous [9]. An ambiguous auditory input forces the system to use the visible information.

Other metaphors build on the idea of combination or integration. Somehow the visible and auditory information is combined, integrated, or joined together. The formalization of this operation is

hotly debated. The two inputs are said to be fused [4], morphed, or converged [6] Fusion and morphing imply some type of blending, whereas convergence appears to be a brain metaphor describing how the different brain systems processing inputs from separate modalities converge for further processing in another brain area. The nature of this blending thus becomes a focal point for investigation and theorizing. We believe that these metaphors must be refined to specify the mathematical manner in which the different modalities are combined.

Of these major theories of speech perception, only the FLMP has provided a formal quantitative description of *how* the auditory and visual sources are processed together to determine perceptual recognition. The FLMP is well-qualified for describing the integration of audible and visible speech because it is centered around the theme of the influence of multiple sources of information. In addition, addressing the nature of the integration process cannot be adequately addressed independently of the dynamics of bimodal speech perception, and it is only the FLMP that takes a stand on the time course of audible-visible speech perception. As shown in Figure 1, the model consists of three operations: feature evaluation, feature integration, and decision. The sensory systems transduce the physical event and make available various sources of information called features. These continuously-valued features are evaluated, integrated, and matched against prototype descriptions in memory, and an identification decision is made on the basis of the relative goodness-of-match of the stimulus information with the relevant prototype descriptions.

During feature evaluation, the features of the stimulus are evaluated in terms of prototype descriptions of perceptual units of the perceiver's language. For each feature and for each prototype, feature evaluation provides information about the degree to which the feature in the signal matches the corresponding feature value of the prototype. Some investigators have argued against this early analysis of the input relative to knowledge in memory. However, it is well-documented that speech perception

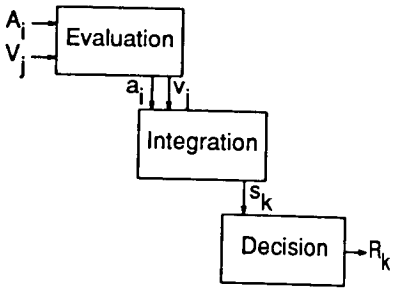


Figure 1. Schematic representation of the three processes involved in perceptual recognition. The three processes are shown to proceed left to right in time to illustrate their necessarily successive but overlapping processing. These processes make use of about prototypes stored in long-term memory. The sources of information are represented by uppercase letters. Auditory information is represented by A_i *and visual information by* V_j. *The evaluation process transforms these sources of information into psychological values (indicated by lowercase letters* a_i *and* v_j*) These sources are then integrated to give an overall degree of support,* s_k, *for a given speech alternative* k. *The decision operation maps the outputs of integration into some response alternative,* R_k. *The response can take the form of a discrete decision or a rating of the degree to which the alternative is likely.*

occurs in real time and we see no justification for some type of temporal delay in the contact of sensory input to information in memory. As an example, we have evidence that speech readers begin accumulating information of a /ba/ even before stop closure [11].

During the second operation of the model, called feature integration, the features (actually the degrees of matches) corresponding to each prototype are combined (or conjoined in logical terms). The outcome of feature integration consists of the degree to which each prototype matches the stimulus. The third operation is decision. During this stage, the merit of each relevant prototype is evaluated relative to the sum of the merits of all relevant prototypes. This relative goodness-of-match gives the proportion of times the

stimulus is identified as an instance of the prototype, or a rating judgment indicating the degree to which the stimulus matches the category. A strong prediction of the FLMP is that the contribution of one source of information to performance increases with the ambiguity of the other available sources of information.

Clarifying the FLMP

Some clarification of the FLMP is necessary because neither real processing nor predicted processing corresponds to a strict single channel discrete flow of information. The three processes shown in Figure 1 are offset to emphasize their temporal overlap. Evaluated information is passed continuously to integration while additional evaluation is taking place. Although it is logically the case that some evaluation must occur before integration can proceed, these two processes overlap in time. Similarly, integrated information is continuously made available to the decision process.

It is also necessary to emphasize that information transformed from one process to another does not obliterate the information from the earlier process. Thus, evaluation maintains its information even while simultaneously passing it forward to the integration process. This parallel storage of information does not negate the sequential process model in Figure 1. What is important to remember is that transfer of information from one process to another does not require that the information is lost from the earlier process. Integrating auditory and visual speech does not compromise or modify the information at the evaluation process. In the FLMP, the representation at one process continues to exist in unaltered form even after it has been "transformed" and transmitted to the following process. As an example, the abstract or amodal categorization of a speech signal does not replace its multimodal sensory representation. The simultaneous maintenance of several levels of information is central to the FLMP. We have shown that perceivers can report modality-specific information being maintained at feature evaluation after this same information has been combined at feature integration [11, 12].

More generally, information in the evaluation process maintains its integrity, and can be used independently of the output of integration and decision. Perceivers are not limited to only the output of integration and decision: they can also use information at the level of evaluation when appropriate. It is well-known, for example, that the relative time of arrival of audible and visible speech can greatly reduce the uncertainty about voicing [13, 14]. We know since the time of Hirsh's seminal studies [15] that perceivers are highly sensitive to temporal onset differences in the two modalities. It would not be surprising, therefore, if perceivers used this temporal asynchrony as a cue to voicing. Furthermore, the temporal asynchrony should be conceptualized as a derived cue that can be integrated with other audible and visible cues.

The FLMP predicts prototypical results of integration, as in the case in which a visual /da/ and an auditory /ba/ produces the percept /ðə/. However, it is not inconsistent with a perceiver's ability to determine the temporal relationship between the auditory and visual input, as in the case when the temporal alignment of the lip movements and auditory tone pulses generated by vocal fold activity can be used as a cue to voicing. This latter phenomenon was used [13] to argue against the independence assumption of the FLMP—that the two sources of information are evaluated independently of one another. By independence, however, we simply mean that the representation of one cue at evaluation is not modified by another cue.

The degrees of support provided by the features from one modality for a given alternative are not modified by the information presented along other modalities. At the same time, the temporal relationship between two modalities might be used as an additional source of information. This comparison could therefore make available "higher-order" multimodal information indicating the temporal relationship between the audible and visible speech. This relative time of arrival could accordingly be used as a cue to voicing, which would be sent forward to the integration process. Comparisons across modalities

could also provide information about the degree to which there was a phonetic discrepancy, and permit perceivers to make some other judgment such as rating the degree to which there was a discrepancy between the auditory and visual inputs [8]. The assumption of independence does not imply that there is no knowledge about what information is available from each modality, and when it is available.

Modality-Specific Representations

We have demonstrated that observers have access to modality-specific information at evaluation even after integration has occurred. This result is similar to the fact that observers can report the degree to which a syllable was presented even though they categorically labeled it as one syllable or another. A system is robust when it has multiple representations of the events in progress, and can draw on the different representations when necessary. In the Massaro and Ferguson [11] study, 20 subjects performed both a perceptual identification task and a same-different discrimination task. There were 3 levels (/ba/, neutral, /da/) of visual information and 2 levels (/ba/, /da/) of auditory. This design gives 6 unique syllables for identification, and there were 20 types of discrimination trials: 6 types of same trials, 6 types of trials with auditory different, 4 types of trials with visual different, and 4 types of trials with both auditory and visual different.

The predictions of the FLMP were derived for both tasks, and the observed results of both tasks were described with the same set of parameter values. For integration in the identification task, the degree of auditory support for the alternative /ba/ in a two-alternative forced choice task is a_i. The visual support for /ba/ is v_j, With just two alternatives /ba/ and /da/, if a visual feature supports /ba/ to degree v_j, then it supports alternative /da/ to degree $(1 - v_j)$, and similarly for the auditory feature. In this case, the overall support for alternative /ba/, S(/ba/), given audible and visible speech, is

$$S(/ba/) = a_i v_j \quad (1)$$

The support for /da/, S(/da/) is equal to

$$S(/da/) = (1 - a_i)(1 - v_j) \quad (2)$$

The predictions of a /ba/ judgment, P(/ba/), is equal to

$$P(/ba/) = \frac{a_i v_j}{a_i v_j + (1-a_i)(1-v_j)} \quad (3)$$

Given the FLMP's prediction for the identification task, its prediction for a same-different task can also be derived. Faced with a same-different task, we assume that the observer evaluates the difference along both the auditory and visual modalities and responds different if a difference is perceived along either or both modalities. Thus, the task is basically a disjunction decision within the framework of fuzzy logic. The perceived difference, d_v, between two levels j and j+1 of the visual factor is given by

$$d_v = v_j - v_{j+1}. \quad (4)$$

Analogously, the perceived difference d_a, between two levels i and i+1 of the auditory factor is

$$d_a = a_i - a_{i+1}. \quad (5)$$

Given two bimodal speech syllables, the perceived difference, d_{va}, between them can be derived from the FLMP's assumption of a multiplicative conjunction rule, using DeMorgan's Law,

$$d_{va} = d_v + d_a - d_v d_a. \quad (6)$$

It is also assumed that the participant computes the degree of sameness from the degree of difference, using the fuzzy logic definition of negation. In this case, the degree of sameness, s_{va}, is equal to

$$s_{va} = 1 - d_{va}. \quad (7)$$

The participant is required to select a "same" or "different" response in the discrimination task. The actual same or different response is derived from the RGR. The probability of a different response, P(d), is thus equal to

$$P(d) = \frac{d_{va}}{d_{va} + s_{va}} = d_{va}, \quad (8)$$

where d_{va} is given by Equation 3.

The predictions of the FLMP were fit to both the identification and discrimination tasks of each of 20 subjects. For each subject, all 26 points were fit with the same set of parameter values. The simultaneous prediction of identification and discrimination insures parameter identifiability, even when only the factorial conditions are tested. There 6 unique syllables in identification, and there were 14 types of different trials and 6 types of same trials. These 26 independent observations were predicted with just 5 free parameters, corresponding to the 3 levels of the visual factor and the 2 levels of the auditory factor. The FLMP gave a good description of the average results, with an RMSD of .0805.

An alternative model was formulated to test the idea that the auditory and visual sources are blended into a single representation, without separate access to the auditory and visual representations. The only representation that remains after a syllable is presented is the overall degree of support for the response alternatives. What is important for this model is the overall degree of support for /ba/ independently of what modalities contributed to that support. In this six-parameter model, it is possible to have two syllables made up of different auditory and visual information, but with the same degree of support for /ba/. For example, a visual /ba/ paired with an auditory /da/ might give a similar degree of overall /ba/ as a auditory /ba/ paired with an visual /da/. When formulated, this model gave a significantly ($p < .001$) larger RMSD of .1764. These model fits provide evidence that the auditory and visual sources of information are maintained independently of one another in memory, even after integration has occurred.

METHODS FOR TESTING MODELS

Grant and Walden (this volume) test the FLMP and the prelabeling model of Braida [17] against confusion matrices of individual subjects. The Prelabeling model (PM) putatively outperformed the FLMP in terms of accounting for bimodal performance as a function of unimodal performance. We believe, however, that for several reasons the test of the prelabeling model against the FLMP has been inadequate and biased, and the results of the comparison incorrect. The limitations are a) the FLMP was fit with no free parameters whereas the fit of the PM allowed many parameters to vary, b) the PM, as used by its adherents, has been biased in favor of fitting bimodal data, and c) the fit of the FLMP and PM has wrongly assumed that the unimodal data are noise free estimates of performance. We now discuss these related issues in greater detail.

In our model tests, the free parameters are adjusted to maximize the overall goodness of fit between the entire data set and a model's predictions. In the PM theorists tests of the FLMP, the bimodal performance was predicted directly from the unimodal performance. The FLMP predicts that the probability of a response to a unimodal condition is equal to the truth value supporting that response. Thus, it seems reasonable to set the free parameters in the FLMP equal to the unimodal performance levels. For example, if a participant correctly identified a visual /ba/ 85% of the time, the the visual amount of /ba/-ness given that visual stimulus would be .85. This value would be used along with the other parameters derived in the same manner to predict the bimodal performance. This would be valid test of the FLMP, however, only if the unimodal identifications are noise free measures and have very high resolution. The first requirement is certainly wrong: behavioral scientists have yet to uncover a noise free measure of performance. The second requirement is also important when confusion matrices are generated. Many cells of the confusion matrix might be 0 or 1 simply because of a relatively small number of observations per condition. Both of these factors can lead to a poor fit of the FLMP when the. unimodal probabilities are used to predict the bimodal responses. To determine the truth of a theory, we believe that it is necessary to measure how well it accounts for the entire pattern of results, rather than how well some conditions predict others. In our model tests, the optimal parameter values are used to predict the entire data set. We do not reserve this technique for tests of the FLMP but allow each competing model to do its personal best.

Braida [17] fit Erber's [18] severely impaired (SI) and profoundly deaf (PD) confusion data. In the fit, the unimodal data was first fit using a KYST technique

to find stimulus and response centers in a multidimensional space based on the unimodal data. In a second stage, these centers are further adjusted to improve the fit of observed and predicted data. This space and categories were then used to predict the bimodal performance. This fit was contrasted with the fit of the FLMP when the unimodal proportions were taken as estimates of the truth values. In this case, the argument is made that comparable methods are being used to fit the PM and the FLMP. However, in addition to having adjustable parameters, the PM's predictions of the unimodal results was not optimal. A multidimensional representation was derived that gave a good fit of the bimodal results at the expense of a poor description of the unimodal data. With the Erber SI data, for example, the coordinates used by Braida yielded RMSDs of .0522 for the visual condition, .0367 for the auditory condition, and .0366 for the bimodal condition. For the PD results, the RMSDs were .0651 visual, .0756 auditory and .0400 for the bimodal. However, when the fit to the unimodal data is maximized, the RMSDs for the SI data are .0255 visual, .0288 auditory, and .0443 bimodal. For the PD data set, the RMSDs are .0299 visual, .0343 auditory and .0435 bimodal. Thus, optimizing the fit of the unimodal data decreases the accuracy of the bimodal predictions. In contrast, the FLMP yielded RMSDs of .0385 and .0509 for the SI and PD bimodal data when the parameters were fixed by the unimodal data. When the parameters were estimated from all of the results, the RMSDs dropped to .0121 and .0114.

As stated earlier, fitting the FLMP on the basis of the unimodal judgments makes the necessarily inaccurate assumption that the unimodal observations are noise free. In order to illustrate the fallibility of this method, we carried out a simulation using some previous results. In this study, participants identified auditory, visual, and bimodal syllables in which the visual syllables were either /ba/, /va/, /ða/, or /da/ and the auditory syllables fell on a ten-step continuum between these same syllables. The FLMP was fit to each subjects results and provided an excellent description of performance, with a mean RMSD of .0198. In this case, the FLMP was fit to 46 individual subjects data by estimating the free parameters using *all* of the observed results from the expanded factorial design. These predictions of the FLMP were used to generate 10 simulated subjects from each of the original 46. Rather than taking the noise-free predictions of the FLMP, each predicted point was assumed to be a value from a binomial distribution with a variance based on the number of observations (16) in the actual experiment. This new set of points now corresponded to a simulated subject. The FLMP was now fit to the simulated subjects, using Grant and Walden's method of predicting the bimodal results with the unimodal observations. Using their method, the FLMP gave a very poor description of the bimodal results of the simulated FLMP subjects. The fit of the bimodal results derived from the FLMP predictions with added noise had a mean RMSD of .0478 for the 460 pseudo-subjects. This demonstration exposes the limitations of testing models by using unimodal performance to predict bimodal results.

MOTOR REPRESENTATIONS

Robert-Ribes, Schwartz, and Escudier (this volume) advocate an amodal motor representation to account for the integration of audible and visible speech. We believe that this account suffers from many of the same problems posed for motor theories of speech perception more generally, such as accounting for the influence of higher-order linguistic context. Furthermore, it is not obvious how this model can account for the cue value of the temporal arrival of the two sources of information—a result they use against the FLMP. Some type of representation is necessary to account for the joint influence of audible and visible speech but we see not compelling reason that this representation should be a motor one.

ACKNOWLEDGMENT

The research reported in this paper and the writing of the paper were supported, in part, by grants from the Public Health Service (PHS R01 NS 20314) and the National Science Foundation (BNS 8812728).

REFERENCES

[1] Diehl, R. L., & Kluender, K. R. (1987). On the categorization of speech sounds. In S. Harnad (Ed.) *Categorical perception* (pp. 226-253). Cambridge: Cambridge University Press.

[2] Massaro, D. W. (1987). Psychophysics versus specialized processes in speech perception: An alternative perspective. In M. E. H. Schouten (Ed.) *The psychophysics of speech perception* (pp. 46-65). Amsterdam: Marinus Nijhoff Publishers.

[3] Fowler, C. A., & Dekle, D. J. (1991). Listening with eye and hand: Cross-modal contributions to speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 17,* 816-828.

[4] Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry.* Hillsdale, NJ: Lawrence Erlbaum Associates.

[5] Liberman, A, Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21,* 1-33.

[6] Mattingly. I. G., & Studdert-Kennedy, M. (Eds.) (1991). *Modularity and the motor theory of speech perception.* Hillsdale, NJ: Erlbaum.

[7] Remez, R. E., Rubin, P. E., Berns, S. M., Pardo, J. S., & Lang, J. M. (1994). On the perceptual organization of speech. *Psychological Review, 101,* 129-156.

[8] Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound.* Cambridge, MA: MIT press.

[9] Massaro, D. W., Cohen, M. M., & Smeele, P.M.T. (1995). Cross-linguistic comparisons in the integration of visual and auditory speech *Memory & Cognition, 23,* 113-131.

[10] Massaro, D. W., Tsuzaki, M., Cohen, M. M., Gesi, A., & Heredia, R. (1993). Bimodal speech perception: An examination across languages. *Journal of Phonetics,* 21, 445-478.

[11] Petajan, E. (1985). Automatic lipreading to enhance speech recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.,* 40-47.

[12] Massaro, D. W., & Ferguson, E. L. (1993). Cognitive Style and perception: The relationship between category width and speech perception, categorization, and discrimination. *American Journal of Psychology,* 106, 25-49.

[13] Breeuwer, M. & Plomp, R. (1985) Speechreading supplemented with formant-frequency information for voiced speech. *Journal of the Acoustical Society of America, 77,* 314-317.

[14] Bernstein, L. E. (1989). Independent or dependent feature evaluation: A question of stimulus characteristics. *Behavioral and Brain Sciences, 12,* 756-757.

[15] Hirsh, I.J. (1959). Auditory perception of temporal order. *Journal of the Acoustical Society of America, 31,* 759-567.

[16] Cohen, M. M., & Massaro, D. W. (1995). Perceiving visual and auditory information in consonant-vowel and vowel syllables. In C. Sorin, J. Mariani, H. Meloni, & J. Schoentgen (Eds.) *Levels in speech communication: Relations and interactions* (pp.25-38). Amsterdam: Elsevier.

[17] Braida, L. D. (1991). Crossmodal integration in the identification of consonant segments. *Quarterly Journal of Experimental Psychology, 43A,* 647-677.

[18] Erber, N. P. (1972). Auditory, visual, and auditory-visual recognition of consonants by children with normal and impaired hearing. *journal of Speech and Hearing Research, 15,* 423-422.

AUDITORY, VISUAL AND AUDIOVISUAL VOWEL REPRESENTATIONS: EXPERIMENTS AND MODELLING

Jordi Robert-Ribes, Jean-Luc Schwartz, Pierre Escudier
Institut de la Communication Parlée, Grenoble, France
{jordi,schwartz}@icp.grenet.fr

ABSTRACT

Audiovisual (AV) speech perception exploits the inherent complementarity of the auditory (A) and visual (V) sensors. We provide new data on the expansion of the vowel triangle in the auditory and visual domain, and on the optimal use of the A-V complementarity for fusion. Then we propose a taxinomy of models for AV integration, and we show that the data in the literature are rather in favor of the so-called MR model, recoding the A and V inputs into an intermediary motor space where integration occurs. Finally, we show that the MR model is not only *plausible* but also *functional*, since it efficiently models AV identification for vowels in noise.

INTRODUCTION

It is now largely accepted that speech is a multimodal means of communication that is conveyed by the auditory and visual system ([1], [2], [3], [4], [5]). How do humans fuse the auditory and visual information? How can recognition systems integrate audio and visual information? Answering the first question we will deal with plausibility, while answering the second we will have to do with functionality.

Studies on audio-visual integration generally have one (*and only one*) of these two approaches: (1) engineering approach, dealing with functionality constraints or (2) experimental psychology approach, dealing with plausibility constraints (and most of the time bypassing the conversion process from inputs into internal representations). This paper studies models of audio-visual speech integration taking into account both plausibility constraints *and* functionality constraints, with an application to vowel perception.

Apart from general questions about sensory interactions and cognitive processes, we have been defending for years the idea that perceptual processing cannot be understood without a deep knowledge of the *structure of the stimuli* [6]. Our section 1 will here be concerned with a set of new data about the perceptual expansion of the vowel triangle in the A and V domain. This will clearly show the *complementarity* of the A and V sensors for vowel place of articulation. In section 2 we will propose a taxinomy of models for AV integration in speech perception, with three successive binary questions leading to four categories of models. We will show how experimental data constrain the choice towards one category, the so-called Motor Recoding Model. Finally, we will show that this model is not only plausible, but also functional, since it efficiently models AV identification of vowels in noise.

1. AUDIO-VISUAL VOWELS

1.1. Physical characteristics

We recorded the vowels [i e ɛ y ø œ u o ɔ a] from a French speaker with the ICP "Video-Speech Workstation" [7]. We obtained images of the speaker's face with the corresponding synchronised sounds. We recorded, for each vowel, of 100 realisations of 64 ms of sound with the corresponding image.

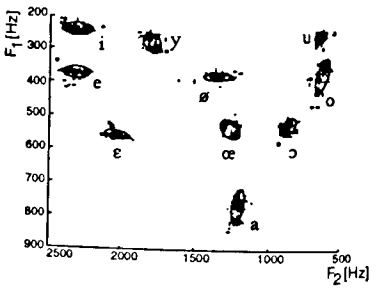


Figure 1. F_1/F_2 representation of the acoustic stimuli used

Their two first formants are presented in Figure 1. We observe the vocalic triangle with vowel [y] close to vowel [i] but far apart from vowel [u]. Notice that even considering higher formants by a global spectral shape analysis, the [i-y] distance remains smaller than the [y-u] one [8].

We extracted from the image the following geometrical parameters of the lip shape: inner-lip horizontal width (A), inner-lip vertical height (B) and inner-lip area (S). Figure 2 presents the stimuli in the A/B plane. We observe a clear separation of rounded vowels ([y ø u o]), semi-rounded vowels ([ɔ œ]) and unrounded vowels ([i e ɛ a]). Notice that vowel [y] is close to vowel [u] but very far apart from vowel [i].

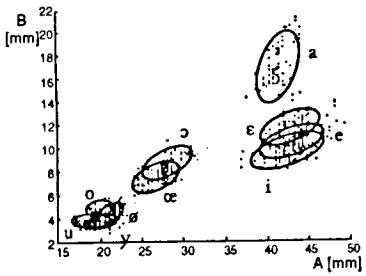


Figure 2. Width (A) / height (B) representation of the optic stimuli used

We summarize the [i y u] acoustical and optical contrasts in Figure 3.

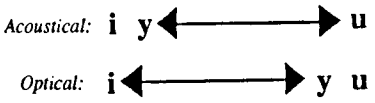


Figure 3. Acoustical and optical contrasts

We see that our stimuli are clearly *complementary*. A number of further analyses of the same stimuli [8] confirm that height contrasts (e.g. [i] vs. [e] vs. [ɛ] vs. [a]) are best represented within the acoustic stimuli. On the contrary, optic stimuli differentiate mainly the stimuli by their rounding (e.g. [i] vs. [y]).

1.2. Audio, visual and audio-visual perception

We carried out a perceptual test on these vowels presented with acoustic noise.

Method

A group of 21 French subjects was presented with 10 realisations of the 7 French isolated (without context) vowels [i e y ø u o a] in audio-visual, visual and audio conditions, with 7 signal-to-noise ratios (SNR). We tested these 7 vowels instead of the 10 available because we didn't want to test the mid-low/mid-high contrast (e.g. [ɛ] vs. [e]) which may be lost in isolation.

Results

Figure 4 shows the correct identification results in percentage corrected to the random level (zero percent means that scores were at random level). This figure shows better audio-visual scores (AV) than audio alone scores (A) and visual scores (V). More detailed analyses based on transmitted information show that this pattern is true for individual phonetic dimensions, namely rounding, height and front-back contrast [8]: we call this the "complementarity rule".

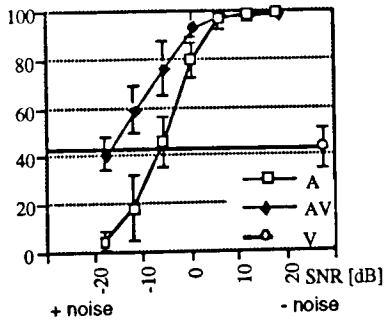


Figure 4. Correct identification scores for the perception test

We also studied the perceptual structure of the auditory and visual perception (for full details, see [8]). We show in Figure 5 a schematic display of the structure we found. We can see that the auditory geometry is stretched in the height dimension ([i] vs. [a]) while the visual geometry is stretched in the rounding dimension ([i] vs. [y]).

Hence our data reveal some audio-visual complementarity: the best information about place of articulation perceived by audition is the worst perceived by vision and vice-versa. Notice that up to now audio-visual complementarity had been rather conceived as an Audition-Mode Vision-

Place complementarity [5]. The complementarity we found in our test is a *complementarity within place of articulation.* The complementarity found in the stimuli (section 1.1) is enhanced by the perceptual system (section 1.2).

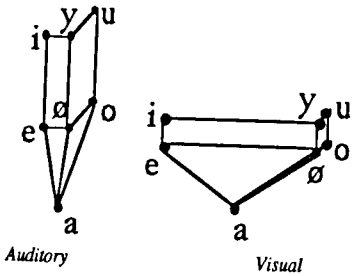


Figure 5. Perceptual structures

2. MODELS FOR AUDIO-VISUAL INTEGRATION

The main theoretical options for modelling audio-visual integration in speech perception have been presented by Summerfield [5]. In the following sections we present these options with slight modifications from our own.

2.1. Four models

Direct Identification (DI Model)

This model assumes a direct identification of the inputs without any transformation. The audio (A) and visual (V) inputs are compiled in a bimodal vector and then classified. Some components of the bimodal vector are A and some components are V (Fig. 6).

Key characteristic of the DI model:

There is *no representation level common to both modalities* between the signal and the percept.

Separate Identification (SI Model)

This model assumes that both inputs have been compared to prototypical forms before the fusion stage. The comparison to prototypes can even lead to a classification of each modality. The two codes (one from the A input and one from the V input) are then combined by means of rules or logical criteria (Fig. 7).

The information at the fusion level can also be continuous (and not discrete).

The key point here, however, is whether it is the result of a comparison to prototypes or not. Therefore, the FLMP (Fuzzy Logical Model of Perception [4], [9]) is one of the SI models because the information at the fusion level is information "indicating the degree of support for one alternative" ([9], p. 743).

Key characteristic of the SI model:

Inputs are compared to prototypes (or even classified) *before fusion.*

Dominant modality Recoding (DR Model)

One of the possibilities that Cognitive Psychology presents for fusing two modalities is the recoding of one modality into the other —supposed to be the dominant modality— [10]. The DR model assumes that the auditory modality is dominant in speech perception. Thus the visual input is recoded into an auditory space where both sources of information are fused [11].

In this model the visual input is used to estimate the vocal tract filter. This estimation is then in some sense averaged with the one derived from auditory processing, while the source characteristics are estimated only from the auditory path. The combined source and filtering characteristics thus estimated are then provided to a phonetic classifier (Fig. 8).

Key characteristic of the DR model:

The *visual modality is recoded into an auditory representation space* where it is fused with the auditory information.

Motor space Recoding (MR Model)

This model assumes that both inputs are projected into a common amodal space where they are fused. This space is amodal because it is homogeneous to neither of the modalities (auditory or visual): it is a motor representation space. This supposes that, in order to perceive speech, we recover the common cause of both the auditory and the visual signals, namely: the motor representation [12] (Fig. 9).

Key characteristic of the MR model:

Both inputs are projected into a motor representation space where fusion occurs.

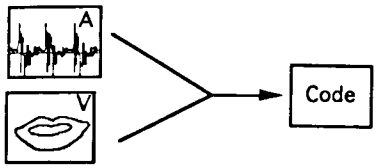


Figure 6. DI Model

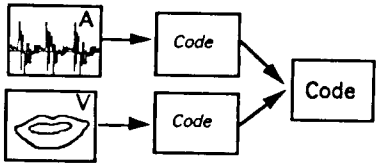


Figure 7. SI Model

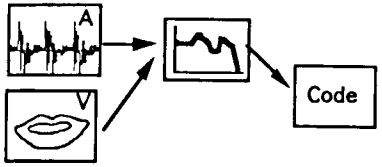


Figure 8. DR Model

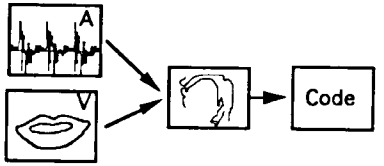


Figure 9. MR Model

2.2. Three questions for a taxonomy

We can attempt to classify the above presented architectures under a synthetic form by comparing them to the general cognitive psychology models. To do so, we can ask three questions.

(1) Does the interaction between modalities imply a *common intermediate representation* or not? The answer allows a first opposition between Model DI (for which there is NO common representation) and the other architectures.

(2) If there is an intermediate representation, does it rely on the existence of prototypes or not? In other words, is it "late" or "early" (see [13], p. 25)?

Integration is considered "late" when it occurs after the decoding processes or the comparison to prototypes (Model SI), even if these processes give continuous data (as with the FLMP). Otherwise, integration is "early" when it applies to continuous representations, common to both modalities, and which are obtained through low-level mechanisms which do not rely on any decoding process nor any comparison to prototypes: It is the case with models DR and MR.

(3) Is there at last any dominant modality which can give a common intermediate representation in an early integration model (DR)? Or is this common representation amodal (such as in the MR model)?

These questions lead to the taxonomy presented on Figure 10.

2.3. Three "plausible" responses to the three questions

About the need for a common representation

In a study on the audio-visual perception of vowels, Summerfield and McGrath [14] showed that subjects detect the incompatibility between the auditory and visual inputs, while they cannot however avoid fusing both inputs.

Even young babies are sensitive to the correspondence between auditory and visual information of a speaking face [15]. When they are presented with two faces and only one sound, babies prefer to look at the face that is articulating the

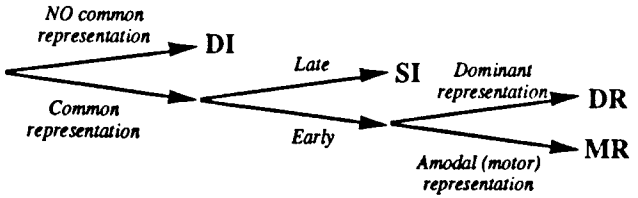


Figure 10. Taxonomy of models

sound they are hearing.

Therefore, the need of a common representation of auditory and visual stimuli is inescapable. Stimuli can be compared in that space before being fused: *stricto sensu*, model DI has to be rejected.

About the need for early integration

Subjects are able to estimate temporal co-ordinations between the auditory signal and the visual signal: they can estimate the VOT audiovisually ([16], [3]). This is clearly incompatible with a late integration model where neither of the two inputs provides information enough to identify the voicing feature, which is hence recovered from the co-ordinations between the auditory and visual signals. It seems that the evaluation of this co-ordination should be done on signals that are not the output of a prototype comparison (in which case the information leading to the co-ordinations would be lost).

In another experiment, the speaking rate perceived audiovisually is the mean of the speaking rate perceived auditorily and the speaking rate perceived visually [17]. In addition, the audio-visual speaking rate can change the phonemic frontier between voiced and unvoiced phonemes [18]. This is hardly compatible with a late integration model because rate is a quantitative information which would be lost after comparison to prototypes.

These facts indicate that subjects can make decisions from audio-visual information, decisions that are impossible to make for each modality independently. The fusion has to be done at an early stage of processing. Late models (SI Model) have to be rejected.

About dominance and complementarity

We have seen in section 1.2 that the best information perceived by audition is the worst perceived by vision and vice versa. The DR Model cannot exploit this complementarity because all the inputs are transformed into an acoustic representation. Thus, audiovisual confusions will be similar to acoustic confusions. Let us develop this idea.

Model DR recodes the visual input into an auditory representation where fusion takes place. From our data in section 1.1, two facts have to be pointed out: (1) Model DR will try to recode visual stimuli distant in the visual space (as [i] and [y]) into close points in the auditory space, and (2) stimuli close in the visual space (as [y] and [u]) will need to be recoded into distant points in the auditory space. This is represented in Figure 11.

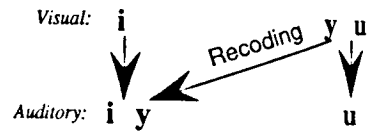


Figure 11. Bad recoding of Model DR

Point (2) results in a lack of stability of the visual-to-auditory transform, which makes the problem of building this association not trivial [19]. Point (1) results in a lack of optimality, which makes the DR Model not very efficient for dealing with vowels in noise (see section 3). But, more seriously, the DR Model predicts that a given V input can modify an A percept only if the V input is conflictual or if the A input is noisy (see an application of the DR Model structure for noisy speech enhancement in [20]). However, some data in a study by Lisker and Rossi [21] show that a V input can bias a congruent and clear auditory stimulus. Their subjects (French-speaking speech researchers) were asked to decide on the rounding category of each vowel. Let us concentrate on the case of [ɯ]. This vowel when presented visually was considered a rounded vowel 1% of the time. When presented auditorily, it was rather considered rounded (60% vs. 40% unrounded responses). Finally, when presented audio-visually the percentage of rounding judgements dropped to 25% (vs. 75% unrounded).

Hence, it is clear that some subjects perceived the vowel [ɯ] as rounded when presented auditorily but they judged it unrounded when presented visually and audio-visually. This fact is hardly conceivable within a DR Model which recodes visual input into an auditory space. Consequently, the DR Model has to be rejected.

Conclusion: plausibility constraints

We have seen that (1) a plausible model of audio-visual integration has to use a common representation between audio and visual inputs, (2) this representation must be placed at an early stage of processing, and (3) this representation cannot be the auditory representation.

The only model that we have not been able to eliminate (Model MR) is in agreement with these three points. We will see in section 3.5 that it can simulate the results found by Lisker and Rossi [21].

In conclusion, the MR Model is "plausible": we will see in the next section that it is also "functional".

3. MR MODEL FUNCTIONALITY

We will present in this section an implementation of the MR Model for the recognition of French vowels. This is the first implementation of this model in the literature. We proved elsewhere ([8], [22]) that the DR Model is less functional than the MR one, namely that it has worse results in a recognition task.

The MR Model implementation is based on simple but controllable tools, which were chosen to allow a good comparison with the DR model (see [8]).

3.1. Motor representation

A crucial choice in the MR Model concerns the definition of the "motor space" in which integration should occur. Since we deal with static vowels, we have chosen articulatory representations based on three parameters, namely X, Y (which are respectively the horizontal and vertical co-ordinates of the highest point of the tongue) and S (the inner-lip area). Of course, X, Y and S respectively provide articulatory correlates of the front-back, open-close and rounding dimensions (see Fig. 12).

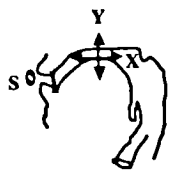


Figure 12. Motor representation

3.2. From the inputs to the motor representation

This stage was implemented thanks to linear associations.

The auditory inputs were 20-dimensional dB/Bark auditory spectra. The outputs were typical values for X and Y for French vowels, and the S value extracted on the corresponding image in the corpus.

The visual inputs were (A, B, S) triplets. X (the tongue front-back position) was supposed to be impossible to estimate from the visual path, and S was directly transmitted from the visual input, hence only the association between (A, B, S) and Y had to be learned.

Notice that all the corpus (e.g. 10 vowel classes) was used in this study.

3.3. Fusion of motor representations

The integration consisted in a weighted sum of the representations obtained by each path (audio and visual). An audio-visual estimate of the (X, Y, S) set was finally derived. The parameter X was estimated only from the acoustic path, while the other two parameters were determined from the corresponding ones provided by both paths. This was performed using the following formulas:

$$Y_{AV} = \alpha_Y Y_A + (1-\alpha_Y) Y_V$$

and $$S_{AV} = \alpha_S S_A + (1-\alpha_S) S_V$$

where index A means auditory, V visual and AV audio-visual. Parameters α_Y and α_S are sigmoidal functions of SNR. The parameter α_Y varied between a value close to 0 for low SNR values (too much noise; almost no available information in the acoustic signal) and a value lower than 1 for high SNR values (no noise; the audio-visual percept is influenced by both the visual and the auditory inputs). On the other hand, the parameter α_S was never higher than 0.3 (indicating that the estimation of S is mainly done from the information of the visual path). The parameters of the sigmoids were learned under a criterion of minimal global error for all learning realisations at all SNR values.

3.4. Vowel identification

Classification was achieved by a Gaussian classifier. We used a Gaussian classifier in the (X, Y, S) space, with a

choice of one between ten classes. The learning corpus for estimating the mean and covariance matrix for each vowel class was based on (X, Y, S) triplets delivered by the auditory path alone on realisations presented at 4 different levels of noise covering a large range between no noise and largely degraded but still partly recognisable stimuli (SNR = 99, 24, 12 or 0 dB).

The whole schema is displayed in Figure 13.

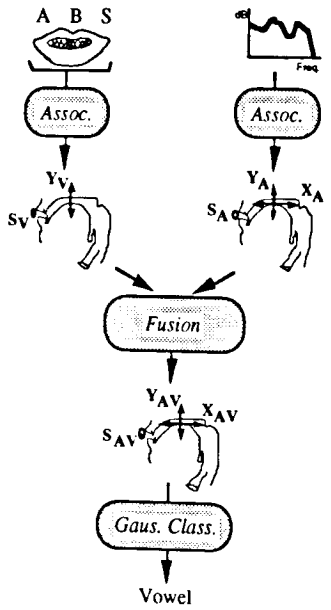


Figure 13. Implementation of the MR Model

3.5. MR Model and complementarity

How could this implementation simulate Lisker and Rossi's data [21] presented in section 2.3? It is known that rounding is highly correlated with parameter S (inner-lip area). We saw that the audio-visual estimate of S by the MR Model mainly depends on the value of S provided by the visual input. However, one value of S is also estimated from the auditory input. Then, a decision about rounding can be taken from the auditory input alone or from the visual input alone. When both inputs (audio and visual) are present, the rounding decision mainly depends on the decision taken from the visual input. This was exactly the case found in the experience of Lisker and Rossi [21] described earlier. Our implementation of the MR Model can simulate this result.

3.6. Results

We present in Figure 14 the identification scores when the audio or the audio-visual stimuli were presented at the input of the implementation. Since dimension X (front-back) cannot be estimated from the visual input, we estimated a visual score by considering (arbitrarily) all rounded vowels as being back-vowels.

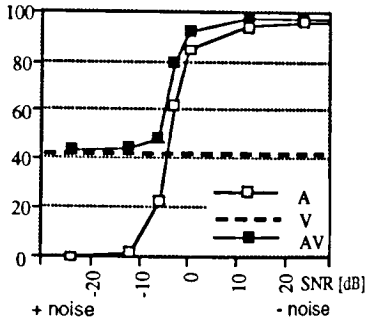


Figure 14. Correct identification for the MR Model

Recognition scores in Fig. 14 are lower than perceptual scores in Fig. 4. This is due to both the higher number of classes in the second case (10 vs. 7) and the simplicity of tools in MR implementation. However, the interesting point is that the visual gain (difference between the A and AV conditions) is high in the model, and it respects the "complementarity rule": transmitted information on each of the three phonetic dimensions is greater for the AV condition than for both the A and V conditions [8].

CONCLUSION

We have shown that a plausible model of audio-visual integration for speech perception requires three characteristics: (1) use a common representation between audio and visual inputs, (2) fuse the modalities at an early stage of processing, and (3) do not use the auditory space as the fusion space.

A model that fuses both informations in the motor space (Model MR) has these three characteristics. We have shown that this model is also functional in a vowel recognition task. We are currently attempting to adapt this model to dynamic stimuli, with more complex processing tools and architectures. We hope that this model should be able to produce some McGurk effect (see [8]).

Acknowledgement

This work has been supported by ESPRIT-BR (6975) funding (Speech Maps Project).

REFERENCES

[1] Erber, N.P. (1975). "*Auditory-visual perception of speech*", J. Speech and Hearing Disorders 40, 481-492.

[2] McGurk, H. and MacDonald, J. (1976). "*Hearing lips and seeing voices*", Nature 264, 746-748.

[3] Breeuwer, M. and Plomp, R. (1986). "*Speechreading supplemented with auditorily presented speech parameters*", J. Acoust. Soc. Am. 79, 481-499.

[4] Massaro, D.W. (1987). *Speech perception by ear and eye : a paradigm for psychological inquiry*. London: LEA.

[5] Summerfield, Q. (1987). "Some preliminaries to a comprehensive account of audio-visual speech perception", in Dodd, B. and Campbell, R. (Eds.) *Hearing by eye : the psychology of lipreading*. (pp. 3-51). London: LEA.

[6] Schwartz, J.L. and Escudier, P. (1991). "Integration for extraction: What speech perception researchers can learn from Gibson and Marr," in *XII Congrès International des Sciences Phonétiques*. (Vol. 1, pp. 68-72).

[7] Lallouache, M.T. (1990). "Un poste 'visage-parole'. Acquisition et traitement de contours labiaux," in *XVIII Journées d'Études sur la Parole*. pp. 282-286).

[8] Robert-Ribes, J. (1995). *Models of audiovisual integration*. PhD Thesis Institut National Polytechnique de Grenoble.

[9] Massaro, D.W. (1989). "*Multiple book review of speech perception by ear and eye: A paradigm for psychological inquiry*", Behavioral and Brain Sciences 12, 741-794.

[10] Hatwell, Y. (1986). *Toucher l'espace. La main et la perception tactile de l'espace*. Lille: Presses Universitaires de Lille.

[11] Yuhas, B.P., Goldstein, M.H., and Sejnowski, T.J. (1989). "*Integration of acoustic and visual speech signals using neural networks*", IEEE Communications Magazine Nov. 89, 65-71.

[12] Fowler, C.A. and Rosenblum, L.D. (1991). "The perception of phonetic gestures", in Mattingly, I.G. and Studdert-Kennedy, M. (Eds.) *Modularity and the motor theory of speech perception*. (pp. 33-59). Hillsdale (NJ): Erlbaum.

[13] Vroomen, J.H.M. (1992). *Hearing voices and seeing lips: Investigations in the psychology of lipreading*. PhD Thesis Katolieke Univ. Brabant.

[14] Summerfield, Q. and McGrath, M. (1984). "*Detection and resolution of audio-visual incompatibility in the perception of vowels*", Quarterly J. Experimental Psychology : Human Experimental Psychology 36A, 51-74.

[15] Kuhl, P.K. and Meltzoff, A.N. (1982). "*The bimodal perception of speech in infancy*", Science 218, 1138-1141.

[16] Rosen, S., Fourcin, A.J., and Moore, B. (1981). "*Voice pitch as an aid to lipreading*", Nature 291, 150-152.

[17] Green, K.P. and Miller, J.L. (1985). "*On the role of visual rate information in phonetic perception*", Percep. and Psychophysics 38, 269-276.

[18] Green, K.P. and Kuhl, P.K. (1989). "*The role of visual information in the processing of place and manner features in speech perception*", Perception and Psychophysics 45, 34-42.

[19] Robert-Ribes, J., Lallouache, T., Escudier, P., and Schwartz, J.L. (1993). "Integrating auditory and visual representations for audiovisual vowel recognition," in *Proc. 3rd Eurospeech*. (pp. 1753-1756).

[20] Girin, L., Feng, G., and Schwartz, J.L. (1995). "Noisy speech enhancement with filters estimated from the speacker's lips," in *Proc. 4th Eurospeech-95*.

[21] Lisker, L. and Rossi, M. (1992). "*Auditory and visual cueing of the [±rounded] feature of vowels*", Language and Speech 35, 391-417.

[22] Robert-Ribes, J., Schwartz, J.L., and Escudier, P. (in press). "*A comparison of models for fusion of the auditory and visual sensors in speech perception*", Artificial Intelligence Review.

PREDICTING AUDITORY-VISUAL SPEECH RECOGNITION IN HEARING-IMPAIRED LISTENERS

Ken W. Grant and Brian E. Walden (Army Audiology and Speech Center, Walter Reed Army Medical Center, Washington, DC 20307-5001)

ABSTRACT

Individuals typically derive substantial benefit to speech recognition from combining auditory (A) and visual (V) cues. However, there is considerable variability in AV speech recognition, even when individual differences in A and V performance are taken into account. In this paper, several possible sources of subject variability are examined, including segment perception, AV integration skill, and context usage. When these sources of variability are accounted for, predictions of AV speech recognition of nonsense syllables for normally-hearing and hearing-impaired listeners are excellent (R^2=0.96). Predictions for AV sentence recognition, however, are much poorer (R^2=0.44). These data will be discussed as part of a generalized model of AV speech recognition which includes the use of A and V unimodal cues, the integration of A and V cues, and the use of phonemic and semantic context.

[Work supported by NIH Grant DC 00792 and the Department of Clinical Investigation, Walter Reed Army Medical Center].

INTRODUCTION

In most communication settings, speech perception involves the integration of both auditory (A) and visual (V) information [1-4]. Further, auditory-visual (AV) speech perception is almost always better than either hearing or speechreading alone. This is especially true when the auditory signal has been degraded due to hearing loss or environmental noise.

Figure 1 shows fairly typical results obtained from intelligibility tests using low-context sentences [5] presented in a background of speech-shaped noise (S/N=0 dB) to hearing-impaired subjects. The hearing-impaired subjects tested had a variety of hearing-loss configurations ranging from mild to severe. For convenience, subjects are arranged along the abscissa in order of ascending A scores. As shown in the figure, all subjects demonstrated benefit from the addition of visual cues, but some subjects derived substantially more benefit than others, independent of their A score. For comparison, normally-hearing subjects tested under the same conditions achieved scores of 90%, 10%, and 98% for A, V, and AV conditions, respectively.

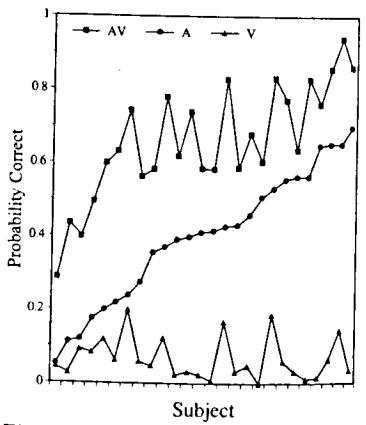


Figure 1. *A, V, and AV sentence recognition by individual hearing-impaired subjects.*

The exact amount of observed benefit depends on a number of variables. Among these are the individual's ability to recognize phonetic (e.g., consonants and vowels) and prosodic (e.g., intonation, duration, and stress) cues, the ability to integrate A and V cues, the difficulty of the speech materials, the physical conditions under which the speech is presented (e.g., noise, reverberation, lighting, viewing angle, etc.), and the individual's knowledge of the language and ability to make use of contextual constraints. Although much is known about the benefits of combining speechreading with audition, the relative effects of each of these factors on AV benefit is largely unknown. In this paper, we discuss some of the primary factors that are important to understanding how hearing and vision are combined in speech recognition.

SEGMENT RECOGNITION

Models of auditory-visual speech perception typically include auditory analyses, visual analyses, and more central processes common to both A and V modalities [6]. Since the classic study by Miller and Nicely [7], the recognition of speech segments (i.e., consonants and vowels) has typically been analyzed in terms of acoustic, phonetic, and articulatory features. Application of these analyses to AV recognition has shown that vision and hearing are often *complementary* in speech recognition under conditions of auditory signal degradation.

Figure 2 shows mean data for consonant feature recognition by normally-hearing subjects as a function of S/N for A, V, and AV conditions [8]. The top panel shows the data for voicing, whereas the middle and bottom panels show the data for manner of articulation and place of articulation, respectively. A and AV feature scores are shown by the dashed and solid lines. V feature scores are shown by the AV values displayed at -15 dB S/N. Notice that voicing cues obtained under AV conditions are determined by audition; that is, there is virtually no difference between AV and A conditions. In contrast, place-of-articulation cues are determined by speechreading. For this feature, AV scores remain virtually constant despite changes in A performance.

Although it may appear that both modalities contribute significantly to the recognition of manner-of-articulation cues, our analyses suggest that this feature is also determined by audition. It is important to remember that voicing, manner and place cues are not independent. Performance on one feature alters the expected chance performance on another. For example, if we assume that place-of-articulation cues are transmitted visually and that responses within place categories are distributed uniformly, then we would predict a visual manner score of 66% by chance alone. This is very close to the score for speechreading alone shown in the middle panel of the figure (depicted by the AV score at S/N = -15 dB). Thus, for these conditions, manner-of-articulation cues were derived primarily from the A condition.

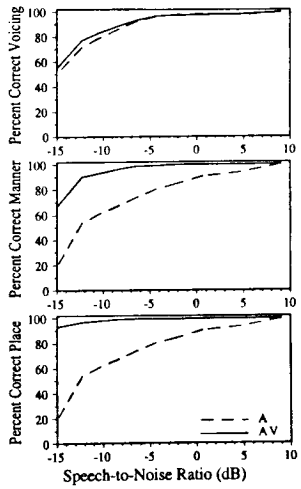


Figure 2. *A, V, and AV feature scores as a function of S/N.*

The complementary relation between auditory manner and voicing cues with visual place cues in speech recognition

has often been cited as a primary reason for the large advantages observed in AV consonant recognition relative to either A or V alone. In previous work, Walden, Prosek, & Worthington [9] developed a measure of AV redundancy that was able to account for a substantial amount of variability observed in AV consonant recognition by hearing-impaired listeners.

In a recent study, Grant and Walden [8] evaluated A, V, and AV consonant recognition by normally-hearing listeners under 12 different filtered-speech conditions. The filters were designed to create a range of A intelligibility scores with different patterns of perceptual confusions across A conditions. Confusion data obtained from each of the A filter conditions were subjected to a Sequential Information Feature Analysis [10], and the proportion of manner-plus-voicing information relative to the total amount of information received was calculated. This proportion represents, to a first approximation, the degree to which A and V information are complementary and shows the proportion of auditory information not obtainable by speechreading alone.

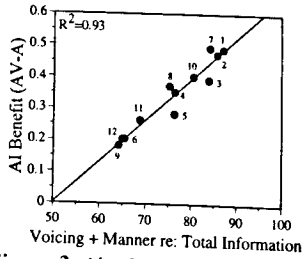


Figure 3. Absolute AI benefit as a function of AV complementation.

Figures 3 and 4 show the results of this analysis after converting percent correct scores to articulation index (AI) units. Figure 3 shows the results for *absolute* benefit (AV-A), while Figure 4 shows the results for *relative* benefit [(AV-A)/(1-A)]. In both figures, the abscissa shows the amount of information received in the A condition for the combined voicing-plus-manner feature expressed as a percent of the total amount of information received. There is a strong relation evident for both benefit measures, indicating that the degree to which A and V conditions complement each other is highly predictive of AV benefit. It should be noted that relating AV benefit to overall A intelligibility resulted in substantially weaker correlations.

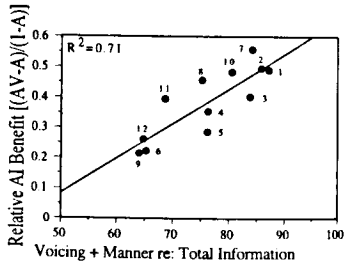


Figure 4. Relative AI benefit as a function of AV complementation.

Models of AV integration which make use of the entire A-alone and V-alone confusion matrices, such as the fuzzy logical model of perception (FLMP) proposed by Massaro [1] or the pre- and post-labelling models (PRE, POS) proposed by Braida [2], have been shown to predict AV consonant recognition more accurately than feature-based models when applied to data averaged across subjects.

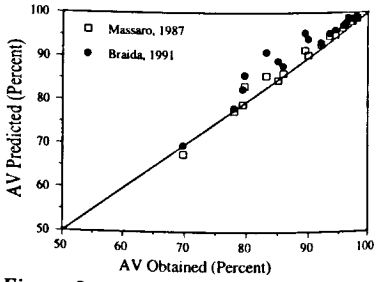


Figure 5. FLMP and PRE-model predictions for normal-hearing subjects.

Figure 5, for example, shows FLMP and PRE model predictions for the 8 S/N and 12 filtered-speech conditions shown in the previous figures. As can be seen, both models account for a large percentage of the variability in the obtained data (R^2 = 0.98 and 0.93 for the FLMP and PRE, respectively).

AUDIOVISUAL INTEGRATION

Whereas the ability to recognize segmental information from A and V speech signals is undoubtedly fundamental to predicting AV recognition, the ability to integrate A and V speech cues is another essential determinant of AV performance [11,1]. With the development of recent quantitative models of multisensory integration [1-2], it is now possible to estimate a listener's integration ability, independent from their ability to recognize A and V speech cues. These models predict AV recognition based on the pattern of segmental confusions obtained for each separate modality. It should be noted, however, that some AV cues, such as the relative timing of lip movements to voicing onset, are multimodal, in that they exist only as inter-modality timing cues. McGrath and Summerfield [12] have suggested that better lipreaders may be sensitive to these cues. Given the accuracy of the FLMP and PRE models, intermodal-timing cues may play only a small role in AV speech perception.

Unlike feature-based models, the FLMP and PRE integration models attempt to make use of all available data obtained in separate A and V identification tasks, and are potentially optimum-processor models. Ideally, a subject's AV performance should never exceed predicted performance. Subjects who perform as predicted are able to make use of all of the available information derived from the unimodal conditions. On the other hand, subjects who perform more poorly than predicted fail to make optimal use of A and V cues.

Our initial efforts to apply the FLMP and PRE models as a gauge of subject integration ability suggests that the PRE model may be more suitable. With the FLMP, stimuli identified correctly in one modality but incorrectly in the other, are predicted to be incorrect in the combined AV condition. As Braida [2] noted, the FLMP does not properly account for structured errors and relies too heavily on unimodal accuracy. In contrast, the PRE model focuses more on the consistency of unimodal responses and not necessarily on accuracy.

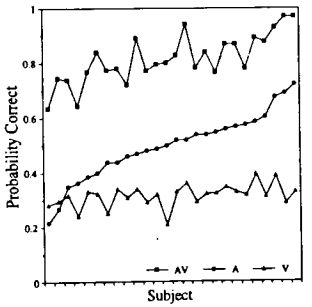


Figure 6. A, V, and AV consonant recognition by individual hearing-impaired subjects.

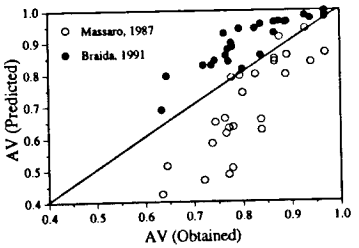


Figure 7. FLMP and PRE model predictions for hearing-impaired subjects.

Figures 6 and 7 show the results of a recent experiment examining consonant and sentence recognition in noise by 26 hearing-impaired subjects. Figure 6

shows A, V, and AV scores for each subject. As with Figure 1, subjects have been ordered along the abscissa according to A performance. Note first the large variability in AV recognition scores across subjects and the moderate correspondence between A and AV scores. This is additional evidence that the overall A score (or V score) does not allow accurate predictions of AV performance.

Predictions made by the FLMP and PRE models for the data shown in Figure 6 are displayed in Figure 7. The modeling results for individual subjects show that predictions of overall accuracy for are far less than perfect. FLMP predictions (R^2=0.71) consistently underpredicted performance when the input unimodal scores were low. Predictions by the PRE model on the other hand (R^2=0.77), were either equal to or greater than obtained performance. Thus, the PRE model, unlike the FLMP, behaves more like an optimal integrator. Some subjects are able to achieve this level of performance, whereas others fall short.

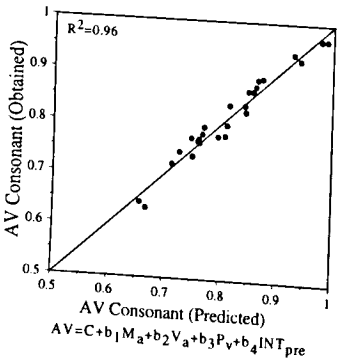


Figure 8. AV consonant predictions using a 4-factor feature-based model.

We have used this feature of the PRE model to estimate the extent to which individual subjects are able to integrate A and V cues optimally. Specifically, the difference between obtained and predicted AV performance was used as an index of AV integration skill. The combination of auditory voicing (V_A), auditory manner (M_A), visual place (P_V), and derived AV integration estimates (INT_{PRE} or INT_{FLMP}) was used in a 4-factor model to predict AV consonant recognition for hearing-impaired subjects. A and V feature scores were obtained from a SINFA analyses of the unimodal conditions and expressed as the percent of conditional information received for that feature relative to the total amount of all information received. For comparison, integration estimates were derived from AV predictions made by both PRE and FLMP models.

Figure 8 shows the results using the PRE model integration estimates. The excellent correlation obtained is impressive considering the simplicity of the model; that is, a linear addition of three unimodal feature scores and a measure of subject integration. Similar attempts to predict AV recognition without integration estimates, or with estimates derived from the FLMP model, led to significantly smaller correlations (R^2=0.822 and R^2=0.823, respectively).

WORD AND SENTENCE PROSODY

Even if AV segment recognition for individual subjects could be predicted perfectly, there would still be the problem of relating segment scores to word and connected-speech scores. One obvious difference between segmental recognition tasks and word and sentence recognition tasks is that the latter two contain important prosodic information related to stress, intonation, and rhythmic structure. The basic function of prosody in speech is to provide information about lexical, grammatical, and emotional aspects of the spoken message [13-14]. Further, individual differences in the ability to extract prosodic information appears to be an important factor in determining AV performance for words and sentences.

Acoustic analyses of prosody have shown that the cues for syllabification, stress, intonation, and phrasing include variations in fundamental frequency, segment and syllable duration, and amplitude envelope [15-20]. In general, speechreaders are not very good at extracting these cues. F0 variations are largely undetectable, and acoustic durational cues signifying segment lengths and intervocalic closure durations are often visually blurred or incompletely specified due to articulator movements that are either too rapid to follow visually or that occur behind the teeth [21-23]. Thus, as with voicing and manner-of-articulation cues, prosodic contrasts detected through audition are highly complementary to speechreading.

An important question related to the use of prosody in speech perception is whether subjects demonstrate variability in their judgements of prosodic contrasts. In a recent study, Grant and Walden [24] measured the ability of normally-hearing listeners to identify syllable number, syllabic stress, intonation, and rhythmic phrase structure in filtered words, phrases, and sentences. The filters used were approximately equal in intelligibility (AI=0.1), and spanned the frequency range from 300-5000 Hz. For some subjects, prosodic features were reliably extracted throughout the frequency spectrum. Other subjects, however, had considerable difficulty identifying sentence intonation and phrase structure from high-frequency speech regions.

Although Grant and Walden did not measure AV performance, the variability which they observed in subjects' abilities to extract suprasegmental cues from various parts of the frequency spectrum may be related to the variability observed in AV speech recognition. It is well known, for example, that there are substantial differences in AV performance among hearing-impaired observers who have the same average auditory recognition scores [25-26]. Given the highly complementary nature of acoustic prosody and speechreading cues, it is possible that some of the variability observed in AV word and sentence recognition tasks may be related to individual differences in the auditory recognition of prosodic and rhythmic cues.

WORD AND SENTENCE CONTEXT

Words and sentences provide listener's with many additional cues besides the usual segmental and suprasegmental cues. For example, identifying *nonsense* syllables requires that each separate consonant and vowel segment be received accurately. However, with *meaningful* words, lexical constraints make it possible to identify words correctly without having to resolve all of the individual segments. Similarly, words presented in isolation typically require more information than if the words were presented in sentences. In order to achieve the desired relationship between segment scores and word and sentence scores, these contextual variables need to be taken into account.

Following Boothroyd and Nittrouer [27], phonemic and semantic constraints can be represented quantitatively by using simple power-law equations. In Equation 1, the recognition of a CVC word is assumed to be equal to the recognition of its component parts. If each of these parts is statistically independent then,

$$P_w = P_p^n, \quad (1)$$

where P_w is the probability of recognition of the whole word, P_p is the probability of recognition of each independent segment, and n is the number of segments in the word. However, in real words, the segments are not independent and it is not required that all segments be received for the word to be recognized. Therefore, for real words,

$$P_w = P_p^j, \quad (2)$$

where $1 \leq j \leq n$. For monosyllabic words, j is approximately 2.5 [4,27].

Equation 3 relates words in isolation to words in sentences,

$$P_s = 1 - (1 - P_w)^k, \quad (3)$$

where P_s is the probability correct for words in sentences, P_w is the probability correct for words in isolation, and k is a free parameter (greater than one) reflecting the degree of predictability or context of the sentence materials. For low-context sentence materials such as the IEEE/Harvard set [5], k is approximately 1.14. For sentence sets with a higher degree of predictability (e.g., CUNY sentences), k is approximately 4.5 [4].

Combining equations 2 and 3 with appropriate estimates of j (≈ 2.5) and k (≈ 1.14), and substituting P_c for P_p gives

$$P_s \approx 1 - (1 - P_c^{2.5})^{1.14}, \quad (4)$$

where P_c is the proportion correct for consonants and P_s is the proportion of words correct for IEEE/Harvard sentences.

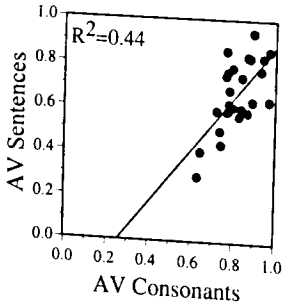


Figure 9. Relation between AV consonant recognition and sentence recognition.

Figure 9 shows the relation between AV consonant and AV sentence recognition for hearing-impaired subjects. Application of k- and j-factors appropriate for IEEE sentences (as described above) would adjust the range of consonant recognition scores to better match the range of observed sentence scores, but does nothing to reduce the variability across subjects. To accomplish this, individual differences in context-usage must be taken into account. Studies to estimate k- and j-factors for individual subjects, as opposed to sets of speech materials, are currently underway. Additionally, other measures of word and sentence context effects are being explored.

SUMMARY

Predictions of AV speech recognition ultimately depend on an understanding of how lexical access is affected by information provided by auditory and visual sources, the processes by which information is integrated, and the impact of top-down contextual constraints. Our efforts thus far to evaluate these factors in individual subjects have been limited mainly to consonant recognition, the recognition of certain prosodic contrasts, and segmental integration skills. Ongoing efforts to expand this work to include vowel recognition, sentence integration, and semantic context usage will no doubt improve our overall understanding of AV speech perception.

REFERENCES

[1] Massaro, D.M. (1987). *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*. Hillsdale, NJ: Lawrence Earlbaum Assoc.

[2] Braida, L.D. (1991). "Crossmodal integration in the identification of consonant segments," Quarterly J. Exp. Psych. 43, 647-677.

[3] Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by Eye: The Psychology of Lip-Reading* (pp. 3-52). Hillsdale NJ: Lawrence Erlbaum Associates.

[4] Rabinowitz, W.M., Eddington, D.K., Delhorne, L.A., & Cuneo, P.A. (1992). "Relations among different measures of speech reception in subjects using a cochlear implant," J. Acoust. Soc. Am. 92, 1869-1881.

[5] IEEE (1969). IEEE recommended practice for speech quality measurements. *Institute of Electrical and Electronic Engineers*, New York.

[6] MacLeod, A. and Summerfield, Q. (1987). "Quantifying the contribution of vision to speech perception in noise," British J. Audiol. 21, 131-141.

[7] Miller, G.A. and Nicely, P.E. (1955). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. 27, 338-352.

[8] Grant, K.W., & Walden, B.E. (1993). "Evaluating the articulation index for auditory-visual consonant recognition," J. Acoust. Soc. Am. 94, 1887.

[9] Walden. B.E., Prosek, R.A., & Worthington, D.W. (1974). "Predicting audiovisual consonant recognition performance of hearing-impaired adults," J. Speech Hear. Res. 17, 270-278.

[10] Wang, M.D., Reed, C., and Bilger, R. (1978). "A comparison of the effects of filtering and sensorineural hearing loss on patterns of consonant confusions," J. Speech Hear. Res. 21, 5-36.

[11] Erber, N.P. (1975). "Auditory-visual perception of speech," J. Speech Hear. Res. 40, 481-492.

[12] McGrath, M. & Summerfield, Q. (1985). "Intermodal timing relations and audio-visual speech recognition by normal-hearing adults," J. Acoust. Soc. Am. 77, 678-685.

[13] Crystal, D. (1979). "Prosodic Development," In P. Fletcher & M. Garman (Eds.), *Language Acquisition* (pp. 33-48). Cambridge: Cambridge University Press.

[14] Kent, R., & Read, C. (1992). *The Acoustic Analysis of Speech*. San Diego: Singular Publishing Group.

[15] Lieberman, P. (1965). "On the acoustic basic of the perception of intonation by linguists," Word, 21, 40-54.

[16] Lehiste, I. (1976). Suprasegmental features of speech. In N.J. Lass (Ed.), *Contemporary Issues in Experimental Phonetics* (pp. 225-239). New York: Academic Press.

[17] Christie, W.M. (1974). "Some cues for syllable structure perception in English," J. Acoust. Soc. Am. 55, 819-821.

[18] Streeter, L. (1978). "Acoustic determinants of phrase boundary perception," J. Acoust. Soc. Am. 64, 1582-1592.

[19] Scott, D.R. (1982). "Duration as a cue to the perception of a phrase boundary," J. Acoust. Soc. Am. 71, 996-1007.

[20] Smith, M.R., Cutler, A., Butterfield, S., & Nimmo-Smith, I. (1989). "The perception of rhythm and word boundaries in noise-masked speech," J. Speech Hear. Res. 32, 912-920.

[21] Risberg, A. (1974). The importance of prosodic speech elements for the lipreader. In H.B. Nielson & B. Klamp (Eds.), *Visual and Audiovisual Perception of Speech VI. Danavox Symposium* (pp. 153-164). Scand. Audiol. (Suppl. 4).

[22] Risberg, A., & Lubker, J.L. (1978). "Prosody and speechreading," Speech Transmission Lab - Quarterly Progress Status Report, 4, 1-16.

[23] Boothroyd, A. (1988). "Perception of speech pattern contrasts from auditory presentation of voice fundamental frequency," Ear and Hearing, 9, 313-321.

[24] Grant, K.W. and Walden, B.E. (1992). "The transmission of prosodic information via selected spectral regions of speech," J. Acoust. Soc. Am. 92, 2300-2301.

[25] Montgomery, A.A., Walden, B.E., Schwartz, D.M., & Prosek, R.A. (1984). "Training auditory-visual speech reception in adults with moderate sensorineural hearing loss," Ear and Hearing, 5, 30-36.

[26] Walden, B.E., Busacco, D.A., & Montgomery, A.A. (1993). "Benefit from visual cues in auditory-visual speech recognition by middle-aged and elderly persons," J. Speech Hear. Res. 36, 431-436.

[27] Boothroyd, A. and Nittrouer, S. (1988). "Mathematical treatment of context effects in phoneme and word recognition," J. Acoust. Soc. Am. 84, 101-114.

CAN THE DEFINITION OF EACH SPEAKER BE EXPECTED TO COME FROM THE LABORATORY IN THE NEXT DECADES?

Francis Nolan

Department of Linguistics, University of Cambridge, UK

ABSTRACT

The symposium of which this paper is a part address questions that arise from speaker identification in forensics. The focus is on what can and cannot be expected of forensic speaker identification, and on directions for future research. The first three sections of this paper set the background for the symposium, and the subsequent sections suggest possible innovations. The overall theme is the need to explore new methods of imposing structure on the data used in speaker identification in order to understand variation. These methods include alternative phonological models, and a more explicit role for articulatory modelling.

1 INTRODUCTION

The primary concerns of phonetics have been to do with the realisation of language in the sound medium, but the scope of phonetics is much wider. A broad view of phonetics might see it as the discipline which answers the questions 'what can we tell when a person speaks, and how?' As soon as someone speaks, listeners are able to infer a wide variety of information other than that contained in the linguistically encoded 'message'. Much of that information is about the producer of the message. Listeners can infer (with a fair degree of reliability) the sex of the speaker, they can induce information about his or her health, and they can often identify the speaker as a person previously heard.

This last ability, the inference of identity, must lead us to assume that information about individual identity is convolved with the other information in the speech signal. This conclusion emerges too from other areas of the phonetic sciences: the difficulties of creating a reliable speech recognition system which is speaker-independent demonstrate that significant speaker-specific information is blended into the acoustic speech signal.

For automatic speech recognition, this speaker-specific information is unwanted noise, to be neutralised if at all possible. But in another domain, that of *speaker* recognition, it is the raw material, the structured variability and underlying regularity of which need to be determined, just as phonetics has done for the linguistically determined aspects of the speech signal. Applications of knowledge about speaker-characterising features of speech include Automatic Speaker Verification, which a massive market awaits in fields such as telephone banking, and, more controversially, forensic speaker identification. The latter provides the focus for this session.

2 DEFINING EACH SPEAKER

In 1934 Twaddell [1] cited Bloomfield as saying 'The physical (acoustic) definition of each phoneme of any given dialect can be expected to come from the laboratory within the next decades'. With the hindsight of six decades of acoustic speech analysis such a statement, if intended literally, might be seen as betraying a certain naivety, not least about the relation between phonological categories and the physical signal.

On the other hand there is a real, practical sense in which Bloomfield's prophecy has been fulfilled. Not only do we have, thanks both to extensive acoustic analysis and to advances in the acoustic theory of speech production, a very good understanding of the acoustic properties which realise phonemes of different types, but we also have advanced statistical models (such as HMMs) which, in some cases speaker-independently, can learn to recognise the realisations of each phoneme in the speech signal.

Whatever the correct assessment of Bloomfield's statement, it may be the case that we are in a similar position *vis-à-vis* individual speaker quality today as Bloomfield was in the 1930s in relation to phonemic quality. We have an analytic construct, speaker quality, for which (if we adopt an appropriately 1930s terminology) we have behavioural evidence, in the ability of listeners to identify speakers. We even have a fairly well worked out phonetic model, parallel to that provided by traditional phonetic analysis for the phoneme, of at least part of speaker quality: Laver's (1980) framework for the analysis of voice quality [2], for instance, can be seen as a model of that part of speaker quality which is under the speaker's control. But we do not have a comprehensive answer to the question: 'What defines an individual in the acoustic signal?'

As a starting point for this session on forensic speaker identification then we can therefore re-phrase Bloomfield's dictum, and debate the proposition that 'The definition of each speaker can be expected to come from the laboratory in the next decades'.

3 TWO STRANDS

There are perhaps two strands to consider in this proposition. The first is the nature of 'speaker quality'. What dimensions are involved? How much variation does an individual exhibit? And, most crucially, does each individual human being occupy a unique location in acoustic space? Or is there instead a significant degree of 'overlapping', by which an individual shares part, or indeed all, of his or her location with others, rather as the English phonemes /ɛ/ and /æ/ may share the phonetic realisation [æ] in words such as *well* and *gag* respectively, as a result of contextually induced allophonic variation? The answers which emerge to questions such as these about speaker quality will inform the issue of what we might mean by 'the definition of a speaker'.

The second strand to the proposition is the implication that it is specifically in the laboratory that progress will be made towards finding the definition of a speaker. Of course, if consideration of the first strand results in the conclusion that we have no viable theory of speaker quality, and if we take a somewhat purist view of empirical science to the effect that measurements and experiments cannot usefully be carried out in the absence of testable hypotheses generated by a theory, then there is no point in going into the laboratory. But it seems unlikely that both these negative conditions would hold. We probably do have the beginnings of a theory of speaker quality; and even if not, it may be that what we most need in this field are large-scale, pre-theoretical, 'taxonomic' studies of between- and within-speaker variation. If we accept that work in the laboratory is appropriate, we can then indulge in informed speculation about the kind of analyses and methodological developments which are likely to bring greater understanding of speaker quality.

Although Bloomfield's proposition is here newly adapted to speaker recognition, the debate which its adaptation encapsulates is already underway. Baldwin and French [3] address essentially the same proposition. Interestingly, the two authors arrive at diametrically opposed views. In Chapter 3, French writes 'For various theoretical reasons, I cannot forsee a day when phoneticians will be able to identify a speaker with the degree of certainty associated with the matching of fingerprints or DNA profiles' (p.62). Baldwin, in the final chapter, despite having taken throughout the book a generally negative stance towards the present-day contribution of acoustic phonetics to forensic speaker identification, writes more optimistically of the future: '... I positively believe there will one day be a "voiceprint", i.e. a print-out from some sort of, not necessarily electronic, device which will be able uniquely to identify an individual speaker' (p.126).

The fact that the authors hold disparate views on such a fundamental matter is, as the foreword to the book (p.iv) points out, potentially productive if the disagreements are rationalised. The shortcoming is perhaps that so little is said about the grounds for the disagreement that it is not clear what the framework for any discussion might be. It is hoped that this session will help to set out the parameters of such a discussion.

4 FUTURE PROGRESS

The other speakers in this symposium provide clear summaries of problems and

methods in forensic speaker identification, and point to ways of improving the methods. Braun focuses on the phonetician's role, while Broeders discusses how much solutions from Automatic Speaker Verification might contribute to the forensic task. Hollien presents a framework of requirements for objective speaker identification, in which nevertheless the ultimate decision is a human one.

Perhaps, though, because all three speakers are closely involved in the day-to-day work of speaker identification, they have chosen to concentrate on improvements and extensions to current approaches and conceptualisations. What I will try to do in the following sections is to suggest more radical departures from current thinking in the area.

The theme which links this speculative train of thought is the need to find new ways of coping with variability, a problem which Broeders draws attention to in his Section 2. No two utterances are identical, even if they are by the same speaker, and so speaker identification cannot proceed on the basis of rejecting a 'match' every time a difference is detected. Where there is a difference we need to understand what lies behind it. To achieve this understanding it will be argued that we need the most comprehensive account available of how phonetic material is structured by phonology, and we need to take advantage of those constraints on variability which are imposed by an individual's speech production mechanism.

5 PHONOLOGICAL THEORY

Like many areas of applied phonetics, speaker identification might be regarded by many contemporary phonologists as stuck in a time-warp. There is little evidence of any view other than one which assumes that phonological analysis is done in terms of phonemes, which receive a variety of realisations according to segmental context. Dialect spotting, measurement of acoustic values, and even more specialised concerns such as coarticulation, are discussed within this framework. Prosody is generally treated as an optional accessory, to be ignored most of the time, but, if dealt with, then treated not as an aspect of the phonological system but purely as an unstructured physical aspect of the signal, in terms of parameters such as mean fundamental frequency or overall perceived pitch. Conspicuously absent from work in speaker identification are concepts and representations taken from schools of phonology such as Autosegmental, Metrical, Dependency, and Government.

Does this matter? After all, the phonetic stuff is there, and the task is to distil out the speaker-specific essence from the signal; and it is hardly going to be important what phonological model one adheres to. But in fact it may matter, because one's phonological prejudices may influence where and how one looks for the speaker-specific essence.

For instance, if one's phonological model incorporates a prosodic hierarchy, with syllables and feet at the bottom, and intonational phrases at the top, it may lead one to be more choosy as regards which events one treats as phonetically equivalent than if one sees speech as a linear string of (phoneme-sized) beads. English /t/ is simply /t/, but an awareness of prosodic structure might restrain one from treating all the vowels in *debility* /dɪbɪlɪtɪ/ as equivalent. Again, speaker identification must have at its disposal accurate descriptions of dialect or accent differences within a language. Some of these are extremely complex, such as English plosive allophony (glottalisation, 'flapping', etc.; see e.g. [4] and [5] for Metrical and Government accounts), and adequate descriptions may only be possible in models embodying a rich phonological mechanism, including syllables, feet, prominence relations, and so on.

Similarly, without a well worked out model of intonational phonology, potentially speaker-specific phenomena may escape investigation. For instance, a search for differences between speakers in the realisation of prosodic categories only arises if one incorporates some prosodic phenomena into the phonological description. If one's phonological model incorporates an autosegmental-metrical representation of intonation in terms of high (H) and low (L) tones, for instance, as in much recent intonational work, it is more likely that the question will arise as to whether speakers may differ in their preferred alignment of the tones to the segmental material. Or an intonational model which includes the notion of downstep will allow of the question whether some speakers use downstep more than others, and of those who do use it, whether there are differences in the implementation of it.

The general point is that the evolution of phonological theory is driven, at least in part, by imperfection in the fit of previous models to the facts concerning the sound structure of language. In order to understand variation in the speech signal, the forensic phonetician needs the best available model. Good forensic phonetic practice is currently immeasurably better than that of the sound engineers mentioned in Braun's contribution to this symposium who compare waveshapes with no regard to the identity of the vowels those portions of signal are realising. But the possibility of further progress through the adoption of more sophisticated phonological models needs to be explored. In a sense, then, some of the means for progress towards the definition of the speaker lie outside the laboratory.

6 ACCENT ANALYSIS

Much of the contribution of the forensic phonetician today is of a kind which pre-dates instrumental analysis of speech samples. It is, in effect, practical dialectology; and when the question is whether two samples of speech were produced by the same human being, a sensible first step is to see whether they manifest the same linguistic properties by comparing their pronunciation. If the pronunciations are grossly different, the samples are unlikely to get as far as the forensic phonetician – those responsible for the legal side of the case will use their own judgment and conclude that a speaker with a London accent in one sample is unlikely to be the same individual as the Scottish speaker in another. The role of the phonetician will normally be to adjudicate in cases where the samples are already superfically similar. The specialist skills which a traditional phonetic training provides will allow the phonetician to notice, and classify, differences between samples which are more subtle than would be noticed by most untrained listeners. Although there are many problems to do with the linguistic variability which the speech of one person undergoes as a result of factors such as style, speaking context, and accommodation to interlocutors, close phonetic analysis can often reveal patterns of difference between samples which make it unlikely that they come from the same source.

How far one can go in the opposite direction and treat the absence of differences in pronunciation as evidence pointing towards the samples coming from the same speaker is a contentious matter. It rests, ultimately, on the question of how finely the 'isoglosses' of a dialect map can be drawn. The strongest position (see e.g. [6]) is that each individual speaks an 'idiolect'. However, even if it could be demonstrated no two individuals share a complete set of linguistic phonetic properties, it is doubtful whether the finite (and often short) samples available in forensic cases would allow a safe extrapolation from 'sameness of sample' to 'sameness of speaker'. As argued in [7] linguistic phonetic sameness may licence conclusions of 'possibly' the same, but not 'probably'.

Nonetheless linguistic phonetic analysis is an important element in the forensic phonetic approach, and must surely enter into 'the definition of the speaker'. Since such analysis requires a well-trained phonetic ear, surely it is unrealistic to expect progress on this aspect of the speaker's definition to come from the laboratory?

It is in fact far from unrealistic. For one thing, in the everyday practice of forensic phoneticians, acoustic analysis already supplements auditory analysis of what are, in effect, dialect features. But looking at the issue more fundamentally, auditory phonetic descriptions are necessarily abstractions, and rely on categories identified by selected perceptually salient characteristics of a sound. It has been shown in other phonetic areas that impressionistic descriptions may be only partially accurate. Production studies, for instance, have shown (cf. [8]) that segmental 'assimilation' can be an articulatorily gradient phenomenon, contrary to the implication of many segmental descriptions. Similarly, doubt has been cast [9] on speech error work based on

impressionistic observation. EMG monitoring of muscle activity reveals that far from involving only discrete segmental effects, speech errors range along a continuum of muscle activation.

The analogies of these findings as far as accent and dialect are concerned must as yet be a matter of speculation; but perhaps they are to be found in the effects which lie outside the static 'frozen frame' ([10], p.108) on which segmental phonetic description is based, and which implicitly or explicitly focuses on a characteristic 'target' for a segment. So whilst we might traditionally describe two accents as having a 'dark' syllable-initial realisation of /l/, and a close-mid realisation of /e/, the description of the accents might be refined by the discovery of systematically different coarticulatory treatments when the segments are juxtaposed. Fine details of intra- and inter-syllabic timing, elusive to auditory analysis, might be highlighted. And differences in intonational features, such as the alignment of pitch peaks relative to segmental material mentioned in the previous section, might be revealed as accent-specific. There is no evidence that the description of an accent is exhausted by what the ear can hear in the context of a classical phonemic framework.

Such instrumentally-mediated detail would, in a sense, provide a finer 'mesh' for the dialect grid which traditional phonetics imposes on the speech community. Given that the forensic phonetician is likely to be sent samples which are at least fairly similar in accent, the finer that mesh is, the more he or she can add to what the lay person can hear. Different speakers might be discriminable by 'sub-auditory' secondary dialect differences too subtle to be consciously manipulable; and if even the fine mesh fails to separate the samples, the odds against them being from the same individual are shortened (though one must still guard against the temptation to claim they are the same speaker merely because, in some more precise way than before, the samples share the same accent).

Another issue is whether the task of the phonetician in 'dialect spotting' will, in the future, be automated. As far as I am aware, little work has been done towards this goal. In the context of a multi-dialect automatic speech recognition system, however, [11] reports a technique which automatically assigns a speaker to one of four major English regional accent groups on the basis of several pre-determined utterances. These are chosen to contain diagnostics for the different accents. In a sentence containing the words *father*, *path*, and *car*, a similar vowel quality for all three ([ɑ:]) suggests Southern British, different for all three ([ɑ], [æ], [ɑᴵ]) General American, and so on. The crucial events are identified in the input utterance by time-warping it to a segmented reference utterance, but all spectral comparisons are internal to the input signal, so that no normalisation for individual speaker characteristics is needed prior to the accent decision.

This method requires the production of agreed speech material, and so even if its accuracy and discriminatory ability were vastly increased it would not threaten the role of the phonetician, whose knowledge and skill often permit an assessment of dialect similarity or difference on the basis of short samples of differing content. But in future decades a semi-automated and vastly improved version might have a role to play where amounts of material are large. Orthographic transcripts of long recordings could be searched automatically for words with dialect-sensitive vowels. These words could be located automatically in the acoustic signal by ASR techniques, with manual correction if necessary. Acoustic parameters would be extracted, and used firstly for 'sample-internal' dialect spotting, as described above; and secondly for direct comparison with values from another sample.

To suggest a procedure of this kind is not to ignore the difficulties – the effects of prosody, segmental context, and so on – but given the extremely powerful signal processing techniques available even today it is not too early to speculate as to how they might be applied in a phonetically informed way to the problem of speaker identity.

In this section, then, I have suggested that the laboratory should provide new approaches to the definition of accent characteristics, and to the detection of accent, which have up to now been a field predominantly for auditory phonetics.

7 VOCAL TRACTS SHAPE

Current research on defining the speaker involves measuring values such as formant frequencies associated with particular phonological events, and deriving estimates of between- and within-speaker variation. This is a vital kind of data collection, and needs to be pursued on as large a scale as possible. But the work tends to treat the measured values as independent, and as varying in a purely statistical fashion, rather than as varying in a lawful way governed by the nature of their source. Only by referring back to the source can the significance of variation begin to be assessed.

To put it another way, an individual's vocal tract shapes, and imposes strict (though by no means absolute) constraints on, the sound he or she can produce; and by considering measured acoustic values not in isolation but in relation to their source we may gain a more powerful grasp on variability.

It is already possible to estimate vocal tract lengths from formant frequencies, and, using for instance linear prediction, to estimate cross-sectional area functions for particular vowels. We can also use vocal tract synthesis models to compute formant frequencies for different tube shapes, and we can restrict the range of tube shapes broadly to those which are anatomically plausible. Source inference, and articulatorily realistic vocal tract synthesis, may prove powerful tools in the interpretation of variation.

For instance, two tokens of a vowel taken from different recordings turn out to have similar first and second formant frequencies, but a less similar third formant frequency. What is the threshold we use to decide 'different speaker'? Though clearly one would never answer the question 'are the recordings from the same speaker' on the basis of one vowel, the 'threshold' problem arises however many factors are taken into account. At present, the threshold would have to be a purely statistical one: from databases, we might estimate that a speaker's F3 frequency will vary by a given percentage for a particular vowel. But if the first recording contains enough material for a reasonably accurate vocal tract model of the speaker to be derived, it may be possible to say something like 'it is highly unlikely that the source of the first recording could achieve the specific combination of formant frequencies found in the vowel from the second recording.'

In this way acknowledging the mechanism producing the speech would allow us in our decision making to go beyond purely statistical treatments of the variability of acoustic data .

Such progress, if it is made, will come not simply from the laboratory, in the sense of empirical discoveries, but from the application of the acoustic theory of speech production.

8 ARTICULATION MODELS

A greater general awareness of the source of the speech signal may permit other novel insights. Speech does not originate from a tube producing a static set of formant frequencies, but from a dynamic complex of articulators working in close coordination to achieve the phonological requirements of an utterance. Generally we do not assume that every phonetic dimension is crucial at every instant in an utterance. Rather, we hypothesise that some 'target' events are more crucial than others. If this is the case, speakers may evolve individual articulatory strategies for achieving and moving between such targets. Such a view is implicit in studies of coarticulatory idiosyncrasy [12], [13].

But the relation between phonological requirements and articulation is not theory-neutral; nor are potential sources of between- and within-speaker variation totally independent of theoretical assumptions; and so it would be negligent for researchers in speaker identification to ignore theoretical and practical developments in articulation modelling.

Perhaps the most radical current view of the relation between phonological specifications and speech is Articulatory Phonology [14,15], whose phonological primitives are 'gestures' such as 'labial closure'. The notion of a gestures is taken from work on the control of skilled actions in a framework called 'Task Dynamics' [16]. Gestures, unlike features or segments, inherently possess dynamic characteristics, and they permit the computational modelling of articulator movements. A 'gestural score' specifies the relations between gestures needed for

particular utterances, and is seen both as a lexical-phonological representation, and a representation of the implementation of the utterance. It is possible to synthesise the speech signal from the gestural score via the task dynamic modelling of articulatory interaction and a vocal tract synthesiser, making it possible directly to predict the acoustic effect of constellations of gestures.

In Task Dynamics coordination of gestures is not represented straightforwardly in the time domain, but in the 'phase-plane', which depends on viewing articulatory movements as oscillations (damped or undamped). It has been claimed ([16], p.41 ff) that representation in the phase-plane may reveal consistency of gestural organisation across differences of rate and stress which is obscured by representations in the time domain. The phase-plane might, in effect, reduce apparent within-speaker variation in the timing of events. If the phasing turned out to differ across speakers (as some studies have implied, e.g. [17]), a better separation of speakers might be achieved than is possible in purely acoustic data.

Importantly many kinds of phonetic variation associated with changes in rate and style, and which are often modelled as the output of phonological rules, are said to emerge automatically from the gestural account as a result of general processes of increased overlap between gestures (presumably some inter-gestural phasing relations do change) and reduction in magnitude of gestures [14]. In the sentence 'He said *a fan could surprise you*', which might be realised as [...ə fæn kəd səpraɪz jʊ] or, more rapidly, as [...ə fæŋ kəd spraɪʒʊ], the apparent change of the alveolar nasal to a velar would result from the velar gesture for the plosive overlapping the alveolar gesture and masking its acoustic consequences (cf. [18]); the 'deletion' of the first syllable of 'surprise' would be an automatic consequence of the labial closure overlapping the [s]; and the [ʒ] would result from the competing effects of two gestures (for [z] and [j]) overlapping and competing for the same articulator.

Faced with two very different acoustic stimuli, let us say a fast and a slow utterance of '*fan could*', or '*surprise you*', from different recordings, it might be possible to determine whether the different acoustic properties of the second one (changed formant trajectories, durations, etc.) are compatible with it being a speeded up version of the first one, or are the product of a different articulatory mechanism. That is, if we have enough speech at one rate in the first recording to be able to replicate the speaker using articulatory synthesis, and if rate change turns out to involve similar articulatory strategies across speakers, we could change the rate of articulation of the synthesised version of the first and test whether the acoustic properties of the second recording are compatible with it having been produced by the same speaker.

The implementation, let alone practical application, of such a procedure, lies a long way off. In particular the inference of articulatory activity from the acoustic signal, which is a prerequisite to the suggestions above, would require a very sophisticated method probably involving analysis-by-synthesis. But such an approach is not unimaginable, as it would have been until relatively recently, and it is the kind of ambitious goal which might stimulate fundamental laboratory research towards taming the variability problem in speaker identification.

9 TWINS SPEAK

The linchpin of any investigation is control. If we are to understand the ways in which speakers differ, and we assume that the differences can broadly be categorised as dependent on 'organic' and 'learned' factors, it would be useful to be able to control one or other of these factors. Nature provides such a control in the case of identical twins, for whom it is a reasonable hypothesis, though not a certainty, that they will have extremely similar vocal mechanisms. This natural control case must surely figure prominently in future research into the definition of a speaker.

Recently a pilot study of three pairs of university-age identical twins, brought up in shared environments, was carried out in Cambridge by Tomasina Oh. The twins recorded lists of words with /l/ and /r/ before a range of vowels, as in [14], one of the objects being to discover if members of a pair had different coarticulatory strategies. Interestingly, consistent differences of various kinds did emerge between the members of each pair. In the most striking case, /r/ was realised by different articulations: as [ɹ] by one member and as [ʊ] by the other, with consequent acoustic and coarticulatory differences. In another pair, one member consistently palatalised /l/ more than the other. In the third pair, whose words showed in general a high degree of similarity, one twin showed greater fronting of /u:/ after /l/ than the other. On the other hand the prediction that there might be distinct coarticulatory strategies was not borne out in general.

This particular study demonstrates that identical twins do not have identical speech. But, more generally, studies of twins provide the possibility of studying the extent to which speaker characteristics are behavioural rather than anatomical.

10 CONCLUSION

It is certain that progress towards the definition of the speaker will involve the laboratory. What I have argued here is that to tackle the central problem of between- and within-speaker variability, it will not be sufficient (though it will be necessary) to carry out acoustic measurement studies on large populations, and to continue only to apply current techniques of analysis. Rather, theoretical and technical innovations of various kinds are needed; and our goals need to be, perhaps, more ambitious than at present.

In what sense 'the definition of each speaker can be expected to emerge from the laboratory' will have to wait for an answer until the results of such innovations begin to materialise. Whilst I share French's more cautious view that speaker identification will never be like fingerprinting (section 3 above), I believe we are far from having reached the bounds of what is possible in speaker characterisation.

REFERENCES

[1] Twaddell, W.F. (1935) On defining the phoneme. *Language Monograph* 16.
[2] Laver, J. (1980) *The Phonetic Description of Voice Quality*. Cambridge: CUP.
[3] Baldwin, J., and French, P. (1990) *Forensic Phonetics*. London: Pinter.
[4] Gussenhoven, C. (1986) English plosive allophones and ambisyllabicity. *Gramma* 10, 119-41.
[5] Harris, J. & Kaye, J. (1990) A tale of two cities: London glottaling and New York tapping. *Linguistic Review* 7, 251-74.
[6] Baldwin, J. (1979) Phonetics and speaker identification. *Medicine, Science and the Law* 19, 231-32.
[7] Nolan, F. (1991) Forensic phonetics. *Journal of Linguistics* 27, 483-493.
[8] Nolan, F. (1992) The descriptive role of segments: evidence from assimilation. In: G. Docherty and D.R. Ladd (eds.), *Laboratory Phonology 2*, 261-280. Cambridge: CUP.
[9] Mowrey, R.A. & MacKay, I.R.A. (1990) Phonological primitives: Electromyographic speech error evidence. *J. of the Acoustical Society of America* 88 (3), 1299-1312.
[10] Laver, J. (1994) *Principles of Phonetics*. Cambridge: CUP.
[11] W.J. Barry, C.E. Hoequist, and F. Nolan (1989) An approach to the problem of regional accent in automatic speech recognition. *Computer Speech and Language* 3, 355-66.
[12] Su, L-S., Li, K-P., and Fu, K.S. (1974) Identification of speakers by use of nasal coarticulation. *Journal of the Acoustical Society of America* 56, 1876–82.
[13] Nolan, F. (1983) *The Phonetic Bases of Speaker Recognition*. Cambridge: CUP.
[14] Browman, C.P. & Goldstein, L. (1989) Articulatory gestures as phonological units. *Phonology* 6, 201-251.
[15] Browman, C.P. & Goldstein, L. (1992) Articulatory Phonology: an overview. *Phonetica* 49, 155–80.
[16] Kelso, J., Saltzman, E., & Tuller, B. (1986) The dynamical perspective on speech production: data and theory. *Journal of Phonetics* 14, 29-59.
[17] Perkell, J. & Matthies, M. (1992) Temporal measures of anticipatory labial coarticulation for the vowel /u/: within- and cross-subject variability. *JASA* 91(5), 2911–25.
[18] Byrd, D. (1992) Perception of assimilation in consonant clusters: a gestural model. *Phonetica* 49, 1–24.

THE FUTURE OF SPEAKER IDENTIFICATION: A MODEL

Harry Hollien
Institute for Advanced Study of The Communication Processes
University of Florida, Gainesville, FL USA

ABSTRACT

One of the major areas within Forensic Phonetics, and to some extent Phonetics in general, is that of speaker recognition--and especially speaker identification. To date, most of the problems attendant to this issue have escaped resolution. The chaos here may be due to the fact that several types of professionals (Phoneticians, Engineers, Psychologists and the Police) all are working in the area but in a fairly uncoordinated manner. The strengths and abilities each bring to it are incomplete and often are further degraded by their weaknesses. The result is that no robust method of speaker identification currently exists. The following presentation will provide a review of the basic problems in the field, its boundaries, past approaches to the problem, the strengths and limitations of the relevant specialists and a model which could lead to its resolution in "the next decades." Forensic Phoneticians are central here; however, the model specifies that objective means must be employed if a valid and effective speaker identification system is to become a reality.

1. INTRODUCTION

One of the fairly new--and certainly exciting--areas within the Phonetic Sciences is that of Forensic Phonetics. Specialists with-in this area are making significant contributions both to relevant research and in response to problems faced by members of the legal and law enforcement communities. This interface ranges from speech enhancement and/or decoding to tape authentication, from detection of stress in voice to the vocal cues which signal intoxication. However, of all the problems encountered, that of speaker identification-is probably the most challenging and (perhaps) the most important. For one thing, it involves issues that are fundamental to the Phonetic Sciences; indeed, it appears appropriate to state that research here should claim a measurable portion of our time and energy. Second, it holds substantial social significance.

By now we should have defined and structured the issue. If we could not "solve" it, we should have, at least, approached it in a coordinated manner so that the relevant relationships could be systematically researched. Unfortunately, we have not done so. As a discipline, our members have tended more to react to positions taken by, or requests from, members of other disciplines rather than to have organized the necessary models and carried out appropriate research. Whether we like it or not, this area of our field is in near chaos.

In the preceding paper, Nolan has provided most of the basic definitions relative to the speaker identification task and has outlined certain of the specific problems and difficulties we face. In the effort to follow, an attempt will be made to supplement his perspective and provide a model which could lead to a solution "in the next decades". To do so, several issues must be addressed; they include reviews of 1) the bases for speaker identification; is it possible to do it in the first place? 2) the boundaries of, and approaches to the problem; which of the available approaches may ultimately lead to a successful resolution? 3) the other classes of professionals who are relevant to the area; what are their responsibilities and what contributions can they make? 4) guidelines for future speaker identification efforts; i.e., the proposed model.

2. BASES FOR SPEAKER IDENTIFICATION

Two related questions may be asked about the identification of speakers by voice. They are: 1) does each human speak in a manner so idiosyncratic that, overall, he/she is different from all others and 2) is inter-speaker variability always greater than intra-speaker variability? The answer to both of these questions is a resounding "probably not!" Worse yet, it is to the discredit of our discipline that we have not already researched these fundamental issues to any great extent. Indeed, nearly all the authors of the over 700 presentations on speaker recognition listed by Hollien and Alderman [1] have addressed only narrow issues or relationships--and many of them involve "application." Application? At first glance it would appear counterproductive (if not ludicrous) to attempt the "solving" of a problem before its nature is understood. None-the-less, this situation functionally constitutes the present State-of-the-Science re: speaker identification.

What is needed, of course, is a major research thrust in the basic areas. For one thing, researchers should attempt to determine if talkers actually do exhibit unique enough characteristics to permit universal speaker identification to be developed. At the very least, an effort of this type would establish the limits and boundaries of the problem and, possibly, lead to techniques and/or procedures which would permit a valid, if restricted, response.

In reality, there is little-to-no possibility that such a massive effort would be supported by any agency or group. This is surprising as there is no question but the need for valid speaker identification and verification methods is a critical one. It exists in nearly every sector of society. What is lacking is the foresight by any of the relevant agencies to see beyond an end-product. Sadly enough, the need is for basic research; about all that will be supported is "product development."

Given the unlikelihood that basic speaker identification issues can be addressed in any meaningful way, only a single recourse appears available. That is, all that may be possible is to generate a working model by the synthesis and interpolation of current information as supplemented by research conducted on a piece-meal basis. Actually, some of the necessary relationships have been established--at least enough of them for researchers to attempt advances in this area. Useful data already can be found in a number of published articles and reports; four books (Baldwin and French, 2; Hollien, 3; Küenzel, 4 and Nolan, 5) provide summaries of most of the important relationships; further, they suggest some useful models. It now appears evident that the two questions cited above must be answered in the negative **only** if a binary answer is required. It also appears evident that a given talker may be differentiated from other individuals within specific sets of speakers if a critical number of his or her features are measured and appro-priate metrics (in multidimensional space) are established and compared. On a simpler level, it appears that establishing speaker profiles may very well be of merit. What appears both needed and realistic is completion of a number of investigations in which attempts are made to identify and validate those individual parameters--plus constellations of parameters--which are robust to the task. Subsequently, the resistance of these individual parameters and profiles to forensic type degradation can be studied. However, it should be noted that this element within the model does not reject human decisions for some mathematically derived metric or group of metrics. The fundamental focus here still would be on human performance and it's assessment by humans.

3. THE BOUNDARIES OF SPEAKER IDENTIFICATION

As was pointed out in the prior paper by Nolan (see also Hollien, 3 for a definition), speaker identification is only one element within the general rubric of speaker

recognition. Here the task is to determine if a given (and known) speaker is the same person as the one who produced the target utter-ances (i.e., the "unknown" talker). This task is a very difficult one primarily due to the nature of speech and the situation within which it exists. The speech will be noncontemporary; all sorts of channel and speaker distortions may (and usually do) exist. For example, 1) there may be many competing speakers, 2) the unknown talker may have provided only a limited speech sample, 3) the process is an "open" one (i.e., the unknown may not be in the suspect pool), 4) speakers usually are uncooperative, 5) poor recordings may exist, and so on. In this milieu, the scientist (or practitioner) will have little control over the available signals.

On the other hand, speaker verifica-tion is a process where an attempt is made to authenticate the identity of a given speaker by comparing his utterances to those in a closed set of voices of which he is a member (that is, unless he is an imposter). Because of the high control practitioners enjoy in this situation (a closed set, speaker cooperation, continual updates, exten-sive samples, sophisticated equipment, etc.) the challenge of speaker verification is a much less rigorous one than is that of identification. As would be expected, far more research has been carried out in the verification area (than on identification) as it is easier to manage and can lead to substantial monetary returns. What few people appear to realize is that, due to the severe challenge created by the identifi-cation task, there is little chance that even successful verification approaches can be applied to "solve" identification. It also is unfortunate that very few people understand that any method which is successful for speaker identification will simultaneously solve the verification problem.

Boundaries to speaker identification also have been established in other domains. In one case, it is the decision-making process that is controlling. There are three "entities" which can be involved in this process: 1) laymen, 2) professionals (usually Phoneticians) and non-humans (i.e., computers, other machines). These divisions, while even more complex than they seem, have essentially been defined in Courts-of-Law. They will be discussed in turn.

3a Laymen.

The Courts have pretty much established, defined and limited acceptable behavior for the first of these cohorts, i.e., laymen. Ordinarily, the process involved takes one of two forms. In the first instance, an individual who can demonstrate a close familiarity with the unknown talker is allowed to testify that he or she can recognize and identify him or her as the (otherwise) "unknown" speaker. In support, there is very good research evidence as a basis for this postulate; that is; people who really know a talker usually can identify that person from speech samples at very high degrees of accuracy. On the other hand, there is no research available which will allow predictions to be made about how often a given individual will be correct in a specific situation. Moreover, the question must be asked as to whether or not this procedure is part of the speaker identification milieu? Of course it is. While not central at all to the fundamental requisites of the area (or the model to follow), it is the responsibility of Phoneticians to study such behaviors and to define them and their limits.

The second subgroup within the untrained cohort involves people who do not know the talker but have heard him. We know from research that untrained individuals, while usually not particularly good at this task, exhibit great variation in their natural ability and that environmental circumstances may have a substantial affect in upgrading or degrading their perform-ance. Additionally, the process here often culminates in what are referred to as earwitness lineups or "voice parades." These lineups are a reality and cannot be

ignored. The procedures used in their conduct vary wildly both with respect to their nature and quality; currently, a lively controversy exists as to who should control the earwitness process in the first place--Phoneticians, the police or relevant Psychologists. Again the problems associated with earwitness lineups are rather peripheral to core speaker identification. Nevertheless, the issues here are the responsibility of the Forensic Phonetician. Relevant procedures are, and will continue to be, employed by law enforcement agencies and the courts. While they cannot be central to our model, they must be taken researched
and understood.

3b Professionals.

The second group includes trained professionals--usually Forensic Phoneticians--who are responsible for the judgements about a speaker's identity. Since the professional but rarely knows the unknown talker, their procedures must involve systematic comparisons of some type. They certainly require that a stored sample of the unknown talker, plus one for the suspect, are available. As is well known, Phoneticians often employ panels of trained and untrained auditors to perceptually judge whether a particular unknown voice was produced by the same person as was the "known" voice. This procedure usually involves direct comparisons of samples which are embedded in a field provided by foils or controls. The Phonetician uses the resulting scores to aid him or her in making decisions; machine processing also may be carried out for the same purposes. But what he or she most commonly does is listen to samples of the unknown voice plus that of the suspect (possibly within a field of foils) over and over again. This process ordinarily involves assessment of these talkers' specific features (dialect, fundamental frequency, voice quality, articulation, etc.,) one at a time. It has been shown that techniques wherein the segmentals and suprasegmentals of speech are systematically evaluated work pretty well and they do so under a variety of conditions. Nonetheless, not much is known about the efficiency or, even, the validity of these approaches. There does not even appear to be a methodological consistency among the Forensic Phoneticians who work in this area. Which of these professionals is better at it than others, what are their "hit" rates, how do the various techniques stack up against each other, how does effectiveness vary as a function of different situations? The questions are many but the answers few. Moreover, this area is absolutely central to the speaker identification process.

3c Machines.

The third approach is that of machine processing of the speech signal for speaker identification purposes. Again the procedures employed take two directions. The first involves traditional signal processing techniques such as axis crossings, HMM, LPC, Cepstral approaches and/or related methods. In the second, researchers attempt to duplicate human auditory processing of the signal; they seek out those features that auditors employ in making identification decisions, attempt to develop appropriate algorithms and, subsequently, program computers to mimic the process. These several approaches have a longer history than is generally appreciated. Early attempts at development reach back to the World War II era and are contemporary with the "voiceprint" technique (i.e., subjective pattern matching of time-frequency-amplitude sound spectrograms). Some of the early attempts were sited at government or industrial laboratories; others were commercial efforts at speaker verification. Unfortunately the thrust was primarily on system development rather than on data gathering relative to basic identification. Hence, a number of excellent beginnings were abandoned when field trials proved disappointing. Even the few sustained,

long-term programs have progressed but slowly. It must be said that even the near-magic of modern technology is not inadequate to the task when the establishment of applied techniques is required before the basic relationships are understood.

Therein lies the functional challenge to the forensic application of speaker identification--or to speaker identification procedures developed for any reason. It will be difficult to establish any kind of effective system until at least reasonable information is available about the natural boundaries of this area and the inter- and intra-speaker variability confusions resolved. Once relevant relationships here have been established, the ways by which application can be carried out also will become available.

4. PARALLEL BUT UNCOORDINATED EFFORTS

A second rather serious problem also exists in the speaker identification area. It results from the well intentioned, but sometime misguided, efforts of the three major groups of professionals working on speaker identification problems. They are the Phoneticians, Audio-Engineers and relevant Psychologists. The insularity and narrowness within each of these groups is creating a serious impediment to orderly progress in the area.

For example, the expertise of Psychologists and Phoneticians overlap in the earwitness identification area. Of course, the Psychologists appear to be almost exclusively concerned with voice parades, whereas Forensic Phoneticians have tended to downgrade this procedure as a risky one at best. It is only very recently that each of these groups has become aware of the relevant philosophies and activities of the other. Procedures here certainly would benefit from a melding of the behavioral skills/knowledge of the Psychologists (and their research/experience with eye-witness lineups) and the Phonetician's fundamental understanding of hearing and aural-perceptual speaker identification. Further, an even more active role, by Psychologists, directed at other speaker identification issues should result in better understanding of all of the behaviors involved.

A problem with even more serious consequences is that which exists between Phoneticians and relevant Engineers. Many Phoneticians are quite unwilling to extend their identification efforts beyond the traditional aural-perceptual techniques and employ modern technology. On the other hand, Engineers often view the identification process as a simple signal analysis exercise and do not seem to understand how the effects of social pressures, the enormous variability in human behavior and the vagaries of the forensic milieu itself can disrupt machine processing of any type. Accordingly, with but few exceptions, the Phonetician's computer-based efforts have been rather feeble and Engineers' attempts to fit their procedures into the real world have been equally disappointing. On the one hand, many Phoneticians refuse to accept the possibility that the only solution to the speaker identification challenge will involve the use of modern technology. Yet, the reality here is quite apparent. On the other hand, the Engineer typically cites what he or she perceives as inadequate quantitative skills on the part of the Phonetician as well as the contradictions-confusions to be found in their literature. Engineers suggest that the answer is in the signal and a good solution can be easily achieved if they only were allowed to address the problem. Perhaps so. However, if this is true, why is it that progress is relatively nonexistent when the much more malleable issue of speaker verification is considered? More important, even after decades of great effort, closure still has not been realized with respect to the challenge of speech recognition by machine. Perhaps it is because Engineers have not been willing to address problems related to speech and speakers as well as the myriad of other

distortions (environmental, channel, speaker) found in the communicative act.

Unlike the difficulties outlined in the previous section, a reasonable solution re: the differences among professionals, may be possible. That is, after nearly a half century of frustration, these groups may be realizing that they need to establish a functional interface with each other. Further, since Phoneticians are central to the problem, it would appear that they bear the primary responsibility in fostering such cooperation. Not an easy task, of course, but one that is mandatory if an effective solution is to be realized.

5. A MODEL

As stated, there currently appears to be only one reasonable solution to the challenge of identifying speakers by voice. It is to develop a machine-based system which can be used to decode and analyze the identity information contained within the speech signal in much the same manner as does the human being. The responsibility for each decision would be the same (i.e., the professional); the primary difference being that software would be substituted for neuroprocessing. One such approach has been to identify, and single out (for processing) those features which people use in this manner (Hollien, 3; Stevens, 6). For example, fundamental frequency level and variability, vocal intensity patterns, prosody plus voice and speech quality are among those elements which have been specified. Segmentals also can be included but they are a little more difficult to process on an automatic or semiautomatic basis. Nonetheless, patterns of vowel formant usage are important here as are articulatory gestures and especially dialect. The advantages of using an approach such as this one is that the data from aural-perceptual speaker identification research can be used to structure the effort; after all humans actually attend to such features and generally are reasonably successful in using them as identity cues. Perhaps even more important, auditors appear capable of carrying out this task even in the face of severely degraded listening conditions. Thus, it should be clear that, if machines can be taught to focus on these same relationships, and process them properly, a reasonable solution to the cited problem should be achievable. Certainly, a given set of procedures can be established, applied and tested; as a result, system strengths and weaknesses can be understood. It is only by this approach, or a similar one, that a valid and effective speaker identification procedure can be developed. Most important, its use would eliminate most, if not all, of the very subjective methods currently being employed. Indeed, it is difficult to understand, much less assess (on any reasonable basis anyway) the effectiveness of Forensic Phoneticians no matter how well trained, talented and motivat-ed they are. Worse yet, some of their techniques may be considered proprie-tary and, hence, cannot be assessed at all.

Please note, however, that it is not being suggested that only machine (computer) assessment of natural speech features is a viable approach to speaker identification. There probably are other elements within the speech signal that can serve as effective identify cues also. The fact that traditional signal analysis approaches have proven grossly inadequate should not preclude efforts to identify still other cues that maybe more robust. Further, it must be remembered that many assumptions must be made even if signal analyses of the natural speech feature type are employed. That is, it is not presently known just how robust each of the "natural" attributes are when environmental distortions (noise, passband, speaker distortions) are present; nor is it known how they can be combined to effect good decisions. While logic and available data will allow a few predictions to be made, it is not possible to specify just how robust each parameter will be under all (or even some) of the conditions which will occur. Nor is it known just how they should be normalized and weigh-ted within

the speaker profile. Of course, **any** signal analysis approach will suffer from these same restrictions. Hence, experiments will have to be carried out to establish these relation-ships before vectors are applied.

Is it possible to proceed even in the face of questions about speaker variability and the differential effects of speakers, recording equipment and the environment? This query probably can be answered in the affirmative if a model is established and safeguards are included in its structure. A suggested model is as follows.

1. It must be assumed first that only digital analysis of the signal will yield a method that ultimately can be established as: a) stable, b) robust, c) efficient and d) universal.

2. The ultimate decisions made must be the responsibility of Forensic Phoneticians and/or other professionals--not the machines themselves.

3. The limitations (cited in the text) must be addressed or, at least, taken into account. That is, compensation for the possibility that intra-speaker variability may exceed inter-speaker variability must be made both with respect to relatively small and very large populations of talkers. The system also must be resistive to channel and speaker distortions.

4. It must be recognized that any attempt to establish a functioning method must be programmatic in nature. That is, it is doubtful that one or even a few experiments will yield information sufficient to develop a working system; a substantial program of research will have to be carried out.

5. The parameters, features and/or vectors (within the signal) which provide the identity cues must be identified and tested. As has been implied, enough information must be gathered about each of them that their behavior can be predicted. The situations in which they are effective and not effective must be established.

6. The ability of a proposed "system" to respond to a variety of situations and challenges should be researched and system robustness inductively specified. Test selection and administration is critical to this process. There is little chance that a system designed even for limited use can be developed unless users have information about the specific types of situations to which it can be successfully applied.

7. The system must be multidimen-sional in nature. Indeed, there probably is no single (or even small group of) feature(s) that will permit a particular speaker to be identified even under the most restricted of circumstances. Further, identification of the number and class of situations in which the method will be effective will require additional analysis--and ultimately the merging of a number of features. The number and class of situations in which the method will be effective will require additional analysis--and ultimately the merging of a number of features. A profile approach should be a effective in this regard.

As may be seen, the model cited specifies that an objective (rather than subjective) approach must be taken if the speaker identification problem is to be resolved. So too must a concerted effort be mounted to permit rational decisions to be made as to what may and may not be a accomplished. This discourse should not be interpreted as one of fault-finding as many researchers have contributed materially to the corpus of information now available. Nor is fair to fault individuals for carrying out finite (rather than programmatic) projects; often the culprit was the simple lack of funding. Perhaps, the only blame to be assessed here is one which can be directed at those practitioners who make sweeping claims about their methods; some show promise but all presently are of limited scope.

The solution also demands a change in the work patterns and philosophies of the specialists involved. Anyone--including Forensic Phoneticians--who believes that good resolution will emerge solely from efforts within his or her specialty, is not being realistic. It will take the combined efforts of members from all three of the cited professions to affect a solution.

REFERENCES

[1] Hollien, H. and Alderman, G. A. (1995) Speaker Identification and Recognition: Current References, *Miszellen Beiträge zur Phonetik und Linguistik*, 64, in press.

[2] Baldwin, J. and French, P. (1990), *Forensic Phonetics*, London, Pinter.

[3] Hollien, H. (1990) *Acoustics of Crime*, New York, Plenum Press.

[4] Küenzel, H. (1987) *Sprechererkennung: Grundzüge Forensischer Sprachverarbeitung*, Jeidelber, Kriminalistik-Verlag.

[5] Nolan, J.F. (1983) *The Phonetic Basis of Speaker Recognition*, Cambridge, UK, University Press.

[6] Stevens, K. N. (1971) Sources of Inter- and Intra- Speaker Variability in the Acoustic Properties of Speech Sounds, *Proceed., Seventh Inter. Cong. Phonetic Sci.*, Montreal, 206-232.

PROCEDURES AND PERSPECTIVES IN FORENSIC PHONETICS

Angelika Braun
Speaker Identification and Tape Analysis Section, Bundeskriminalamt, Wiesbaden, Germany

ABSTRACT

In this contribution, it is argued that the forensic applications of their field should no longer be ignored or denied by phoneticians. The development of forensic phonetics in the last decade including the increased importance of computerized procedures is outlined. Owing to the degradations introduced to the signal by the conditions under which forensic recordings are typically made, however, there is serious doubt that a fully automatic voice identification device will emerge in the near future. Topics for further research are indicated.

1. INTRODUCTION

The forensic application of phonetic sciences is one of the most controversial issues within the phonetics community. The extreme standpoints are probably represented by the successors of the so-called voice print technique in the United States, i.e. the Voice Identification and Acoustic Analysis Subcommittee (VIAAS) of the International Association for Identification (IAI) on the one hand and groups like the British Association of Academic Phoneticians (BAAP) or the Bureau du Groupe *Communication Parlée* de la Société Française d'Acoustique on the other hand. Whereas the former group not only advocates forensic speaker identification unconditionally but also basically holds the view that anyone with a high school diploma can do it after having undergone a two-week training course[1], the latter have taken rather strong positions against forensic phonetics in general and forensic speaker identification in particular by adopting motions to the effect that phoneticians should not engage in such tasks[2][1]. A third view on the subject is represented by the International Association for Forensic Phonetics (IAFP), which was formed in York in 1989. This organization aims to provide a forum for discussion among those who either work in the field of forensic phonetics and/or have an academic interest in it as well as to define and ensure professional standards in this area.

Discussion about the forensic application of phonetics has focussed on two principal issues: (1) Is it ethical for *anyone* to undertake forensic case work at all as long as scientific/empirical proof for the notion of "one speaker - one voice" has not been established?; (2) Are *phoneticians* more qualified than others to do forensic speaker identification?.

This contribution will address these issues which are controversial among phoneticians as well as - in line with the theme of this session - the question of what can be realistically expected to come from the laboratory within the next few decades.

Since a good part of the reservations that many phoneticians have about the forensic application of their field seem to stem from misconceptions about the exact nature of that work and the conditions under which it is done, a brief account of what forensic phoneticians actually do as well as the methods employed will be given. Although virtually any question related to speech or sound may be put to the phonetician in a particular case, this contribution will largely focus on speaker identification.

[1] There are indications to the effect that several members of BAAP now take a different view of forensic applications of phonetics than they did 15 years ago when the motion was passed.

2. WHAT FORENSIC PHONETICIANS DO

Some of the misconceptions of phoneticians about the forensic applications of their field may be due to the voiceprint legacy or other rather rash accounts of cases which are not representative of the state of the art in forensic phonetic work [3]. Specifically, many phoneticians do not seem to be aware of the fact that speaker identification, i.e. the comparison of a speech sample produced by an unknown speaker involved in the commission of a crime to that of one or more suspects, forms an important task for the forensic phonetician but by no means the only one. Other activities include speaker profiling or characterization, the analysis of disputed utterances, the analysis of background noise, the design of voice line-ups as well as interpretation of their results, intelligibility enhancement of noisy tape recordings, and tape authentication. The most relevant of these is speaker profiling, a task which is regularly requested in the early stages of e.g. kidnappings when a recording of the criminal's voice is available. Most of the time the voice forms the only lead at this stage of the investigation, and its analysis with respect to sex, age group, regional accent or dialect, peculiarities or defects in the pronunciation of certain speech sounds, sociolect, mannerisms etc. is of paramount importance for the investigation and thus, eventually, for the victim's life.

Not every forensic phonetician should or would engage in all of the above activities; what people are ready to take on largely depends on their specialization during their education and - as in other fields of expertise - on the amount of insight in the limits of their knowledge. The International Association for Forensic Phonetics has established a Code of Practice in order to ensure that its members will not exceed the limits of their expertise.[4]

In this context it is important to mention that one of the foremost duties of phoneticians is to explain to various groups of people what forensic phonetics **cannot** do, e.g. point to the limitations induced by telephone transmission or by speech samples in a language of which the phonetician does not have perfect command or the impossibility to judge a speaker's sincerity based on phonetic evidence alone.

3. THE FORENSIC ENVIRONMENT

At first glance, any discussion about forensic voice comparison methodology may seem quite dated in view of the fact that very powerful speaker recognition algorithms are available for commercial purposes, i.e. access control. But all of these systems require *cooperative* speakers in the sense that the speaker makes an effort to articulate clearly, that she or he agrees to pronounce a preselected phrase particularly suitable for comparison purposes, and that she or he is prepared to repeat an utterance if necessary. Needless to say, none of these prerequisites are met by forensic recordings.

Furthermore, in commercial speaker *verification* the number of speakers with whom the actual sample has to be compared is by definition finite, whereas forensic speaker *identification* is almost always an open-set task. Thus, even if as many as 20 recordings of suspects are submitted, there is no reason to assume that the offender is among them.

Aside from these principal issues there are some technical factors which preclude the use of commercial speaker recognition techniques in the forensic domain. The most frequent and also the most salient one is telephone transmission, which implies a bandwidth limitation to 300-3400 Hz and a restriction of the dynamic range to 30dB, if the line is good. This loss of frequency and amplitude information can obviously not be compensated for and leaves the phonetician with a limited basis for judgement. Specifically, formants outside the frequency range of the telephone line cannot

be measured, and misarticulations of fricatives like lisps may no longer be detectable.

Finally, the quantity and the quality of the material available for analysis is to a large extent controlled by the offender. Thus, even if as much as one minute of net speech (i.e. not counting pauses, hesitations etc.) is available, it may not be assumed that the material will fully represent the range of that person's verbal behavior.

One way of reacting to adverse conditions like the ones outlined above is, of course, to refrain from doing any forensic phonetic work at all. This seems to be the course of action suggested by BAAP as well as the GCP who have taken a rather strong view against phoneticians' engaging in forensic work by adopting motions to the effect that "phoneticians should not consider themselves expert in speaker identification until they have demonstrated themselves to be so" and "the GCP Bureau affirms that, in its opinion, speaker identification experts have yet to furnish any verifyable proof of their abilities"[2], respectively. On the other hand, there can be no doubt that phoneticians do possess specific knowledge about the human voice and its analysis, and it seems difficult to argue that the knowledge there is should be withheld from the legal community just because it is limited. To put it drastically: If a child has been kidnapped and a recording of the kidnapper's voice as well as that of a suspect were available, it would be absurd to outright refuse to do a phonetic voice comparison for lack of theoretical foundation.

4. METHODOLOGICAL APPROACHES

4.1. Auditory vs. spectrographic

As far as methods employed in forensic speaker identification are concerned, the history of forensic phonetics is a history of extremes. On the one hand, there used to be a very strong aural-perceptual phonetic tradition in Britain [3,5]. The conclusions reached by this method are largely based on a minute dialectological description of the samples in question, along with judgements of segment articulations as well as pitch and intonation. Although auditory phonetic procedures still form an important part of forensic speaker identification and obviously the most powerful tool in speaker profiling, voice comparison reports which are based on listening techniques alone are no longer considered state of the art [6,7].

The other extreme is represented by those who sought to reduce the human factor by applying various partly or fully automatic procedures. The worst facet of this is what has become known as the "voiceprint" technique, first introduced in the United States by Lawrence Kersta in the 1960s [8]. The obviously untenable analogy to the evidential value of fingerprints as well as the lack of theoretical foundation and poor training of the so-called "experts" [1] have done severe damage to forensic speaker identification as a whole. The visual inspection and comparison of spectrograms is obviously neither objective nor superior to aural-perceptual methods - the subjective judgement is merely shifted to the visual domain, and considering the sensitivity of the human ear as compared to the crude resolution of a spectrogram easily reveals the severe shortcomings of the technique as a whole. No claims are made concerning the theoretical validation of the procedures used beyond the - unvalidated - assumption that formant structures and other spectral characteristics which are evident from a spectrogram are different for each individual. Even though this assumption has been shown to be incorrect [9, 10], and voice identifications based on spectrograms were found to be much less reliable than those based on aural-perceptual judgements [11], it has taken decades to convince judges in most, though not all, States of the US to no longer admit "voiceprint" evidence, and it still seems to be practised in some countries including Israel and Italy. A slightly modified form of the voiceprint technique is still adhered to by the VIAAS of the IAI, but the "Voice Comparison Standards" as published by that organization [1] cannot be considered as a basis for serious discussion, as is indicated by the list of required reading for all of its examiners, which consists of 11 titles followed by the suggestion to read the manuals for any equipment used in the examination.

The use of formants as a sole basis for forensic voice comparisons has fairly recently been advocated in a different context by some scientists whose background seems to be in engineering rather than phonetics [12, 13, 14]. They propose to compare formant values and sometimes also pitch [sic]. They argue that all it takes to arrive at exact percentage values for the probability of identity or non-identity of two voice samples is the right statistical procedures. This, however, would only be true if it could safely be assumed that the within-speaker variation with respect to formants and formant-related acoustic parameters is under all circumstances smaller than the between-speaker variation. That this is precisely not the case has been demonstrated in the course of the voiceprint controversy (see above). Thus, approaches like those described so far do not only lack theoretical foundation but run counter to established phonetic knowledge.

4.2. The current approach

Since the early 1980s, an approach to speaker identification which combines traditional aural-perceptual and acoustic phonetic techniques has become increasingly widespread. It emerged from a research project at the German Bundeskriminalamt and has been used in thousands of cases at that institution alone [15]. The first stage in the examination consists in a detailed auditory analysis of the voice samples involved. Much like the profiling of anonymous voices, this part of the analysis pertains to parameters like voice quality, dialect or regional accent, speech defects, misarticulations of sounds, speech rate, intonation, rhythm, but it also includes observations on syntactic, idiomatic, and even paralinguistic features like breathing patterns. The main results of this analysis are documented in a transcript using IPA symbols in order to facilitate a comparison of the results with those of other experts. This aural-perceptual analysis is complemented by an acoustic phonetic examination of the recordings. Thus, several of the parameters used in the report can be quantified or described more precisely than by auditory analysis alone. A good example is formed by the set of parameters concerning voice. A "high-pitched" voice in auditory phonetic terms can thus be described as exhibiting an average fundamental frequency of, say, 158 Hz. What the auditory phonetician might call a well-modulated voice can be characterized as having a standard deviation from the average F0 of, say, 28 Hz. An intonation contour which strikes the auditory phonetician as "unusually stylized", can be described as involving steps of, say, 87 Hz. In the area of articulation, formants as well as e.g. the frequency of a "sharpened" /s/ or a strikingly long aspiration can be measured. Thus, of the parameters studied, as many as possible are documented using the whole set of techniques which are currently available in acoustic phonetics. Some of the algorithms were tailored to the specific needs of forensic material. All analyses are carried out bearing in mind the communicative context and the emotional state of a speaker. Of particular interest are features like those mentioned above, which deviate from the usual. The difficult part for the forensic phonetician is, of course, to define what is "usual" or "norm" and what is "deviation". This is partly done on the basis of statistics showing the distribution of features like average F0 in the relevant population or, if such are not available, on the basis of experience.

Much forensic phonetic research is directed at establishing distributional data for as many parameters as possible. For the time being, however, the subjective element in the formulation of the conclusion cannot be completely eliminated. For the same reason, conclusions are phrased in terms of probability scales instead of percentages. The phrasing of the probability in a particular case will depend on the amount, quality and phonetic-linguistic yield of the material and the rarity of the features which are contained in the voice(s) involved.

5. WHY PHONETICIANS

There are laboratory studies [16, 17] which suggest that trained phoneticians are not significantly (though marginally) better at certain perceptual tasks related to speaker recognition than phonetically naive subjects. Those studies specifically deal with (closed-set) speaker identification and pairing [16], and age estimation [17]. The relevance of these findings with respect to forensic speaker identification, however, is not quite clear, because the experimental design of neither study represents forensic conditions. Furthermore, in an experiment reported by Köster [18], recognition and identification rates were higher for the expert than for the non-phoneticians. On the other hand, there may be perceptual tasks at which phoneticians are not necessarily much better than phonetically naive listeners. One should look very closely whether any experiments carried out in this area test genuinely phonetic skills - let alone *forensic* phonetic skills [6] - or involve intuitive tasks whose underlying mechanisms have not even been fully explained as yet. Certainly, the results of tests like those cited above should not tempt us into thinking that phoneticians are no more competent to describe and analyze voices than non-phoneticians and that therefore forensic speaker identification can be done by anyone. This would almost amount to a denial of phonetics as a scientific field.

Forensic phoneticians have been criticized for not having come up with their own experiments which would demonstrate that they have speaker identification skills which are superior to those of ordinary people [19]. On the other hand, the question is whether there is a fair (to the non-phoneticians) way of comparatively testing genuinely phonetic skills like doing a narrow transcription, describing the laryngeal setting of a certain speaker or explaining why a lisp cannot be detected in a telephone call. Particularly in the courtroom situation, it is of paramount importance that any opinion about voice identity be made explicit in terms of descriptive phonetic parameters. In order to do this, phonetic training is mandatory.

As Bolt et al. point out [10, p.99] there are "two kinds of experience [which] provide knowledge about the problems inherent in voice identification as well as some indication of possible success. The first is the experience of those who have attempted the task in real-life situations. The second is that of laboratory experimenters [...]".The position outlined above is strongly supported by the first type of experience cited. Although no exact account was kept, within the BKA laboratory alone there are literally hundreds of cases in which non-phoneticians have made very strong claims about speaker identity, while the phonetician indicated that the samples originated from different speakers. A typical example occurred in the course of the investigation of a kidnapping. A Turkish boy had been abducted and was still held by the kidnapper(s). Two police officers who had been listening to telephone taps of a particular person implicated in another crime for several months were absolutely convinced that he was also the kidnapper who had phoned to demand ransom. The voices were indeed very similar, but there was also phonetic evidence suggesting that the samples came from two different people. Later in the investigation, another suspect was recorded because he had been identified by witnesses as having made the anonymous phone call, but again there was strong phonetic evidence against identity (i.e. the suspect had a stutter whereas the offender did not).Thus, even without formal testing, there is a lot of evidence from everyday work for the superior performance of phoneticians.

This example can also be used to demonstrate the implications of forensic phonetic work: If the phonetician fails to recognize speaker identity, the kidnapper goes free, and the victim may be killed. If the scientist falsely identifies the wrong person, that person might be physically harmed by members of the special squad trying to make an arrest and free a kidnapped child. In the present author's view, this kind of responsibility should make anyone involved in forensic phonetic work very cautious, but it can hardly be used as an argument against providing expertise to the legal community.

Another reason why it seems difficult for phoneticians to refrain from forensic case work altogether is a political one. With so-called speech analysis packages available for any home computer for less than £100, even people with no specific training in phonetics may set out to do forensic work. French [3, pp. 58-59] mentions two cases from England in which sound engineers failed to distinguish between letters and sounds in their reports. In another country, two former members of the police force set out to do speech enhancement using commercially available signal manipulation software, having to admit that they were not sure what was actually happening when they operated certain controls.

There is an imminent danger that this will happen much more often in the future, particularly in countries like England and the United States whose judicial system is adversarial, i.e. where usually both sides hire their own experts. Under these circumstances, it would seem almost like a moral obligation to speak up against charlatans working for the other side. It should be added at this point that in Germany as well as the Netherlands the conditions under which any forensic expert works are quite different: The judicial systems in these countries can be described as inquisitorial rather than adversarial, this term implying that any expert is appointed by the court rather than by one side. The rôle of an expert within these systems is to supply the court with expertise pertaining to specific areas in which the judges themselves[2] do not feel sufficiently competent. The expert is to be impartial, and she or he *has* to present a full report of her or his findings irrespective of the implications for the trial. Thus, it is extremely uncommon to have more than one expert in a trial, and some of the problems specifically related to the fact that phoneticians may act as "hired guns" simply do not occur. The author would like to add at this point that she is extremely grateful to be working in this kind of framework since she would find it difficult, if not impossible, to be restricted in what she says by either prosecution or defense strategy.

6. ANSWERS FROM THE LAB

Nolan has pointed to the shortcomings with respect to the theoretical foundation of forensic speaker identification 12 years ago [20]. Defining the speaker under laboratory (HiFi) conditions seems to be a vastly different (and in many respects: easier) task than defining what is left of a speaker in terms of information contained in the signal under forensic conditions. In view of the limitations outlined above, there is a possibility that we may never be able to come up with an exhaustive list of speaker-characterizing features at all. Even if it could be demonstrated experimentally that each speaker has a voice which is distinct from

[2]In Germany, there are no jury trials. Instead, for major crimes there is a panel of five judges, two of whom are lay persons. They decide on the question of guilt as well as the sentence by majority vote.

those of all other speakers of that speech community, this does by no means imply that the distinction can always be discovered in the forensic material that happens to be available. On the other hand, there are many areas of empirical research which can help to widen the basis for judgement under forensic conditions.

Thus, a two-way approach is suggested here. Obviously, any laboratory experiment addressing the one-speaker-one-voice issue will be of great interest to anyone involved in forensic phonetics, even though the findings may have no immediate bearing on forensic work, e.g. if articulatory parameters are measured. Of particular interest from the forensic point of view would be attempts to describe the full range of a person's verbal behavior, i.e. changes introduced to the "neutral" way of speaking by psychological (stress, emotion) or physiological (fatigue, smoking, alcohol, medication) factors. Some of these factors have been studied in detail, often with the forensic application in mind [examples are 21-24], but the need for this kind of "top-down" research providing basic data will probably not be met for decades.

On the other hand, there is the necessity to start at the other end, i.e. to ask how, in view of the forensic environment, the procedures currently used in speaker identification can be improved. This "bottom-up" research starts out at the parameters which can still be assessed in degraded recordings and seeks to either quantify parameters which could not be quantified before or to gain information concerning the statistical distribution of certain features in order to be in a better position to assess the frequency of their occurrence. One example for this kind of research is a project currently under way at the Bundeskriminalamt, involving the quantification of a certain type of hoarseness from running speech [25]. A matter of great interest not only to forensic phoneticians would be an exhaustive phonetic description of hesitation markers including questions of intrapersonal vs. interpersonal variability. Another area for research would be the distribution of phonetic and linguistic characteristics in the population, which would enable the expert to weigh that parameter more precisely. An example of this kind of research is currently being carried out as a joint project between the Bundeskriminalamt and the Universities of Marburg and Trier [26]. It consists in establishing a data base of regional varieties of German and will enable the forensic phonetician to listen to samples of up to five min. duration from 450 locations. Narrow phonetic transcriptions of the samples are available. It is also possible to search for specific segments, morphs and words (in different phonetic contexts). A thesaurus component has been built into it which will display the phonological system of the accent or dialect in question. It would certainly be desirable to have similar data bases for e.g. speech defects.

7. CONCLUSIONS

Speaking is such a complex type of behavior that I tend to be sceptical that we may expect an answer to the question asked in the theme of this session any time soon. I am not even sure that the answer is going to be positive, particularly with the complicating factors induced by the forensic setting in mind. No matter how good a definition of a speaker will come from the lab, the forensic application of these findings will always be limited by the amount of information about that speaker which is contained in a recording. This applies both to the technical side, i.e. the amount of frequency and amplitude information available, and the representativeness of the material in terms of the speaker's "normal" voice. There is no doubt, however, that any step that is taken towards the definition of each speaker will make forensic speaker identification an easier task.

8. REFERENCES

[1] Voice Identification and Acoustic Analysis Subcommittee (1991), "Voice comparison standards", *J. Forensic Identification*, vol.41, pp. 373-392.

[2] Bureau du Groupe *Communication Parlée* de la Société Française d'Acoustique (1990), "Motion adopted Sept.7", *NESCA - The ESCA Newsletter*, no. 4, p 39.

[3] Baldwin, J. & French, P. (1990), *Forensic phonetics*. London: Pinter.

[4] "Announcement", in *JIPA*, vol. 22, pp. 80-81.

[5] Ellis, S. (1994), "The Yorkshire Ripper enquiry: part I", *Forensic Linguistics*, vol. 1, 197-206..

[6] Künzel, H.J. (1994), "Current approaches to forensic speaker recognition". *Proc. ESCA Workshop on automatic speaker recognition, identification, verification*, Martigny, pp. 135-141.

[7] Nolan, F. (1990), "The limitations of auditory-phonetic speaker identification". In: H. Kniffka (ed.), *Texte zu Theorie und Praxis forensischer Linguistik*, Tübingen: Niemeyer, pp.457- 479.

[8] Kersta, L. (1962), "Voiceprint identification", *Nature*, vol. 196, pp. 1253-1257.

[9] Hollien, H. (1990), *The acoustics of crime*. New York: Plenum Press.

[10] Bolt, R.H. et al.(1979), *On the theory and practice of voice identification*. Washington, D.C.: Natl. Academy of Sciences.

[11] Stevens, K.N., C.E. Williams, J.R. Carbonell, and B.Woods (1968), "Speaker authentication and identification: A comparison of spectrographic and auditory presentations of speech material", *JASA* vol.44, pp.1596-1607.

[12] Maturi, P. (1990), "Speaker identification in forensics: a simulation experiment". *Proc. ESCA Workshop on speaker characterization in speech technology*, Edinburgh: CSTR, pp.155-160.

[13] Federico, A. & Paoloni, A. (1993), "Bayesian decision in the speaker recognition by acoustic parametrization of voice samples over telephone lines". *Proc. 3rd European Conference on Speech, Communication, and Technology, Berlin*, vol. 3, pp.2307-2310.

[14] Lipeika, A. & Lipeikiene, J. (1993), "The use of pseudostationary segments for speaker identification". *Proc. 3rd European Conference on Speech, Communication, and Technology, Berlin*, vol. 3, pp.2303-2306.

[15] Künzel, H.J. (1987), *Sprechererkennung. Grundlagen forensischer Sprachverarbeitung*. Heidelberg: Kriminalistik-Verlag.

[16] Shirt, M. (1984), "An auditory speaker-recognition experiment", *Proc. Inst. Acoustics*, vol. 6, pp.101-104.

[17] Braun, A. & Rietveld, A. (1995), "The influence of smoking habits on perceived age". *Proc. XIIIth Intl. Congr. Phon. Sc.*, Stockholm, 4 p.

[18] Köster, J.-P. (1987), "Auditive Sprechererkennung bei Experten und Naiven". In: R. Weiss (ed.), *Festschrift für H. Wängler*, Hamburg: Buske, pp.171-180.

[19] Nolan, F. (1991), "Forensic phonetics", *J. Linguistics*, vol. 27, pp.483-493.

[20] Nolan, F. (1983), *The phonetic bases of speaker recognition*, Cambridge: CUP.

[21] Künzel, H.J., Braun, A., Eysholdt, U. (1992), *Einfluß von Alkohol auf Stimme und Sprache*, Heidelberg: Kriminalistik-Verlag.

[22] Hollien, H. (1993), "An oilspill, alcohol and the captain: a possible misapplication of forensic science". *Forensic Science International*, vol. 60, pp. 97-105.

[23] Johnson, K., Pisoni, D.B., Bernacki, R.H. (1990), "Do voice recordings reveal whether a person is intoxicated? A case study, *Phonetica*, vol. 41, pp.215-237.

[24] Braun, A. (1994), "The effect of cigarette smoking on vocal parameters".*Proc. ESCA Workshop on automatic speaker recognition, identification, verification*, Martigny, pp. 161-164.

[25] Wagner, I. (1995), "Jitter-measurements from telephone-transmitted speech". *Proc. XIIIth Intl. Congr. Phon. Sc.*, Stockholm, 4 p.

[26] Künzel, H.J. & Köster, J.-P. (1995), "Forensic Data Base System of Regional Accents of German", *Proc. AAFS Annual Meeting*, Seattle, Wa., p.88.

THE ROLE OF AUTOMATIC SPEAKER RECOGNITION TECHNIQUES IN FORENSIC INVESTIGATIONS

A.P.A. Broeders

National Forensic Science Laboratory, Rijswijk, Netherlands
Department of Language and Speech, University of Nijmegen, Netherlands

ABSTRACT

There are several major differences between the sphere of application of automatic speaker recognition techniques and the conditions attending speaker identification in the forensic context. Some of the factors involved are discussed below. The prevailing view that these differences preclude the introduction of even the more powerful automatic verification techniques in forensic work is questioned, and an indication is given of ways in which this question may profitably be addressed.

1 INTRODUCTION

At first sight, the proposition which serves as the central theme for this session - the definition of the speaker can be expected to come from the laboratory in the next few decades - looks simple enough. It seems to call for either wholehearted support or utter rejection, and it was no doubt phrased with the express aim of provoking such primitive responses. However, the simplicity of the proposition is somewhat deceptive. In its present form, I find myself unable to react to it in unequivocal terms. I have therefore taken the liberty of reformulating it in terms of what, from my perspective anyway, seems to be the real question underlying it: "Can automatic speaker recognition techniques be expected to play a role in the forensic context in the foreseeable future?"

In fact, this question must itself be rephrased in several ways, with each subquestion addressing a different aspect of the central issue. Some of these questions are discussed below, and some indications are given of the way in which they might be resolved.

2 IS THERE A VOICEPRINT?

It is widely accepted today that the term voiceprint is a misnomer for what is basically simply a spectrographic representation of a particular utterance by a particular speaker. Indeed many would argue that the term is better avoided altogether. However, there is a sense in which the term can usefully be employed in a manner which rather more closely resembles the parallel use of the term fingerprint, i.e. to refer to a unique representation of a particular individual. For the sake of the present discussion we could conceive of a voiceprint as a representation, in whatever shape or form, of such acoustic information as will uniquely characterize each individual speaker. This would enable us to address a more specific question, viz. whether a voiceprint in the sense just defined is in fact a real possibility.

Obviously, such a unique representation can only fully serve its purpose if we can rely on the signal under examination to contain the acoustic information that is required for a unique identification. However, we know that on the physical plane speech is marked by constant variation. The representation we are looking for would therefore have to reside in a continuously varying signal. But we know that as ordinary language users, even when dealing with speakers with whom we are very familiar, we are liable to make identification mistakes, especially - but not exclusively - in

situations where we are expecting a particular speaker but are in fact exposed not to the expected speaker but to a close soundalike. We may think we hear a friend answering the phone, only to find that we are talking to his son. This suggests that, for human listeners at any rate, there is a very real sense in which we cannot be sure that no two speakers speak exactly alike (Nolan [1]), and that we must at least consider the possibility that there is not always enough speaker-specific information in the signal to enable us to verify the identity of a familiar speaker, let alone that of an unfamiliar speaker.

Over and above the inherent variability of speech as a physical phenomenon, there is of course the variation inherent in speech on the linguistic plane. Anyone who has been in a position to listen to even a moderate amount of unmonitored speech will have been struck by the wide variety of speech styles used by many speakers in different communicative contexts. As language users we may be able to identify speakers on the basis of utterances produced in a quiet conversational style with reasonable success but we have great difficulty doing this if the utterances are produced with different degrees of intensity. Similarly, we do not feel confident about extrapolating the quality of a speaker's whisper, headvoice or loud voice from speech produced by the same speaker with a modal voice quality (Broeders and Rietveld [2]).

Given the variability of speech in different communicative contexts, it is doubtful whether any representation can be made which will capture the unique acoustic information required for the identification of the speaker from signals as diverse as those found in real-world conditions. Or, phrased differently, it is doubtful whether the speaker-specific information contained in the signal will be sufficiently uniform and consistent across various speech styles to serve as a basis for automatic speaker recognition in situations that are more challenging than the typical closed-set automatic verification context. So it appears that no matter what type of speaker profile we conceive of, it is likely to lack the one property that, together with uniqueness, makes the fingerprint such a powerful means of identification, i.e. invariance.

It is worth noting though that in spite of the lack of reliability that has been shown to be associated with the traditional voiceprint technique in forensic speaker identification (Bolt et al. [3]), one still occasionally comes across unwarranted claims like that recently found in a brochure advertising the 'Kreutler Computerised Speech Lab'. Next to the photograph of a computer screen display which, on closer examination, turns out to bear a more than remarkable resemblance to the Kay CSL-system, the law enforcement and security services type clientele that the brochure seeks to address are offered the following information: *'Forensic analysis is a widely spread technique to identify persons by their voice prints. These voice prints are specific for each person and can not be altered.'* ([4], p. 6)

3 FORENSIC VS COMMERCIAL APPLICATIONS

Various authors, including both Künzel [5] and French [6], have drawn attention to the severe limitations imposed by real-world conditions on forensic speaker identification and discuss the implications this has for the application of automatic speaker recognition procedures, as used in commercial applications. There are five major factors that need to be taken into account here. They are:

Text dependence

In automatic speaker verification

systems the utterances that are used for the verification test can in principle be pre-selected for best performance. In the forensic context the nature (and size) of the contested material is normally entirely beyond the investigator's control, and the nature of the reference material, i.e. the material that is known to have been produced by the known speaker (usually the suspect), is often determined by what happens to be available in a particular case.

Speaker cooperation

Even if reference material can be collected expressly for the purpose of an identification test, the investigator will, even at the best of times, have to be mindful of the observer's paradox (Labov [7]). As it will not normally be legally possible to collect a speech sample without the suspect being aware of it, let alone without the suspect's consent, there is the very real danger that the reference material that is collected does not constitute a representative sample of the suspect's speech.

Obviously, speakers may deliberately set out to systematically alter their speech style, may choose to be less than forthcoming and may more generally try to avoid producing a representative speech sample. However, even cooperative speakers may, as a result of the stressful nature of the situation they find themselves in, produce speech that varies considerably from their usual repertoire. In the automatic speaker verification context however, the situation is unlikely to be experienced as stressful and speakers can normally be relied upon to be cooperative since they stand to gain from a positive result.

Of course, the questioned material itself may show signs of varying degrees of deliberate disguise or, more generally, be of a nature which virtually precludes its being subjected to any type of systematic investigation.

Recording and transmission conditions

Telephone recordings account for a very large proportion of all forensic material. In addition to the major frequency bandwidth reduction of the telephone system, the effect of the handset and various less predictable signal modifications introduced by the telephone system, there are the effects of a wide range of recording equipment to be reckoned with. Between them they may give rise to a variety of signal degradations and distortions which may vary quite considerably from one call to the next. In the verification context, of course, none of these complications will normally arise, since great pains will be taken to control the quality of recording equipment and transmission channels.

Class size

A major problem for the application of automatic speaker recognition techniques in the forensic context lies in the size of the speaker set in real-world forensic conditions. Automatic procedures are typically geared to applications with a known or closed set of speakers. On the other hand, in the forensic context, the unknown speaker cannot be assumed to be one of a small set of speakers but must normally be taken to be one of a class whose membership, if not of indefinite size, may very often be quite large and is typically unknown.

Cost of errors

There is an even more fundamental difference between the usual sphere of application of automatic techniques and the forensic context. As is well-known, in closed-set verification systems, there is a trade-off between the false acceptance of unknown speakers or impostors and the false rejection of known speakers or customers. In a commercial application, the cost incurred by the false acceptance of an imposter in the form of unauthorized access to information, services or facilities can be balanced against the frustration and loss of time generated by the false rejection of a bona fide customer. But in the legal setting, such a cost-benefit analysis in essentially financial terms would be unthinkable. Indeed, it has often been argued that in the forensic context any method that, in addition to correct identifications, will produce even a single incorrect identification is unacceptable, since it conflicts with one of the fundamental principles that any judicial system may be required to subscribe to, which says that it is better to have a guilty person acquitted than an innocent suspect convicted.

4 COMMON PROBLEMS

Although it is fair to say that the factors discussed above present formidable obstacles to the introduction of automatic speaker recognition techniques in the forensic context, this should not be taken to imply that there is no point in investigating conditions in which benefits may be derived from their application. It may well be the case that an automatic speaker identification technique, in the sense of a set of decision procedures that is carried out entirely independently of human interpretation, is an unrealistic scenario but that is not to say that there is no room for these methods at all.

In fact, there may be good reasons for a somewhat more optimistic view than is taken by many commentators. Part of the explanation for the lack of progress may lie in the gap separating what, perhaps somewhat disrespectfully, may be termed the engineering approach as opposed to the linguistically-oriented approach to speaker recognition. Leaving aside the decreasing number of adherents of the voiceprint technique, practising forensic phoneticians, especially those associated with the International Association for Forensic Phonetics (IAFP), are keenly aware of the need to bridge this gap. There are indications that speech technologists too are aware of the need to take more account of both the linguistic and the judicial aspects of forensic speaker identification (Bimbot et al. [8], p. 82). This development may well be aided by a growing awareness that the factors limiting the applicability of automatic procedures do not in fact always constitute absolute impediments.

The text

A good example is text dependence. The use of a limited number of fixed passwords obviously tends to render the older automatic verification systems vulnerable to fraud. After all, with the increasingly widespread availability of low-cost, high-quality digital speech processing technology, it is not too difficult to record the voice of a bona fide customer and subsequently replay it to gain unauthorized access to a particular system or service. So the need arises for text-independent or text-prompted formats. A possible solution is a combination of speech and speaker recognition techniques which allows the system to freely prompt random utterances and to check not only whether the voice is that of the customer but also whether the required text is produced (Furui [9]). On the other hand, there are many forensic situations where the requirement of text dependence, i.e. the availability of identical utterances in both questioned and reference materials, can easily be met.

The speaker

The same applies to speaker cooperation. Again, there are situations when reference material is available whose status is not contested by either party and which also satisfies the major demands that it is representative of the speaker's linguistic repertoire and is produced in a communicative context which is similar to that in which the

questioned material was produced, so that an adequate basis for comparison exists. Conversely, in commercial speaker verification environments, there are obviously also limits to the degree of cooperation with which the speaker can realistically be relied upon to interact with the machine. Ironically enough, the use of such a pre-eminently human faculty as language by machines will often cause frequent users to lose patience with other, less than human characteristics of the machine and to develop a reluctance to adapt their performance to the machine's requirements. Possible effects on speech include a loss of articulatory precision and lower overall intensity.

The telephone line

Telephone transmission conditions do not in actual forensic casework necessarily always vary more than they would in commercial verification applications. In fact, it is quite common for recorded telephone conversations that are the subject of a forensic inquiry to have been made from the same location, through the same extension and on the same day. Recording conditions are also frequently at least potentially controllable to the point where they may be sufficiently uniform to meet the same technical requirements that must be met in commercial applications. Traditional analogue telephone logging and tapping devices are increasingly being replaced with advanced digital facilities, with calls being stored in a digital format.

The speaker set

Class size is probably ultimately the more intractable problem. This is sometimes obscured by the confusion that is created by the use of the terms identification versus verification. In fact, forensic phoneticians are typically involved not in speaker identification but in speaker verification albeit - and here lies the real difference - with an open set of speakers rather than a closed set. But the question that poses itself in the forensic context is essentially a verification, not an identification problem: is the questioned material produced by the same speaker as the reference material? In more concrete terms: were all the questioned calls made by the same person, and if so, do they originate from the person who is believed to have made them?

The complication introduced by the circumstance that in the forensic context the unknown speaker is not normally claimed to be one of a closed set of speakers but must be assumed to be one of an open class creates problems that are essentially of a statistical nature. What an objective forensic procedure would be required to do is not just to quantify the degree of similarity between questioned and reference samples and make a decision based on a comparison with a pre-determined threshold, as occurs in closed-set verification applications, but to give a statistically sound indication of the probability of this degree of similarity occurring by chance. Or, to phrase the question in Bayesian terms, it should allow one to calculate the likelihood ratio of the probabilities that the findings would arise under the two conditions that the defendant was, and was not the unknown speaker (Evett [10]).

The consequences

Finally, there is the cost of error aspect. Obviously, erroneous conclusions can do a great deal of harm, especially if findings are presented without an indication of the reliability of the methodology used with reference to the specifics of a particular case. On the other hand, if our final criterion is that a method be demonstrated to produce no false positives, it may well be unnecessarily strict. What is important is that reliable statistics can be given, or that, if a probability scale is used,

the relative position on this scale of the particular degree of probability arrived at in a particular case is indicated, and a clear statement is given of the limitations of the methodology employed (Nolan [11]). If this requirement can be met, speaker identification evidence does not compare unfavourably with other types of expertise that are regularly sought by courts of law. By the nature of their work, judges are constantly involved in weighing probabilities and uncertainties. Deference to experts of whatever designation is a threat to any judicial system (Nijboer et al. [12]), although the danger may well be greater in adversarial systems where 'rival' experts find themselves in the business of explaining their findings to a jury, whose critical faculties may well be taxed beyond capacity by the level of abstraction required to follow the argument.

Also, there is an as yet largely uncharted demand for forensic speaker recognition expertise for investigative rather than evidential purposes. In large-scale police investigations a degree of uncertainty may be less problematic and an informed use of automatic procedures may improve the quality of decisions and lead to considerable savings in time and staff expenditure.

5 COMBINED RESEARCH

A particularly promising approach is that described by Boves et al. [13]. Within the design of the Dutch POLYPHONE speaker database a number of operational conditions are systematically varied so that their effects can be investigated. The recording platform used to collect the speech of the 5,000 speakers in the POLYPHONE database proper, was also used to collect an additional 2 groups of 50 speakers each, specially selected to examine the effect of variables like kinship and linguistic background. The speakers are 100 adult males, all native residents of two distinct parts of the Netherlands, the cities of The Hague in the West and Nijmegen in the East, who between them form some 50 pairs made up of two or more brothers, or a father and a son. The composition of this speaker set was partly inspired by the sort of questions that are particularly relevant in the forensic real-world context, where the pertinent statistic is not how likely a speaker is to be confused with a random 'imposter' but with a speaker with a similar linguistic background. Forensic phoneticians are rarely asked to compare samples involving clearly different accents but suspects or their barristers may well claim that the speaker in the questioned recording is the suspect's brother, and the circumstances of the case are often such that this possibility cannot be ruled out.

The design makes it possible to investigate a variety of questions that are particularly relevant to the forensic field. The project includes experiments to compare identification performance among the two sets of closely matched speakers with that among the larger group of male POLYPHONE speakers, and to investigate within-dialect as opposed to between-dialect confusions, as well as experiments to study the effect of close kinship on error rates. As all speakers in both sets of 50 each made 8 phone calls using two different handsets, intra-speaker and inter-phone variation can also be studied.

Preparations are also under way to test the performance of the arithmetic-harmonic sphericity measure developed by Bimbot and Mathan [14,15] on the material produced by the two sets of 50 speakers.

6 THE DEBATE CONTINUES

In some countries, speaker identification in the forensic context is a very controversial issue. To some extent,

this may be due to the exaggerated claims made by those who were responsible for the introduction of the so-called voiceprint technique. At the same time though, the short-lived popularity of the voiceprint may serve as a vivid reminder of the need for phoneticians to take an active interest in forensic questions, if only to expose phonetically unsound testimony offered by non-phoneticians of various denominations.

Of course, individual phoneticians must decide for themselves whether they wish to do forensic research or take on actual casework. But, as argued elsewhere (Broeders [16]), it would be wrong for phoneticians or linguists as a body to refuse to be involved in forensic work for the sole reason that they feel their discipline cannot provide incontrovertible evidence. That is nevertheless exactly what the motion adopted by the Groupe Communication Parlée de la Société Française d'Acoustique [17] would seem to advocate, inasmuch as it effectively calls for the withdrawal of all phonetic expertise from the field of forensic speaker identification. However, ironically enough, the overriding importance of the need for speech scientists and phoneticians to collaborate with those with first-hand knowledge of real-world forensic conditions could hardly have been demonstrated more forcibly than by the text of the motion. It reflects a sad lack of understanding of the type of question that poses itself in the forensic context, of the way in which these questions are handled by practising forensic phoneticians in countries like Britain, Germany and The Netherlands, and of the role and the responsibility of the expert witness in a judicial investigation.

That this position is unlikely to stimulate the necessary collaboration between forensic practitioners and other phoneticians and speech scientists is all the more unfortunate as phonetics as a science only stands to gain from the type of questions that emerge from the real-world conditions that apply in the forensic context. Fortunately, though, there are also indications that more and more phoneticians and speech scientists are taking an active interest in the problems posed by forensic speaker identification, with symposia like the present providing an ideal opportunity to exchange views, clear up some of the more persistent misunderstandings and define common research aims.

7 CONCLUSION

Recent developments have led to a situation where closed set speaker verification and open class forensic speaker identification have come to share a greater number of problems than has so far been the case. It follows that there is every reason to look into the possibility of combined research. The projects described in section 5 provide good examples of this approach. It is based on the premise that, in forensic applications too, performance of automatic recognition techniques will be dependent on the amount of control that can be exerted on operational conditions (Doddington [18]). It implies that in carefully controlled forensic conditions automatic procedures may in due course also come to play a role, if only for investigative rather than evidential purposes.

However, even here the process will never be fully automatic. It will always take an experienced phonetician or a linguistically informed speech scientist to decide what parts of the speech samples under examination are linguistically sufficiently similar to be used as suitable test material. Ultimately, then, it is the variation along the linguistic dimension that may well prove to be least amenable to efforts to bring automatic speaker verification techniques to bear on forensic material. In other words, it is unrealistic to anticipate a fully automatic procedure that will be able to extract a sufficiently comprehensive speaker profile from a questioned speech sample, given the variety of speech styles encountered in forensic conditions.

REFERENCES

[1] Nolan, F. (1991), 'Forensic Phonetics', *Journal of Linguistics* 27, 483-493.

[2] Broeders, A.P.A. & A.C.M. Rietveld (1989) 'Segmental Marking as a Cue in Auditory Voice Identification of Telephone Speech', in: J.P. Tubach & J.J. Mariani (eds.), *Eurospeech 89*, CEP Consultants, Edinburgh, 71-74.

[3] Bolt, R.H. et al. (1979) *On the Theory and Practice of Voice Identification*, Washington DC: National Academy of Sciences.

[4] 'Professional Telecommunication Systems', brochure published by Kreutler, Brussels.

[5] Künzel, H.J. (1994) 'Current Approaches to Forensic Speaker Recognition', *Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, Martigny, 135-141.

[6] French, P. (1994) 'An Overview of Forensic Phonetics with Particular Reference to Speaker Identification', *Forensic Linguistics* 1(2), 169-181.

[7] Labov, W. (1972) *Sociolinguistic Patterns*, Oxford: Blackwell.

[8] Bimbot, F., Chollet G., Paoloni A. (1994) 'Assessment Methodology for Speaker Identification and Verification Systems' *Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, Martigny, 75-82.

[9] Furui, S. (1994) 'An Overview of Speaker Recognition Technology', *Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, Martigny, 1-9.

[10] Evett, I.W. (1991) 'Interpretation: A Personal Odyssey', in: Aitken, C.G.G. and Stoney, D.A. (eds.) *The Use of Statistics in Forensic Science*, New York: Ellis Horwood.

[11] Nolan, F. (1992) 'Code of Practice', *Journal of the International Phonetic Association* 22(1 & 2), 80-81.

[12] Nijboer, J.F., Callen, C.R., Kwak, N. (eds.) (1993) *Forensic Expertise and the Law of Evidence*, Amsterdam: North-Holland.

[13] Boves, L., Boogaart, T., Bos, L. (1994) 'Design and Recording of large Databases for Use in Speaker Verification and Identification', *Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, Martigny, 43-46.

[14] Bimbot, F., Mathan, L. (1993) 'Text-free Speaker Recognition using an arithmetic-harmonic Sphericity Measure' *Proceedings of Eurospeech*, Berlin, 169-172.

[15] Bimbot, F., Mathan, L. (1994) 'Second-Order Statistical Measures for Text-independent Speaker Identification' *Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, Martigny, 51-54.

[16] Broeders, A.P.A. (1991) 'Great Debate on....' *Nesca - The ESCA Newsletter* 5, 50-51.

[17] Bureau du Groupe Communication Parlée de la Société Française d'Acoustique (1990), Motion adopted on September 7, *Nesca - The ESCA Newsletter* 4, 39.

[18] Doddington, G.R. (1985) 'Speaker Recognition - Identifying People by their Voices', *Proceedings of the IEEE*, 73(11), 1651-1664.

PROLAB - THE KIEL SYSTEM OF PROSODIC LABELLING

K.J. Kohler
IPDS, Kiel, Germany

ABSTRACT

For the Kiel Corpus of Spontaneous Speech [1] a prosodic labelling system (PROLAB) has been developed. It is based on a prosodic model for German (KIM - The Kiel Intonation Model) and uses a 7 bit ASCII repertoire.

KIM - THE KIEL INTONATION MODEL

The prosody model incorporates the following domains [2,3]:

(1) lexical stress – three levels: unstressed, secondary stress in compounds, and primary stress

(2) sentence stress – four levels: reinforced, accented, partially and completely deaccented

(3) intonation:
- pitch peaks and valleys and their concatenation
- synchronization of pitch peaks and valleys with stressed syllables – three steps: early, medial, late
- downstep of successive pitch peaks/valleys and pitch reset

(4) prosodic boundaries (degrees of cohesion) – three variables: pause duration, phrase-final segmental lengthening, scaling of F0 end points

(5) overall speech rate

(6) disfluencies: pauses, breathing, hesitations

Stress

Within stress we have to differentiate between lexical and sentence stress. At the abstract level of phonological specifications in the lexicon, every German word has at least one vowel that has to be marked as potentially stressable, as being able to attract the feature specifications of sentence stress. Lexical stress is thus not a distinctive stress feature, it only marks a position that can attract such a feature at the sentence level, but need not.

By default, content words are accented and function words completely deaccented. Deviation from default content word stress may be partial or complete deaccentuation, determined by syntax, semantics and pragmatics. Thus, e.g., in 'verb + direct object' constructions the verb is partially deaccented in neutral (non-focussed) accentuation, no matter whether it precedes or follows the object (*Max schreibt einen Brief./Max hat einen Brief geschrieben.*), whereas in 'verb + adverbial' constructions default accentuation is kept (*Max hat täglich geschrieben.*). In either case deviation from this neutral pattern implies focus (of the verb or the adverb, respectively). Complete deaccentuation in the first case introduces focus contrast on the object, which may be strengthened by emphatic reinforcement. Function words, although completely deaccented by default can receive all the content word sentence accent categories by deviation from default.

Intonation

All lexically stressed vowels of words with 'primary' or 'secondary' (= partially deaccented) sentence stress receive intonation features, which may be either 'valleys' or 'peaks', and in the case of 'peaks', they may contain a undirectional F0 fall, or rise again at the end, resulting in a fall-rise. 'Valleys' may have a low rise, to indicate, e.g., continuation, or a high rise, used, e.g., in questions.

All 'peaks' and 'valleys' may have their turning points (F0 maximum in 'peaks', or F0 minimum in 'valleys') early or later with reference to the stressed-vowel onset. For 'peaks' the non-early position may be around the stressed vowel centre (= medial) or towards its end (= late).

Peaks are characterized by a quick F0 rise confined to the vicinity of a sentence-stressed syllable. This rise precedes the onset of the latter, and is usually short and narrow in range, for an early peak; it extends into the first half of the stressed nucleus in the case of a medial peak. In the late peak, it starts after the stressed vowel onset and continues into the second half of the nucleus or beyond; the exact timing of the maximum peak value depends on vowel type (duration according to quantity and quality), subsequent voiced/voiceless consonants and number of immediately following unstressed syllables. There may even be a low stretch of F0 in the stressed vowel before the rise.

Valleys, on the other hand, have a continuous rise, starting before the stressed-syllable nucleus (early) or inside it (non-early) and extending as far as the beginning of the following sentence-stressed syllable. If there are several unstressed syllables between two sentence stresses a valley is thus realised as a more gradual F0 ascent compared with the much quicker rise for a late peak. The less distance there is between stressed syllables the more difficult it becomes to distinguish between a 'valley + peak' and a 'late peak + peak' sequence, especially if there is no F0 dip in between the first and second stress F0 maxima, as in a hat pattern.

In a concatenation of pitch peaks without prosodic boundaries between them, F0 may fall to a low or an intermediate level and then rise again for the next peak. This fall will be effected on intervening unstressed syllables between the two peaks, reaching the lowest point, to start the next rise, in the vicinity of the following stressed syllable, depending on peak position. If there are no unstressed syllables separating the two peaks, the dip can be accommodated between all peak combinations, except for 'late + early/medial', where a hat pattern is created; it combines the rise of the 'late peak' and the fall of the 'early peak' in a two-stress sequence.

This absence of an F0 descent between peaks can also be extended to concatenations with intervening unstressed syllables. In such a hat pattern, an early peak is not possible initially, and a late one is excluded non-initially. If there are more than two stresses incorporated in a hat the non-initial and non-final ones are unspecified as to peak position because they neither have a rise nor a fall but are simply integrated into the downstepped sequence of peak maxima. In the categorization of pitch patterns they are nevertheless grouped together with peaks. If in a two-stress rise-fall it is difficult to decide whether the rise represents a valley, or a late peak in a hat pattern, the latter solution is chosen.

When prosodic boundaries intervene any sequencing of peaks and/or valleys is possible, but the hat pattern is then excluded since it represents a very high degree of cohesion. On the other hand, a late peak with a full F0 descent marks a dissociation from a following peak and will then normally be linked with a prosodic boundary, i.e. final lengthening and F0 reset afterwards.

Unstressed syllables preceding the first sentence stress in a prosodic phrase may be either low or high: they represent different types of pre-head.

Declination, i.e. the temporally fixed decline of F0 has been replaced by downstepping in KIM, i.e. a structurally determined pitch lowering from sentence stress to sentence stress, independent of the time that elapses between them.

Prosodic boundaries

One of the functions of prosody is the sequential structuring of utterances and discourse. Two categories of phrasing have been set up so far: [PG1] corresponding to prosodic sentences and [PG2] related to prosodic phrases. Both are always phonetically signalled by lengthening before them, and usually by F0 resetting after them. Asides and paren-

thetic insertions have no F0 resetting in spite of other clear phrasing marker signals. Contrariwise, F0 resets may occur at other points than the phrasing markers [PG1,2]. [PG1] also coincides with high syntactic structure nodes, whereas [PG2] does not. Both may be further strengthened by the incidence of pauses and intonation patterns. Full F0 peak descents are particularly frequent with [PG1], and [?] as well as [.?] are only associated with this phrasing marker.

SYMBOLIZATION OF THE MODEL CATEGORIES

The symbolic labelling system has to meet the following requirements:

- unequivocal representation of the categories of the prosodic phonology
- integration into 7 bit ASCII segmental label files
- integration into 7 bit ASCII orthographic files of German text
- clear typographic separation from the segmental labelling allowing prosodic notations on the same tier for convenient cross-reference between segmental and prosodic aspects of speech
- mnemonic ease for learning and use.

The application of these guiding principles has resulted in the standardization of the following repertoire and conventions [4] for insertion in orthographic text or segmental phonetic files.

- Apostrophe and quotation mark ['], ["] for lexical stress are put in front of the primary or secondary stress vowel; unmarked vowels are unstressed. In a segmental label file these stress markers are linked to the vowel symbol, in an orthographic file they are inserted in logical order before, and on the same time mark as, the vowel. Function words, marked by suffixed [+], do not get a lexical stress symbol by default; if they receive sentence stress, double apostrophe [' '] is inserted before the vowel of the appropriate syllable.
- Digits [3],[2],[1],[0], when not combined with punctuation marks, refer to sentence stress. They are put in logical order before words that receive the reinforced, accented, partially or completely deaccented sentence stress category. The lexical stress position then determines where F0 contours have to be hooked.
- Punctuation marks [.],[,],[?] refer to pitch peaks, low and high rising valleys, and the character sequences [,] and [.?] to the corresponding fall-rises. They are put in logical order before a prosodic boundary or before the next sentence-stress digit [>1]. [(.)?] can only occur before a prosodic boundary.
- Parentheses [)],[(] refer to early and late peaks or early and non-early valleys and are put after the sentence-stress digit; the medial peak is marked by the absence of these symbols. Digit and parenthesis form a symbolic unit.
- The pitch movement between successive peaks or between a peak and a boundary may be a full or an intermediate F0 descent or a level F0, symbolized by digits [2],[1],[0] before [.]. Digit and punctuation mark form a symbolic unit.
- Downstep is not marked. F0 reset is implied by a prosodic boundary; in the case of its absence, [=] is prefixed to the next digit [=>2]. If reset occurs at other points than boundaries, [+] is prefixed to the next stress digit [=>2].
- A high prehead is marked by [HP] at the beginning of an utterance or after a phrase boundary.
- Prosodic phrasing markers [PG1] and [PG2] are put after punctuation marks at the appropriate places.
- Only speech rate changes in relation to the speed in the preceding prosodic phrasing unit are marked: [RP] and [RM] (= 'rate plus/minus') are put after [PG1,2] (and before [HP]). An absolute rate judgement at the utterance onset may be added at a later labelling stage.
- Disfluency markers are
  - [z:] for hesitation lengthening at the end or inside of a word
  - [/+] or [=/+] for break-offs and resumptions at word boundaries and within words, respectively.
- Markers for segmental phrase-level units are [p:], [h:] (= pause, breathing), [l:], [s:] etc. (= laughing, clicks etc.) [4].
- All non-segmental prosodic markers are without duration; they are put on the same time mark as the beginning of the next segmental unit.

LABELLING PROCEDURE

A labelling platform has been created at IPDS by M. Patzold on an AT, running on UNIX and equipped with a sound card, which accepts segmental label files, generated by the KTH MIX programme, and F0 analysis data as input, allows the display of F0 contours and labels, as well as the insertion, deletion and change of prosodic labels under auditory and visual control. The default sentence stress markers [2] for content words and [0] for function words and a general prosodic phrasing marker [PG] are inserted automatically on the basis of the segmental labels. The manual labelling then proceeds in cycles dealing with one prosodic domain after another. The result is a label file that integrates prosodic labels into the segmental strings. The following orthographic transcript with prosodic annotations (rather than a complete label file, to reduce the amount of information and for greater ease of intelligibility) provides an illustration of the prosodic labelling of a spontaneous dialogue from the Kiel Corpus of Spontaneous Speech [1].

g071a004.s1h
TIS004:
2 <{hm> **PG2 2(** D'ienstag **0** w}rde+
0 mir+ **0** g'ut **1. 2)** p'assen **2. PG1**
2 <{hm> **PG2 0** das+ **2** h'ei~t **, PG2 p:**
2 Mom'ent **1. PG2 2** 'allerdings **,**
2 'erst **z: , PG2 2(** n'achm"ittags **h: 2. PG1 RP HP 0** das+ **0** wird+ **0** dann+
2 wahrsch'einlich **0** 'n+ **0** b'i~chen **0.**
2) schw'ierig **2. PG1 2** D'ienstag **1.**
2 m'ittwochs **z: 1. PG2 RM** <{h> **PG2 p: 0** is=/+ **0** s'ieht **0** das+ **0** bei+
2 m' 'ir+ **z: 0** sch=/+ **2. 2** schw'ierig
0 'aus **2. PG1 RP 0** da+ **0** hab' **0** ich+
=2 tags'}ber **1. 2** Term'ine **1. PG1 RM**
h: 2 <{hm> **PG2 HP 0** wie+ **0** s'ieht
0 das+ **0** bei+ **2** ' 'Ihnen+ **0** am+ **1.**
3 D'onnerstag **0** 'aus **2. PG1**

Prosodic label files can now be the input to the RULSYS/INFOVOX TTS system for German, which also contains an implementation of KIM [3], to test the adequacy of the manual labelling by comparing its rule synthesis with the original. Prosodic modelling, prosodic labelling and prosodic synthesis thus form an integrated framework of prosodic research at IPDS Kiel. The prosodic categories, being related to human sound production beyond the particular language phenomena found in German, are transferable to the description of other languages, and PROLAB may be used more generally in prosodic labelling.

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.

REFERENCES

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

[2] Kohler, K.J. (1991): "A model of German intonation", *AIPUK*, vol. 25, pp. 295-360.

[3] Kohler, K.J. (forthcoming): "Parametric control of prosodic variables by symbolic input in TTS synthesis", 2nd *ESCA/IEEE Workshop on Speech Synthesis*, Sept 1994, New Paltz, N. Y.

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

Multi-language Speech Database: Creation and Phonetic Labeling Agreement

Terri Lander, Beatrice Oshika, Ronald A. Cole, and Mark Fanty

Center for Spoken Language Understanding
Oregon Graduate Institute of Science and Technology
Portland, Oregon USA
tlander@cse.ogi.edu

ABSTRACT

The focus of the paper is the evaluation of inter-labeler reliability on broad phonetic transcriptions when labelers do not necessarily know the language they are labeling. We provide an analysis of label disagreements, presenting results from six languages, English, French, German, Japanese, Spanish, and Vietnamese with a total of 2 minutes of continuous labeled speech. Labeler agreement across languages ranges from 41 percent with detailed label to label comparisons to 91 percent when less fine comparisons were made.

INTRODUCTION

This paper describes research on a large multi-language speech database being collected at the Oregon Graduate Institute (OGI). The Center for Spoken Language Understanding (CSLU) at OGI has been developing multi-language telephone speech corpora for the last 5 years. An earlier corpus [1] contained data from 11 languages with 90 speakers per language. Presently a 22 language corpus with over 200 native talkers per language is being collected with a wide representation of language types: Eastern Arabic, Cantonese, Czech, English, Farsi (Modern Persian), French, German, Hindi, Hungarian, Japanese, Korean, Malay, Mandarin, Italian, Polish, Portuguese, Russian, Spanish, Swedish, Swahili, Tamil, and Vietnamese. The corpus consists of short responses to 21 questions plus extemporaneous responses up to 60 seconds long. The corpus will be donated to the National Institute of Standards and Technology and to the Linguistic Data Consortium.

Each call is verified by two native talkers who verified that the caller followed the instructions to each prompt, and made judgments as to regional accent, language competency (fluency), age of talker, telephone line quality, background noise and call completion.

Up to one minute of spontaneous speech and responses to questions are being transcribed at the orthographic level by two native talkers, with disagreement resolution. A standard method for transcribing continuous speech, including pauses and non-speech sounds has been developed [2]. In addition, trained linguists will label two one-minute sections from each language at the broad phonetic level using Worldbet [3]. [1]

An earlier study [4] compared agreement of broad phonetic labels by both native and non-native talkers of five different languages. Label agreement between native speakers averaged 68%, while agreement between non-natives was much less consistent at 34%. This paper reports on results from labeled speech in six languages, and includes an analysis of phonetic categories on which labelers most disagree, with possible explanations of such variation.

[1] Phonetic label sets were developed by Dr. James Hieronymus, author of [3].

TRANSCRIPTIONS

Transcription was supported by the OGI speech tools [5] which display the waveform and corresponding spectrogram. Transcribers were able to play any part of the waveform multiple times as needed. The labelers used Worldbet, an ASCII rendering of the IPA for broad phonetic transcriptions.

Worldbet attempts to represent phonetic and phonemic distinctions within a single level of transcription. Base symbols generally capture phonetic detail that might otherwise be described by rule, e.g., the Spanish stops /d/ and /t/ are transcribed in Worldbet as explicitly being dental: d[and t[. Diacritics are used to label allophonic variations. A nasalized vowel /i:/ in English would be i:_~ but nasalized vowels which are phonemic in the language, such as the French nasalized vowels, are transcribed A~ where nasalization is part of the base symbol, not a diacritic.

Little prior discussion went into specific labeling and segmentation conventions, although the transcribers did label and compare 10 seconds of speech per language to gain a basic familiarity with each language and speaker. Orthographic transcriptions produced by native speakers were also available to the phonetic labelers to assist in decisions about the choice of base symbol. These were useful when the transcribers were not familiar with the languages.

TRANSCRIBERS

The two labelers are trained in phonetics and acoustics. Both are native speakers of English and are familiar with Spanish. They have less or no knowledge of the other languages labeled. Both of the transcribers have had extensive experience labeling speech.

DATA

The data transcribed for this experiment were a subset of the OGI 22 Language Telephone Speech Corpus described in the introduction. Three 10-second segments of continuous speech were selected for English, German, French, Japanese, Spanish, and Vietnamese. The data selected were gender balanced.

Two ten-second segments of speech in each language (a total of 12 ten-second segments, or two minutes of speech) were labeled independently by the two transcribers.

ANALYSIS

Inter-transcriber agreement was measured in terms of the number of substitutions, deletions and insertions required to map one transcription to another. The "reference" transcription was chosen arbitrarily.

When computing the mapping, overlap in time and phonetic similarity were considered when deciding which segments were substituted, inserted and deleted. This occasionally resulted in a very slightly smaller accuracy than the optimal. However, it results in much more accurate and meaningful confusion matrices. Accuracy was computed as follows:

$$ACC = (ref - sub - ins - del)/ref$$

where ref, sub, ins, and del represent total number of reference segments, substitutions, insertions, and deletions, respectively.

The average accuracy for the set of files in each language was computed using the average number of reference segments, substitutions, insertions and deletions over both of the files.

Six scores were calculated, using the original labels and five less fine sets.

Original Labels To facilitate the anal-

ysis, all non-speech labels were mapped to a single symbol, and adjacent, identical symbols were collapsed.(Table 1 Column 1)

Diacritic Stripping This is a reduced symbol set produced by stripping diacritic information but maintaining the base symbol (Table 1 Column 2).

Broad category We reduced labels into: vowel, plosive, fricative, approximant, nasal, and non-speech (Table 1 Column 3).

Vowel Agreement Additional analysis was performed that clustered vowels by place of articulation. Diphthongs were not included unless the place of articulation fell entirely within the space defined by the cluster. Three different vowel sub groupings were used; all non-vowel sounds were removed from the files so that the scoring algorithm would reflect errors in vowel category only:

1. high, mid, low (Table 2, Column A)
2. front, central, back (Table 2, Column B)
3. high-front, high-back, central, low-front and low-back. (Table 2, column C)

RESULTS

Table 1 displays results for three label comparisons. As expected, agreement improves as the distinctions within the symbol set are reduced.

Table 2 compares agreement between different vowel reductions based on place of articulation.

Table 3 displays the number of base symbols and the number of vowel symbols available per language.

Table 4 displays various usage patterns of original labels.

Table 1: *% Average transcriber agreement at three levels of precision: 1) original labels 2) diacritic stripped 3) broad category*

	1	2	3
Eng	55(143)	67(124)	83(143)
Fre	59(119)	60(97)	83(119)
Ger	41(122)	52(105)	74(122)
Jap	72(143)	77(118)	91(143)
Span	71(107)	78(94)	86(107)
Viet	60(104)	68(84)	84(104)
ave	59	67	84

Table 2: *% Average transcriber agreement for three levels of vowel reduction: A) high, mid low, B) front, central, back, C) high-front, low-front, central, high-back and low-back D) contains the average number of reference segments*

	A	B	C	D
EN	55	54	52	38
FR	62	73	62	45
GE	52	61	54	38
JA	75	78	77	57
SP	85	86	81	47
VT	52	62	56	27
ave	67	71	66	

DISCUSSION

As a follow up to [4] we wanted to do a more careful error analysis of labeler disagreement. In the present experiment, labeler agreement across languages ranges from 41 percent with detailed label to label comparisons to 91 percent when less fine comparisons were made. This compares to 33% and 83% in [4]. Perhaps using orthographies in addition to labeling and comparing test data prior to actual labeling helped to raise over all agreement.

Lower agreement with the full label set (Table 1) seems to result in part from convergence on "preferred" but differing sets of symbols. This happened with various symbols (see Table 4). L1 preferred the devoicing diacritic, using it 24 times more often than L2. L2 used the nasalization diacritic 22 times more often than L1. L1 used closure labels 44 times more often than L2.

Over specificity factored in to some of the disagreements. L1 used an average of 6.1 (5%) more symbols per file than L2, using from -3 (Spanish) to 21 (English) symbols more than L2.

The variability in vowel comparisons (Table 2) seem to be related to the number of vowel labels available to transcribers for each language. Spanish and Japanese, both with relatively small vowel inventories, represented the greatest agreement. Although English and Japanese had the same number of vowels (Table 3), there were actually only 7 places of articulation represented in the Japanese vowel labels, as five of the Japanese vowels differ only in length.

Label inventory seems to influence agreement more than knowledge of the language, because although transcribers were familiar with Spanish and English, they agreed more often in Spanish, with its smaller label inventory.

In the future we plan to expand this experiment by labeling a larger set of languages, more speech per language, and a variety of speakers in each language. We also plan to further analyze the role played by the orthographies for non-native transcribers.

Table 3: *Number of base symbols available (BL); number of vowels and diphthongs available (VL)*

	EN	FR	GE	JA	SP	VT
BL	63	47	72	70	41	64
VL	20	17	30	17	12	28

Table 4: *Specific examples of label divergences: the number of times each symbol (base label(b) or diacritic(d)) was used by each labeler (not necessarily simultaneously.)*

	L1	L2
closure(b)	222	178
schwa(b)	87	64
devoicing(d)	28	4
nasal(d)	3	35
centralize(d)	10	0

ACKNOWLEDGEMENTS

This research was funded by U S WEST, the Office of Naval Research, and the National Science Foundation.

REFERENCES

[1] Y.K. Muthusamy R.A. Cole, and B.T. Oshika, "The OGI Multi-language Telephone Speech Corpus", *The International Conference on Spoken Language Proceedings,* Banff, Alberta, Canada, Oct 1992, pp 895-898.

[2] T. Lander, S.T. Metzler, "The CSLU Labeling Guide", CSLU, Oregon, February, 1994.

[3] J.L. Hieronymus. "Ascii phonetic symbols for the world's languages: Worldbet". AT&T Bell Laboratories Technical Memo, 1994.

[4] R. Cole, B.T. Oshika, M. Noel, T. Lander, M. Fanty, "Labeler Agreement in Phonetic Labeling of Continuous Speech", Proceedings ICSLP94, Yokohama, Japan, September 1993, pp. 2131-2134.

[5] CSLU. "OGI speech tools user's manual," Technical report, Center for Spoken Language Understanding, Oregon Graduate Institute, 1993.

BROAD TRANSCRIPTION IN PHONETIC TRAINING

Michael Ashby, Patricia Ashby§, John Baldwin*,*
Frederika Holmes§, Jill House*, John Maidment**
**Department of Phonetics and Linguistics, UCL*
§University of Westminster

ABSTRACT

In this paper we question some tradtional assumptions made about the status of the broad transcription. We advocate a relaxation of strict phonemic constraints in favour of principles of phonetic salience and recoverability.

1. INTRODUCTION

Part of a training in phonetics involves learning to make 'broad' transcriptions of running speech, based on written or spoken texts. Although inspired by a phonemic approach to phonology, in practice the transcription system taught to students of English phonetics, for example, is usually not a strictly 'phonemic' one. The systems made popular in Britain by Jones and Gimson were never strictly phonemic, and the most recent development of the *de facto* standard [1] is better described as polysystemic. We advocate a more explicit recognition of this development and favour yet further relaxation of the requirement to transcribe 'phonemically'. We explore the difficulties faced by learners of English phonetics and propose an approach to transcription more in line with their real needs.

2. BROAD TRANSCRIPTION

2.1 What is it?

The following sample illustrates a traditional broad transcription:

/ðeə wəz 'wʌns ə jʌŋ 'ræt kɔ:ld 'ɑ:θə | hʊ wʊd 'nevə teɪk ðə 'trʌbl tə 'meɪk ʌp ɪz 'maɪnd | wen'evər ɪz 'frenz 'ɑ:st ɪm ɪf ɪ wʊd 'laɪk tə gəʊ 'aʊp wɪð ðəm | hɪ wʊd 'əʊnlɪ 'ɑ:nsə | aɪ 'dəʊnt 'nəʊ | hɪ 'wʊdnt seɪ 'jes | ən ɪ 'wʊdnt seɪ 'nəʊ | 'aɪðə | hɪ kəd 'nevə 'lɜ:n tə 'meɪk ə 'ʧɔɪs/

The symbols used here are those of the *English Pronouncing Dictionary* (*EPD14*) [2]. The representation of prosodic information is restricted to a simple stress mark and word group boundary symbol. Like many other British phoneticians, we are in the habit of calling this type of transcription 'phonemic' or 'broad', though in terms of Abercrombie's [3] analysis it probably qualifies as neither:

> "A transcription which is made by using letters of the *simplest possible shapes*, and in the *smallest possible number*, is called a SIMPLE PHONEMIC transcription. It is called 'simple' because of the first characteristic, and 'phonemic' because of the second." (p. 17)

The *EPD14* transcription fails to be 'simple' or 'phonemic' on both criteria. Representations such as /aɪ/, /æ/, /ɔɪ/ are not notationally the simplest; some symbol combinations, such as /eɪ/ and /əʊ/, are not even uniquely parsable into segments. Nevertheless, the intention is to represent units identified by classical commutation and substitution procedures, and to provide symbols for all and only those sounds which potentially distinguish words: the phonemes.

The term 'broad' has acquired a range of meanings ([3]: p. 35; [4]: ch.18). The *EPD14* transcription qualifies as 'broad' insofar as relatively little phonetic detail is shown. For example, conditioned variation in the laterals in *like*, *called*, *only* and *trouble* is not explicitly represented in the sample above.

2.2. Why teach it?

Producing a plausible broad transcription requires both phonological awareness and a range of auditory skills:

- segmentation of the speech signal into a sequence of 'discrete' sounds;
- identification of and discrimination between sounds;
- use of a language-specific, finite set of symbols to represent them;
- independence from orthographic prejudice;
- objectivity about accent and style variation;
- awareness of the difference between citation forms and connected speech;
- production of a faithful record of a particular rendering of the passage in question;
- the ability to disregard insignificant phonetic differences, and to group sounds together into functionally equivalent classes.

Auditory skills are required directly or indirectly when transcribing from speech or text. A basic aim of the transcription exercise is auditory training [5].

Analytic skills are involved in tapping intuitions about the phonological system and in rescrutinizing the judgments of sameness or difference which underlie the analysis.

2.3 Who learns it?

The weight attached to each of the theoretical and practical skills outlined above should depend on the purpose for which the student is learning English phonetics. Typically, such students will include specialists in linguistics, speech science, speech pathology, English language, EFL, modern languages, performing arts. The groups may include both native and non-native speakers of English, and speakers of a wide variety of native accents.

For some groups the theoretical aspects will outweigh the practical ones, but for others the situation will be reversed. For example, students of English as a foreign language need be less concerned by phonological theory and may concentrate on using transcription as a tool to improve their pronunciation. Linguistics students, on the other hand, need to develop their abstract analytical skills. Speech and language therapists need to be proficient in both aspects if they are to diagnose phonetic and phonological immaturities and disorders.

3. LEARNERS' DIFFICULTIES

Students new to phonetics differ widely in their natural ability to master transcription skills and in their basic auditory discrimination skills. Some may find difficulty in relating the transcription conventions to their own accent. Others take longer to adjust to the conventions, either because they find them difficult to grasp theoretically, or because of some seemingly arbitrary property they exhibit.

In demonstration conditions, using isolate words in their citation form, most learners can quickly discriminate between linguistically contrastive sounds in their own language or variety, and recognise the need to move away from orthography. Identifying and symbolising the sounds in connected speech takes practice, since knowledge of citation forms may interfere with the direct translation between sound and symbol.

When using text, we encourage native-speaker students to transcribe their own variety of English. This may initially complicate life for the non-RP student who has also to transcribe from dictation by an RP-speaking lecturer, but pays off eventually by reinforcing observations and understanding of the differences between accents. A sizeable proportion of our students are speakers of the near-RP of SE England, and regular exposure to this variety during their stay in London has its influence on accents from further afield too.

The ability to ignore irrelevant phonetic detail is often achieved surprisingly painlessly, suggesting that learners can use intuitions to access and exploit phonological knowledge they already have. But in certain cases there is a tension between finding a symbolically accurate representation of a sound and confining oneself to the contrastive system: notably where allophones in complementary distribution have a highly salient difference in realisation.

4. ARBITRARY CONVENTIONS

In trying to reconcile transcription conventions established largely for RP with their own perceptions, students regularly experience difficulties such as:
(i) selecting a symbol to represent the weak vowel in e.g. *happy* or *mediate*;
(ii) using a /t/ when what they clearly hear is a glottal stop [ʔ], e.g. *what*;
(iii) using /l/ (rather than a close back vowel or /w/) when there is clearly no lateral consonant present, e.g. *milk*.
(iv) selecting a symbol for the vowel in words like *old* (the diphthong with the quality [ɒʊ] is often attributed to the phoneme /ɒ/ rather than /əʊ/).
(v) using the diphthong /eə/ for the sound they produce and perceive as a long monophthong [ɛ:] in words such as

bare and *bared.*

A new set of conventions has become established for (i), involving the use of weak /i/ rather than /iː/ or /ɪ/ (assuming that otherwise redundant length marks are used in the basic symbol list). Introduced by Gordon Walsh in [6], this practice has been extended to cover the use of /u/ rather than /uː/ and /ʊ/ in [7], [1] and [8]. In encouraging it, we are accepting the use of a symbol which is not on the usual phoneme list, but which represents a realisation which could be a neutralisation between /iː/ and /ɪ/. This clearly violates the strict phonemic criteria of traditional practice.

Deviation from traditional practice in respect of (ii) - (iv), where we are arguably dealing with allophones in complementary distribution which are appropriately represented with the same symbol, is not yet widely accepted. Unlike the situation in (i), there is always a possible phonemic solution in transcription. But difficulties will still arise: for some, [ʔ] may at times be a neutralisation between /p/, /t/, /k/; for others, words like *doll* and *dole* are genuine homophones, making it intuitively unsatisfying to symbolise the vowel differently, even where a difference remains for other forms, such as *dolling* and *doling.*

We have to consider whether it is helpful to insist on a transcription which is an exercise in phonemic theory, or whether we should applaud the ability of learners to identify and symbolise more precisely the sounds they hear.

In seeking to justify relaxing the traditional requirements to maintain a strict distinction between phonemic and allophonic levels, we should examine the conventions which have long been accepted for other problem areas in broad transcription.

The theoretical purity of the phonemic transcription is a myth. Phenomena where contrast, distribution and native speaker intuitions do not lead to a unique solution will remain. Since such problem areas have been abandoned rather than solved theoretically, the conventions established for dealing with them must be treated as arbitrary. Students are therefore learning to fall in with theoretically dubious conventions.

Let us look more closely at the arbitrary 'solutions' proposed for a couple of these areas of conflict.

(1) status of schwa: it can be argued that schwa in RP should be analysed as a weak, non-contrastive variant of some other vowel in an unstressed context. At a lexical level, the allophonic status of the vowel may not be transparent, unless alternating forms, or alternative pronunciations, suggest that the strong form of the vowel would be different under stress. Compare *photograph* /ˈfəʊtəgrɑːf/ and *photography* /fəˈtɒgrəfi/. The case may be more clearly made with respect to *weak forms*: schwa can be regarded as a conditioned variant of /ʊ/ in *from*, of /æ/ in *have*, of /ʌ/ in *but*, and so forth. Transcribing such forms with /ə/ thus explicitly incorporates allophonic variation in a broad transcription (though schwa can of course be in contrast with other weak vowels). However, to deny its use in transcription would fundamentally alter the status of the broad transcription in English, since the use of schwa is not predictable in all unstressed contexts, and the source pronunciation would cease to be reliably recoverable.

(2) assimilations: where the output of an assimilation corresponds to a realisation consistent with a different phoneme, we conventionally encourage students to show this in broad transcription: thus *ten men* is represented /ten men/ etc. But at the same time we ignore other assimilatory processes where the output, though phonetically distinct, does not cross a phoneme boundary -- an arbitrary distinction which obscures the theoretical generalisations relevant to assimilation. Disallowing assimilations in broad transcription, on the grounds that they are contextually determined, would seemingly be more consistent with phoneme theory, but would greatly impoverish the transcription's explicitness.

Furthermore, what are we saying by allowing an assimilation like *ho*/p/ *potato* in a broad transcription, but disallowing *ho*[ʔ] *potato* (insisting on /t/)? In the latter case, students are being asked to disregard auditory evidence and a freshly discovered ability to discriminate between different articulations in favour of a theoretical point. By clinging too hard to the theoretical point we lose

explicit recoverability.

5. CONCLUSIONS

Relaxation of the requirement of strict adherence to phonemic theory in broad transcriptions undertaken for the purposes of phonetic training should be guided by two principles which we call **phonetic salience** and **recoverability**.

Phonetic salience refers to situations where the sound perceived or produced by the learner is markedly different from that normally represented by a symbol chosen from the range available for a strict phonemic transcription: for example, the use of [ʔ] for phonemically sanctioned /t/, or a vocalic segment of the [o] variety for phonemic /l/.

Recoverability can best be explained by reference to the ideas put forward in [3] regarding the interaction between the text of a transcription and the conventions necessary for its interpretation: "any departure from a simple phonemic transcription has the effect of transferring information to the text from the conventions" (p. 23). The type of convention recognised by Abercrombie which is relevant to our argument is what we would probably now call an allophonic rule: it specifies the contextually determined interpretation of a phonemic symbol.

What Abercrombie does not propose is that this sort of convention should be further subdivided into those which are exceptionless and those which are variable. For example, while it is the case for many accents of English that /l/ should be interpreted as [l] before vowels and [j], and as [ɬ] elsewhere, and that this variation is entirely predictable, the use of [ʔ] vs [t] is much less certain for many speakers. On a given occasion it may be impossible to recover which variant was used without including the information in the text of the transcription.

Our proposal then, for south-eastern English and near-RP, is that the usual set of symbols employed for a broad transcription should be augmented to allow for the explicit symbolisation of phonetically salient variants which are not recoverable by general, exception-free rules. Exactly how great the increase in the number of symbols should be will depend on the experience and particular requirements of the learner. Inclusion of [ʔ], [o] and [ɒʊ] in the symbol set appears to us an indispensable minimum for most groups of students. Introducing further modifications for other optional variants may well be worth considering.

The following incorporates some of the innovations discussed above.

/ðɛː wəz ˈwʌns ə jʌŋ ˈræʔ kɔːod ˈɑːθə | hu wʊd ˈnevə teɪk ðə ˈtrʌbo tə ˈmeɪk ʌp ɪz ˈmaɪnd | wenˈevər ɪz ˈfrenz ˈɑːst ɪm ɪf i wʊd ˈlaɪk tə gəʊ ˈaʊʔ wɪð ðəm |hi wʊd ˈəʊnli ˈɑːnsə | aɪ ˈdəʊnʔ ˈnəʊ | hi ˈwʊdnʔ seɪ ˈjes | ən i ˈwʊdnʔ seɪ ˈnəʊ | ˈaɪðə | hi kəd ˈnevə ˈlɜːn tə ˈmeɪk ə ˈʧɔɪs/

The same principles of phonetic salience and recoverability should apply to the transcription of other varieties of English and of other languages. If we cease to pay lip-service to the idea of a phonemic analysis as the basis for our transcription, the decisions about what to include do not in fact become entirely arbitrary. A broad, but principled transcription can be guided by the criteria outlined above.

REFERENCES

[1] Wells, J.C. (1990), *Longman Pronunciation Dictionary*, London: Longman.

[2] Ramsaran, S. (1988), *Everyman's English Pronouncing Dictionary*, 14th edition. Reprinted with revisions and a supplement by S. Ramsaran, London: Dent.

[3] Abercrombie, D. (1964), *English Phonetic Texts*, London: Faber & Faber.

[4] Laver, J. (1994), *Principles of Phonetics*, Cambridge: Cambridge University Press.

[5] Jones, D. (1948), "The London School of Phonetics", *Zeitschrift für Phonetik* 2, pp. 127-35.

[6] *Longman Dictionary of Contemporary English*, 1st edition, (1978), Harlow and London: Longman.

[7] Roach, P. (1983/1991), *English Phonetics and Phonology: a practical course*, Cambridge: Cambridge University Press.

[8] Ashby, M. (1994), (Phonetics editor) *Oxford Elementary Learner's Dictionary*, Oxford: Oxford University Press.

AGREEMENT IN CONSENSUS TRANSCRIPTIONS OF TRAINED AND UNTRAINED TRANSCRIBERS

W.H. Vieregge and A.P.A. Broeders***

*Department of Language and Speech, University of Nijmegen, Netherlands
**National Forensic Science Laboratory, Rijswijk, Netherlands

ABSTRACT

Consensus transcriptions were made by trained as well as untrained transcribers of several segmental variables in Dutch. A randomly selected subset of these variables was transcribed twice by both groups. Two hypotheses were tested: the degree of agreement between non-contemporary consensus transcriptions is a measure of their validity; trained transcribers reach higher consistency levels than untrained transcribers.

1 INTRODUCTION

In her discussion of the meaning of the terms validity and reliability as applied to phonetic transcription, Cucchiarini [1] suggests that, in the absence of a proper benchmark for the estimation of the validity of a transcription, the consensus transcription may serve as a viable alternative. The consensus transcription is often proposed as a procedure which will reduce errors in transcriptions and increase agreement among transcribers (Shriberg et al. [2]). We have found that the consensus transcription can serve as a suitable format for the analysis of intra- and interspeaker variation in the realization of certain segmental variables in Dutch (Vieregge and Broeders [3]). However, we are not aware of the existence of studies in which the agreement between consensus transcriptions was examined to see if this would produce a more satisfactory measure of transcription validity.

2 AIM OF THE STUDY

The main aim of the investigation was to look into the possibility of testing two hypotheses, both of them inspired by our experience with the consensus transcription and following from the claim that this transcription procedure tends to reduce errors due to inattention, and leads to greater agreement between transcribers (Ting et al. [4]). If this is true, the degree of agreement found in consensus transcriptions made at different points in time should provide a good measure of the validity of these transcriptions. In other words, we hypothesize that consensus transcriptions are more valid as they are replicated with greater consistency.

On the assumption that trained transcribers may be expected to be more competent than untrained tran-scribers, a second hypothesis can be formulated, viz. that trained transcribers will reach a higher degree of consistency than untrained transcribers.

In order to test these hypotheses consensus transcriptions made as part of a study of inter- and intraspeaker variation in the realization of segmental variables in Dutch were used.

3 THE SPEAKERS

The speech samples were produced by 7 educated speakers of Dutch, hailing from various parts of the country. The amount of regional variation in their speech varied from hardly any to quite marked. All speaker were male, with ages ranging between 25 and 50. The speech style could be described as quasi-spontaneous: all seven speakers were asked to give a description of what they saw in three drawings, showing a street scene, some shops and a living-room respectively. The duration of their descriptions varied from 2 minutes to 2 minutes and 45 seconds. The material forms part of a larger corpus collected for a different purpose by our colleague Van Bezooijen, who kindly made the recordings available to us.

4 THE VARIABLES

The segmental variables used in this investigation form a random subset of the larger set of variables transcribed as part of a study to look into the inter- and intraspeaker variation of certain segmental variables in Dutch (Broeders and Vieregge [5]. They are presented in Table 1.

Table 1. Variables used in the investigation (N: the number of tokens per variable in the subset).

Variable	N
/x/	21
/z/	14
/v/	14
schwa-insertion after /r,l/	13
assimilation of voice before /b,d/	14
n-deletion after schwa	14

The variables themselves were selected as part of the earlier study on the basis of their expected variability in Dutch. The subset of tokens used in the present study was picked at random.

5 THE TRANSCRIBERS

Consensus transcriptions were made by two trained transcribers, the present writers, and nine pairs of untrained transcribers. The latter were all Language and Speech Pathology students of the University of Nijmegen, all of them qualified speech therapists, who made the transcriptions in part fulfilment of the requirements of a 120-hour course in phonetic transcription taught by the first author. They were instructed to produce a consensus transcription in accordance with the IPA conventions [6], which they were told would later be assessed by their teacher.

6 PROCEDURE

The trained transcribers made the second transcription of the random subset several months after the first. For the untrained transcribers both transcriptions were made as part of a single transcription assignment but the work was structured in such a way that, unlike the trained transcribers, they may be assumed to have been unaware of the fact that they were transcribing (some of) the variables twice.

7 RESULTS

The results are presented in Table 2. Transcriptions were considered to be in agreement if the same phonetic symbol plus any of a limited number of diacritics was used on both occasions. They are expressed as the percentage agreement reached per variable. The percentages given for the untrained transcribers are averaged for the 9 pairs.

Table 2. Variables used in the investigation (U: untrained, T: trained transcribers; N: number of tokens per variable in the subset).

Variable	U	T	N
/x/	64.6	76.2	21
/z/	78.6	92.9	14
/v/	69.0	92.9	14
schwa-insertion	80.3	92.3	13
assimilation	69.0	71.4	14
n-del	90.5	85.7	14

8 DISCUSSION

It appears that, with the exception of the last variable, trained transcribers achieve considerably more agreement than untrained transcribers. The difference in the amount of agreement found between trained and untrained transcribers is significant ($t = -2.44$; $p < 0.05$; one-tailed).

At first sight, the results seem to confirm the second hypothesis that trained observers reach higher consist-

ency levels than untrained transcribers. However, inspection of the actual transcriptions suggests that there are one or two complicating factors at work whose effects, while undeniably present, are difficult to quantify. On the one hand, there is the fact that some of the variables are essentially binary (n-deletion, schwa-insertion). Obviously, all other things being equal, agreement is likely to be higher if the number of options is small and vice versa. On the other hand, there are variables like /x/ that easily run into as many as 5 different symbolizations, each combining with several diacritics. Of course, in principle this embarras de choix applies to trained and untrained transcribers alike. In practice, however, it must be expected to work against the trained transcribers, as their greater familiarity with the phonetic symbol set and greater experience as trained listeners should make more options available to them. By the same token, untrained listeners are likely to reach higher agreement between transcriptions because they have a smaller set of symbols to choose from. On balance though, the results lend support to our second hypothesis: agreement between consensus transcriptions is higher for trained than for untrained transcribers.

However, in the course of the discussion we have seen that there are strong indications that our first hypothesis is not tenable as it stands. Agreement per se is a necessary but not a sufficient criterion for validity. It is simply not the case that the consensus transcription that happens to show the highest degree of agreement is for that reason also the more valid one. What is essential of course is that the consensus transcriptions are made by competent transcribers. If agreement is high between non-contemporary replications of consensus transcriptions by experienced transcribers it is reasonable to assume that these can be used as a criterion against which the quality of other transcriptions can be measured.

9 A VALIDITY CRITERION

If we revise our hypothesis in the light of these observations, we are in a position to judge the quality of the consensus transcriptions made by the pairs of untrained transcribers, using the consensus transcriptions of the trained transcribers as our criterion (Vieregge [7], p. 31). Obviously, this will only be possible for those cases where the trained transcribers produced identical transcriptions in the two consensus sessions. While it is clear that this introduces a degree of inaccuracy in those cases where the trained transcribers disagree between the two sessions, it is safe to assume that the effect of this is marginal. After all, for most variables the agreement scores reached by the trained transcribers are quite high, and what discrepancies do arise will by and large occur in respect of the transcription of the rather more problematical variables, on which untrained transcribers would be unlikely to do better in the first place.

10 THE VALIDITY CRITERION APPLIED

If we apply the above criterion to the transcriptions made by the untrained transcribers this yields two types of information. First, we can calculate the score for each variable averaged over the nine pairs of untrained transcribers. This figure expresses the extent to which the untrained transcribers, on average, produced transcriptions that are identical to those of the trained transcribers. It gives an indication of how well the variable in question was transcribed by the untrained transcribers. The results are presented in Table 3, which also specifies the number of tokens for each variable transcribed identically by the trained transcribers and used in the validity criterion. It is worth noting that on average the transcription of the variables /x/, /v/ and *assimilation* deviates in the majority of cases from that of the trained transcribers, which may be taken as an indication of the difficulty these variables present.

Table 3. Variables used in the investigation (Mean: average score per variable; N2: number of tokens per variable used in validity criterion; N1: total number of tokens per variable in the subset).

Variable	Mean	N2	N1
/x/	46.9	16	21
/z/	70.1	13	14
/v/	41.4	12	14
schwa-insertion	67.8	12	13
assimilation	43.9	10	14
n-deletion	90.3	12	14

Alternatively, we can calculate the performance of the separate pairs of untrained transcribers for each variable, again using the identical transcriptions of the trained transcribers as our criterion. The results are presented in Table 4. It appears that average performance scores vary between 53 and 69%.

Table 4. Average scores per pair over all the tokens used as part of the validity criterion (For reasons of space, numbers are rounded off to the nearest integer; P: pair; V: variable; s'a: schwa-insertion; ass: assimilation; n-del: n-deletion).

P\V	/x/	/z/	/v/	s'a	ass	n-del
1	41	62	50	63	20	83
2	28	81	42	42	75	83
3	28	73	17	88	70	92
4	50	62	38	75	60	100
5	44	81	38	83	25	63
6	50	58	58	88	45	100
7	84	73	46	79	30	100
8	53	89	42	25	40	100
9	44	58	42	67	30	92

11 CONCLUSION

The results of the study lend support to our hypothesis that trained transcribers reach higher consistency levels in replicated consensus transcriptions than untrained transcribers.

It also appears that, while agreement between consensus transcriptions is not a good validity criterion per se, high agreement between non-contemporary consensus transcriptions made by trained transcribers can be used as a measure of transcription validity.

REFERENCES

[1] Cucchiarini, C. (1993) *Phonetic Transcription: A Methodological and Empirical Study*, Nijmegen.

[2] Shriberg, L.D. et al. (1984) 'A procedure for Phonetic transcription by Consensus: A Research Note', *Journal of Speech and Hearing Research* 27, 456-465.

[3] Vieregge, W.H. and Broeders, A.P.A. (1993) 'Intra- and Interspeaker Variation of /r/ in Dutch', in: *Proceedings of Eurospeech 93*, 267-270.

[4] Ting, A. et al. (1970) 'Phonetic Transcription: A Study of Transcriber Variation', *Report*, Madison: Wisconsin University.

[5] Broeders, A.P.A. and Vieregge, W.H. (1991) 'Intraspeaker Variation on the Segmental Level: a Transcription-based Approach', in: *Proceedings of the XIIth International Congress of Phonetic Sciences*, Aix-en-Provence: Université de Provence, Vol 5: 46-49.

[6] (1993) 'The International Phonetic Alphabet' *Journal of the International Phonetic Association* 23(1), center pages.

[7] Vieregge, W.H. (1987) 'Basic aspects of Phonetic Segmental Transcription', in: Almeida, A. and Braun, A. (eds.) *Probleme der phonetischen Transkription*', Stuttgart: Franz Steiner Verlag.

MEASURES OF THE GLOTTAL AIRFLOW WAVEFORM, EGG, AND ACOUSTIC SPECTRAL SLOPE FOR FEMALE VOICE

Eva B. Holmberg, Robert E. Hillman, Joseph S. Perkell
Massachusetts Eye and Ear Infirmary and Massachusetts Institute of Technology

ABSTRACT

Comparisons were made among aerodynamic, electroglottographic, and acoustic spectral measures for syllable production and sustained vowel phonation in comfortable and loud voice of 20 women with normal voices. Measures differed significantly between tokens having harmonic energy versus noise in the F3 region. Spectral measures added useful information to glottal waveform data about abrupt versus gradual vocal fold closing.

INTRODUCTION

The objective of this study was to examine whether acoustic spectral measures of voice production could be used to supplement measures obtained from the inverse filtered oral flow waveform that are especially sensitive to technical difficulties [1]. We focused on measures that have been found salient for vocal intensity [2] and glottal aperture [3]. Our flow-based inverse filtering technique is not without problems: 1) An accidental air leak between the subject's face and the transducer mask lowers amplitude-based flow measures that include the DC flow. 2) Unsuccessful inverse filtering that results in formant residuals superimposed on the glottal waveform can make time-based measures unreliable. 3) Low-pass filtering, used in our inverse filtering algorithm, has the undesirable effect of rounding of waveform discontinuities, such as at the instant of vocal fold closure.

Measures made from the acoustic spectrum may also assist in an objective evaluation of voice quality, such as degree of perceived breathiness. A breathy voice is the result of incomplete vocal fold closure and increased subglottal coupling [4]. Thus, this study also examines relationships between measures of the acoustic spectral slope and glottal waveform measures that are believed to correlate with increased subglottal coupling [3]. In addition, qualitative observations were made of the energy content (noise versus harmonic energy) in the third formant frequency region [5].

A measure of an "adduction quotient" on the electroglottographic (EGG) signal [6] was also included, in order to determine whether any useful information could be obtained from the EGG-based quotient that was not available from the analogous flow-based quotient.

Finally, two different elicitation materials were used: strings of repeated /pæ/ syllables and sustained phonation of /æ/ vowels. Our intention is to combine measures from the two speech tasks.

METHODS

Detailed descriptions of recording procedures, signal processing, data extraction and analyses procedures are presented in previous publications [7, 8]. In brief: twenty American females, age 20 to 43 years, with healthy voices served as subjects. They produced two different speech tasks in comfortable and loud voice: (1) strings of five repetitions of the syllable /pæ/, and (2) the vowel /æ/, sustained for 2-3 seconds. Recordings were made of oral airflow with a "Rothenberg mask" (Glottal Enterprises); intraoral air pressure with a thin catheter connected to a differential pressure transducer (Glottal Enterprises); sound pressure, with a small microphone (Sony model ECM 50) attached at a fixed, reproducible distance of 15 cm from the subject's lips; and EGG, using a laryngograph (Glottal Enterprises). The flow signal was low-pass filtered at 1100 Hz and inverse filtered to remove effects of the first formant. The EGG signal was low-pass filtered at 1710 Hz. Appropriate calibration signals were recorded for air pressure, airflow and intensity. The recorded signals were sampled at different rates, digitized simultaneously, demultiplexed and processed further in software. Measures were extracted algorithmically, with interactive monitoring, at a vowel midpoint location.

Estimates were made of *average transglottal air pressure*, (the driving force for phonation, cm H_2O). *SPL* was calculated from the RMS of the speech signal. *Glottal airflow waveform measurements* were made of: *DC flow* (the unmodulated flow, l/sec), and *flow-adduction quotient* (closed time/T) using a 30% amplitude criterion level [9]. *EGG-adduction quotient* (vocal fold contact time/T), measured at an (arbitrary) 65% criterion amplitude level. *Amplitude differences (dB) were calculated from the acoustic spectra* between: the first two harmonics (AH1-AH2); the first harmonic and the peak harmonic in the first formant (AH1-AF1); the first harmonic and the spectral peak of the third formant (AH1-AF3); the peak harmonic of the first formant and the spectral peak of the third formant (AF1-AF3). *Qualitative observations were made of the energy content in the frequency region of F3*, whether the spectrum consisted predominantly of harmonics, noise, or a mixture of harmonics and noise. *Statistical analyses* were performed to examine: differences between /pæ/ and /æ/; pairwise linear relationships between parameters; and the extent to which the parameters differed between tokens with F3 harmonic energy and those with F3 noise.

RESULTS

SPL was higher for the vowel in the syllable strings than in the sustained phonation. Analysis of covariance (ANACOVA, $p<0.05$, Bonferroni corrections, $p=0.0045$) with SPL as the covariate showed that there were no significant differences in other parameters between /pæ/ and /æ/ productions after adjustment for SPL.

Relationships between Flow- and EGG-Adduction Quotients.

Pearson product moment correlations calculated between the flow- and EGG adduction quotients, showed strong relationships ($r>0.70$) for individual speakers, whose signals were strong and noise free. The results suggest that the quotients measured at the amplitude levels of 30% (flow) and 65% (EGG) were highly related.

Relationships between Glottal Airflow Measures and the Spectral Slope.

Pearson product moment correlations between measurements of glottal airflow waveforms and spectral slope showed a relatively strong relationship between flow-adduction quotient and AH1-AH2 ($r=-0.69$). The results suggest that the degree to which the glottal waveform has a sinusoidal shape, and inversely, the degree of glottal adduction, was reflected relatively well in AH1-AH2. AH1-AH2 was also relatively strongly correlated with SPL ($r=0.69$). The relationship between flow-adduction quotient and AF1-AF3 was significant for tokens with predominantly F3 noise, but non-significant for tokens with predominantly F3 harmonic energy. Other relationships were weak in the group data. The relationships between acoustic spectral measures and glottal waveform measures were examined also for each individual speaker. A majority of the individual speakers displayed strong relationships ($r>0.70$) between flow-adduction quotient and all the spectral measures, with the exception of AF1-AF3. The data suggest that the degree to which the glottal waveform had a sinusoidal shape was reflected in ratios that included the amplitude of the fundamental, but not in the ratio which included only the higher frequency region.

Relationships between F3 Spectral Energy Content and Loudness Condition.

Simple tallies were made of the number of tokens with F3 harmonic energy, tokens with F3 noise, and tokens with mixed noise and harmonic F3 energy, in comfortable and loud voice respectively. Most tokens (122 of 240) in comfortable voice displayed a mix of harmonic energy and noise in the F3 region, followed by tokens with predominantly F3 noise (84). Few tokens (34) displayed predominantly F3 harmonic energy in comfortable voice. In contrast, in loud voice, most tokens (144 of 240) displayed harmonic F3 energy, followed by tokens with a mix of F3 harmonic energy and noise (68). Few tokens (28) in loud voice displayed F3 energy with predominantly noise.

Differences in Acoustic and Glottal Waveform Measures between Tokens with F3 Spectral Energy and F3 Noise.

Analyses of variance ($p<0.001$) showed that tokens with predominantly noise in the F3 region were associated with significantly: lower SPL; larger values of AH1-AH2, AH1-AF3, and AF1-AF3; smaller values of AH1-AF1; lower subglottal air pressure; smaller adduction quotients (flow and EGG); and higher DC flow.

DISCUSSION

The finding of higher SPL for /æ/ in the /pæ/ syllable strings than for /æ/ in sustained phonation was most likely due to differences in location of the mid-vowel data extraction; in the syllables the mid-vowel point occurs shortly after initiation of vocalization, while in sustained phonation mid-vowel occurs well into the vowel, at a point where SPL was stabilized and somewhat lower than at the beginning of the sustained vowel.

The literature has suggested that the EGG waveform contained a number of interesting features and events that could be useful for a better understanding of the underlying vocal-fold vibration pattern [6]. However, we seldom find such clear events in the EGG waveforms. In addition, we have experienced difficulties in recording EGG, for example weak and noisy signals for women, and intermittent disruptions of the signal due to gross movements of larynx that accompany changes in vocal effort. However, for speakers with strong EGG signals and clean glottal waveforms (without formant residuals), the flow and EGG quotients, measured at 30% and 65% amplitude criteria levels respectively, were highly correlated ($r>0.85$). This finding suggests that quotients from clean samples from one signal can complement the other if necessary.

A particular focus of this study was relationships between flow and acoustic parameters that have been found reflective of glottal aperture, with the goal of cross validating these measures. Gradual closing movements of a somewhat abducted membranous portion of the vocal folds should result in relatively sinusoidal glottal waveforms and small adduction quotients [3]. The result of relatively strong (negative) correlation between flow-adduction quotient and AH1-AH2 suggests that measurements of adduction quotient at the 30% level criterion was sensitive enough to differentiate among waveforms with gradual and abrupt closing, and that AH1-AH2 could be used as a substitute for adduction quotient in case of unsuccessful inverse filtering. The strong (negative) relationships between flow-adduction quotient on the one hand and AH1-AF1, AH1-AF3, and SPL on the other for the individual speakers suggest that gradual vocal fold closures resulted in an increased amplitude of the first harmonic, reduced amplitude of the first formant, steeper overall spectral slope, and reduced SPL, in agreement with previous research [2].

Ideally, high vocal fold closing velocities and abrupt reduction of the airflow should result in glottal waveforms with sharp corners between the closing and closed portions. However, a detrimental effect of low-pass filtering at 1100 Hz is a "rounding" of waveform discontinuities, which could have an influence on waveforms associated with high vocal fold closing velocities. Waveforms which result from more gradual closing movements (therefore already rounded) would be relatively uninfluenced by the low-pass filtering rounding effect. These waveforms, with F3 excited by noise, have a significant relationship between flow-adduction quotient and AF1-AF3. In contrast, waveforms with F3 excited by harmonic energy have a non-significant relationship between flow-adduction quotient and AF1-AF3. The filter-induced rounding of waveforms with more abrupt closures may account for the lack of significant correlation between adduction quotient and AF1-AF3. In other words, the sharp corners in those waveforms were obscured, and reliable adduction quotient measurements were precluded by the low-pass filtering at 1100 Hz. However, the effects of the sharp discontinuities were preserved in the AF1-AF3 value derived from the full-bandwidth acoustic spectra. These results suggest that the spectral measurement of AF1-AF3 may serve as a useful complement to the flow-based adduction quotient, especially when there are high vocal fold closing velocities [10].

In normal phonation, DC flow in the glottal waveform has been assumed to reflect airflow that passes through a posterior glottal "chink". A large chink would increase the subglottal coupling with reduced high frequency energy and reduced SPL. However, neither the relationship between underlying physiology and the DC flow [11], nor the acoustic effect of the DC flow is completely understood. None of the acoustic measures varied systematically with DC flow. Thus, DC flow data must be interpreted with caution.

CONCLUSIONS

The following conclusions could be drawn from the results of this study:

1) Comparisons between measures obtained from the vowel in /pæ/ syllables and those obtained from sustained /æ/ phonation can be made, as long as SPL differences are controlled for.

2) Adduction quotient, measured at a 30% amplitude level on the glottal waveform is sensitive enough to differentiate among waveforms with gradual and abrupt closing portions in data for individual subjects.

3) Measurements of the amplitude difference between the two first harmonics (AH1-AH2) may be used as a substitute measure for flow-adduction quotient, in cases of unsuccessful inverse filtering that make measurements of adduction quotient unreliable.

4) The flow- and EGG adduction quotients, measured at 30% and 65% levels respectively, may serve to complement one another.

5) AF1-AF3 may serve as a useful complement to measurements of maximum flow declination rate, especially in voices with very high closing vocal fold velocities that cannot be reflected accurately in a flow waveform that is low-pass filtered at 1100 Hz.

ACKNOWLEDGMENTS

This study was supported by a grant from The National Institutes of Health.

REFERENCES

[1] Holmberg, E.B., Hillman, R.E., Perkell, J.S., Guiod, P.C., and Goldman, S.L. (in press), "Comparisons among Aerodynamic, Electroglottographic, and Acoustic Spectral Measures of Female Voice", J. Speech and Hear. Res.

[2] Fant, G. (1979), "Glottal source and excitation analysis", Quarterly Report, 1, Speech Transmission Laboratory and Music Acoustics, The Royal Institute of Technology, Stockholm, pp. 85-107.

[3] Stevens, K.N. (1977), "Physics of laryngeal behavior and larynx modes", Phonetica, 34, pp. 264-279.

[4] Södersten, M., & Lindestad, P-A. (1990), "Glottal closure and perceived breathiness during phonation in normally speaking subjects", J. Speech and Hear. Res., 33, pp. 601-611.

[5] Klatt, D.H., & 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] Childers, D.G., Hicks, D.M., Moore, G.P., Eskenazi, L., & Lalwani, A.L. (1990) "Electroglottography and vocal fold physiology", J. Speech and Hear. Res., 33, pp. 245-254.

[7] 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. 84, pp. 511-529.

[8] Perkell, J.S., Holmberg, E.B., & Hillman, R.E. (1991), "A system for signal processing and data extraction from aerodynamic, acoustic, and electroglottographic signals in the study of voice production", J. Acoust. Soc. Am., 54, pp. 1777-1781.

[9] Colton, R.H., Brewer, D.W., & Rothenberg, M. (1985), "Vibratory characteristics of patients with voice disorders", Oral presentation, Symposium and voice Acoustics and Dysphonia, Gotland, Sweden, 1985.

[10] Hillman, R.E., Holmberg, E.B., Perkell, J.S., Walsh, M, & Vaughan, C. (1989), Objective assessment of vocal hyperfunction." J. Speech and Hear. Res., 32, pp. 373-392.

[11] Hertegård, S., Gauffin, J. (1995), "Glottal area and vibratory patterns studied with simultaneous stroboscopy, flow glottography, and electroglottography", J. Speech and Hear. Res., 38, pp. 85-100.

SUB-GLOTTAL RESONANCES IN FEMALE SPEAKERS AND THEIR EFFECT ON VOWEL SPECTRA

Helen M. Hanson (1) and Kenneth N. Stevens (2)
(1) Div. of Applied Sciences, Harvard Univ., Cambridge MA 02138, USA
(2) Research Lab of Electronics and Dept. of Electrical Engineering and Computer Science, MIT, Cambridge MA 02139, USA

ABSTRACT

Resonances of the subglottal system often influence the acoustic characteristics of vowels. These influences are manifested as extra peaks in vowel spectra and as discontinuities in apparent formant movements. Data from vowels produced by a number of female speakers show that the magnitudes of these effects are correlated with acoustic measures indicating the degree of glottal abduction used by the speakers during phonation.

1. INTRODUCTION

We often observe prominences in the spectra of vowels that are not attributable to formants that are natural frequencies of the vocal tract [1]. An example of a vowel spectrum with these extra prominences is given in Fig. 1. In this example we see an extra spectral peak at about 1600 Hz. Adjacent to this spectral peak is a valley indicating the presence of a zero or antiresonance in the vocal-tract transfer function. These irregularities in the spectrum result from acoustic coupling, through the glottis, between the vocal tract and the trachea.

We report here some data on the prominences arising from the tracheal resonances for a number of female speakers, and we relate these data to other spectral measurements reported previously for these speakers [2][3]. We turn first to some theoretical background.

2. THEORY

The subglottal and supraglottal systems are coupled through a narrow glottal opening which has a resistance R_g and an acoustic mass M_g. As a first

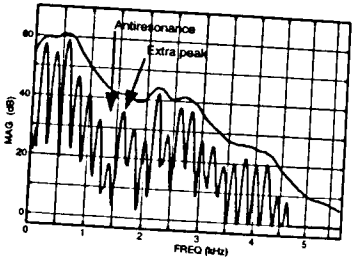


Figure 1: *Spectrum of vowel /ɛ/ produced by a female speaker, showing extra peak and antiresonance due to acoustic coupling to the trachea. Time window of spectrum is 22.3 ms.*

approximation we represent the acoustic source as paired volume-velocity sources U_s as shown in Fig. 2. Z_t and Z_v are the impedances looking into the trachea and vocal tract.

The transfer function U_m/U_s is characterized by poles, which are the natural frequencies of the coupled system, together with zeros at frequencies for which $Z_t = \infty$. These zeros are the natural frequencies of the subglottal system when the glottis is closed.

Measurements of the subglottal resonant frequencies have been reported by [4][5] and others. Typical values of the lowest three of these frequencies

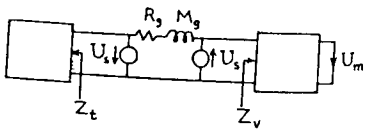


Figure 2: *Equivalent circuit showing vocal tract and subglottal system connected to sources and coupled through the glottis. See text.*

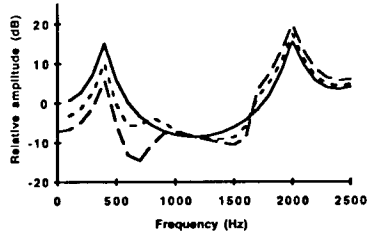


Figure 3: *Calculated transfer function U_m/U_s of the vocal tract when there is no acoustic coupling to the trachea (solid line), when $A_g \approx 0.02$ cm² (dotted line), and when $A_g \approx 0.05$ cm² (dashed line). The tracheal resonances are at 650 and 1600 Hz.*

with a closed glottis are estimated to be 700, 1700, and 2300 Hz for adult female speakers (cf. [5]). The bandwidths of these resonances are about 200 Hz [4]. The poles f_p of the transfer function due to the subglottal cavity are close to the zeros f_z noted above, and the amount of separation between a pole and a zero in a pair depends on the size of the glottal opening, i.e., the values of R_g and M_g. For different speakers, the average glottal area, and hence the average values of R_g and M_g, may be different. For some speakers, this difference is due to the fact that the glottis does not completely close during the so-called closed phase of vibration.

In order to estimate the effect of the glottal opening on the spectrum of the radiated sound, we can calculate the transfer function U_m/U_s for various values of the average glottal area A_g. We assume that the impedance looking into the vocal tract is small compared with the impedance of the glottis. This assumption is reasonable as long as the subglottal resonance is not too close to a natural frequency of the vocal tract (a formant). The frequency of the pole is estimated to be the natural frequency of the subglottal system when it is terminated by the glottal impedance.

Calculations of the vocal-tract transfer function for a typical front-vowel configuration with formants well separated from the tracheal resonances are shown in Fig. 3 for two different glottal areas. When the glottal area

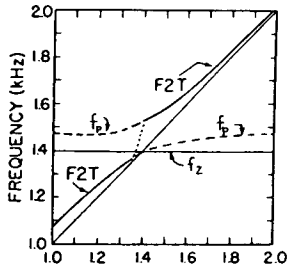


Figure 4: *Estimates of the frequencies of the two poles and zero in the vicinity of F2 when there is coupling to the trachea through a partially open glottis. The zero f_z is assumed to be fixed at 1400 Hz. F2T represents the pole corresponding to F2, shifted by the influence of the tracheal system. The solid lines indicate the most prominent spectral peak, which shows an abrupt jump in frequency (dotted line) when F2 is just below f_z. The dashed lines represent f_p which is less prominent in the spectrum.*

is larger, substantial additional prominences appear in the vowel spectrum, whereas for the smaller area the effect of the tracheal resonances is small.

If the frequency of a formant is close to a subglottal resonance, the subglottal coupling will have an influence on the spectral representation of the formant. For example, if a formant passes through the region of a subglottal resonance, interference is expected. This interference effect is illustrated in Fig. 4. The abscissa is the frequency $F2$ that would exist if the glottis were closed, and the ordinate is the actual frequencies of the poles and the zero for the coupled system. The second formant frequency $F2$ increases from 1100 to 1800 Hz, passing through the tracheal resonance f_z, which in this example is fixed at 1400 Hz. When $F2$ is well separated from f_z, there is a small upward shift in the pole representing $F2$, and there is a pole-zero pair f_z and f_p due to tracheal coupling. When $F2$ approaches f_z, the pole-zero-pole combination creates two nearby spectral peaks. When $F2 < f_z$, the lower of these peaks is dominant, but when

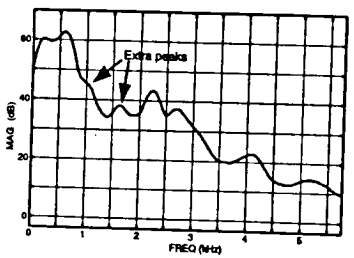


Figure 5: *Average spectrum of vowel /ɛ/ produced by the speaker represented in Fig. 1. The average was obtained over five repetitions of the vowel (in the word* bɛd*). Evidence for extra peaks due to tracheal coupling is shown.*

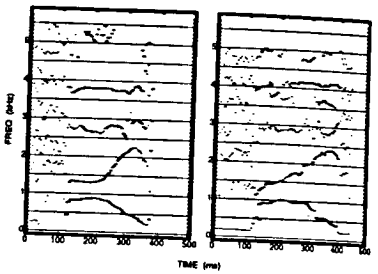


Figure 6: *Examples of trajectories produced by an LPC-based formant tracker for the word* bide *produced by two female speakers. On the left, the F1 and F2 tracks are smooth, with no discontinuities, but on the right, there are discontinuous jumps as the formants pass through tracheal resonances.*

$F2$ passes upward through f_s the upper peak becomes dominant. Thus there is a discontinuous upward jump in the dominant spectral peak as $F2$ increases through the subglottal resonance. A similar effect occurs when $F1$ passes through the lowest tracheal resonance.

This theoretical analysis suggests, then, that there are two kinds of acoustic evidence for acoustic coupling to tracheal resonances: one is the presence of spectral prominences in addition to the prominences due to vocal-tract resonances or formants, and the other is the disruption of prominences due to formants as they pass through frequency ranges of tracheal resonances. The latter effect should be observable in diphthongs like /ai/, where $F1$ traverses downward through the lowest tracheal resonance and $F2$ follows an upward-moving trajectory through the second tracheal resonance.

Theory also predicts that tracheal resonances should be more evident in vowel spectra for individuals who phonate with a glottis that remains partially open throughout a glottal cycle. Such individuals are known to exhibit a greater high-frequency tilt in the glottal spectrum and a greater $F1$ bandwidth due to increased acoustic losses at the partially open glottis.

3. EXPERIMENTAL DATA

Two kinds of acoustic data were obtained from vowels produced by 22 speakers. From spectra of the vowels /ɛ æ ʌ/ in CVC words, estimates were made of the degree of perturbation by extra peaks and valleys that could be ascribed to tracheal resonances. The spectrum for each vowel was an average spectrum over the vowel portions of five repetitions of the words, using a short time window (7 ms) calculated every millisecond. An average spectrum of the vowel /ɛ/ for the speaker of Fig. 1 is shown in Fig. 5. It was thought that such an average spectrum should be effective for showing prominences (such as those due to tracheal resonances) that remain relatively fixed in frequency over time. The deviation of each vowel spectrum in terms of extra prominences was rated by two observers on a scale from 0 to 2.

A second type of acoustic data examined the tracking of the first and second formant peaks in the diphthong /ai/ in the utterance bide. Formant tracks obtained using a standard LPC algorithm are shown in Fig. 6 for two speakers. For one of the speakers, the formants are tracked smoothly, except for minor ripples due to the interaction of the fundamental frequency and the formants. For the other speaker, there is an abrupt discontinuity in both formant tracks, presumably due to the influence of tracheal resonances.

The $F1$ and $F2$ tracks for this diphthong produced five times by each speaker were examined, and cases with a significant discontinuity in either track were noted. To qualify as a discontinuity induced by a subglottal resonance, it must occur in the frequency range 500-1000 Hz for $F1$ and 1500-2000 Hz for $F2$. Each speaker was rated by the number of such discontinuities, ranging from 0 to 10.

These two measures — spectral deviations caused by extra prominences (EP), and discontinuities in formant tracks for a diphthong (DF) — were examined in relation to other acoustic measures. These other measures are theoretically related to the size of a fixed opening in the glottis during the "closed" phase of the glottal vibration cycle, and should increase as the cross-sectional area of the opening does [2][3]. The measures are: (1) $H1$-$A1$, the difference (in dB) between the amplitude of the first harmonic and the amplitude of the largest harmonic in the vicinity of the first formant. This difference is related to the bandwidth of the first formant. (2) $H1$-$A3$, where $A3$ is the amplitude of the third formant peak. This is a measure of spectral tilt. (3) The bandwidth $B1$ of the first formant, as determined by the rate of decay of the $F1$ waveform during the initial (most closed) part of the glottal cycle. (4) Estimates N_w and N_s of noise excitation in the $F3$ waveform and high-frequency spectrum, respectively [5].

The correlations between EP, DF and the spectral measures are summarized in Table 1. The correlations between DF and the spectral measures are all quite high, particularly DF and spectral tilt. The correlations for EP are smaller, possibly due to the subjectivity of this measure. It is clear from these correlations that when spectral measures indicate a significant glottal opening or "chink," evidence for tracheal resonances appear in the vowel spectrum. The effect of the tracheal resonances on the spectrum increases as the size of the opening increases.

Table 1: *Correlations between two measures of the prominences of tracheal resonances and several spectral measures (see text) obtained from vowels produced by 22 female talkers.*

Measure	*EP*	*DF*
H1-A1	0.50	0.68
H1-A3	0.70	0.83
N_w	0.68	0.82
N_s	0.57	0.79
DF	0.62	1

4. CONCLUSION

Tracheal resonances can introduce significant modifications in the vowel spectra for some speakers. These are speakers for whom other spectral measures such as spectral tilt indicate some glottal abduction during the "closed phase" of glottal vibration. Tracheal resonances can interfere with the estimation of formants from vowel spectra and thus have implications for formant tracking and speech recognition systems. The effects of these resonances on both formant location and prominences can also influence vowel space and quality. Finally, our observations of these effects suggest that the simple source-filter theory may not always be adequate, even for modal phonation.

ACKNOWLEDGEMENTS

This work was supported in part by NIH Grant DC00075.

REFERENCES

[1] Fant, G. (1972), "Subglottal formants," *STL-QPSR 1*, Royal Institute of Technology, Stockholm, 1-12.

[2] Stevens, K.N. and H.M. Hanson (1995), "Classification of glottal vibration from acoustic measurements," in O. Fujimura and M. Hirano (eds.) *Vocal Fold Physiology: Voice Quality Control*, San Diego: Singular, 147-170.

[3] Hanson, H.M. (1995), "Glottal characteristics of female speakers," Ph.D. Thesis, Harvard Univ., Cambridge MA.

[4] Ishizaka, K., K.M. Matsudaira, and T. Kaneko (1976), "Input acoustic-impedance measurement of the subglottal system," *J. Acoust. Soc. Am.*, Vol. 60, 190-197.

[5] 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.*, Vol. 87, 820-857.

EXCITATION-SYNCHRONOUS GLOTTIS INVERSE-FILTERING BY MEANS OF A SELF-EXCITED THRESHOLD AUTO-REGRESSIVE MODEL OF THE SPEECH SIGNAL

*J. Schoentgen * and Z. Azami †*
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

We propose to carry out excitation-synchronous glottis inverse filtering by means of a compound auto-regressive model of the speech signal. The model consists of two linear auto-regressive models that are excitation-synchronously fitted by means of an auxiliary signal which has the same period as the speech signal and a single peak per cycle. Auxiliary and speech signals need not be aligned.

INTRODUCTION

Glottal inverse filtering is the estimation of the glottal waveform from the speech signal. Generally speaking, glottal inverse filtering consists of filtering the speech signal by means of an "inverted" transfer function estimate in which poles are replaced by zeros. Difficulties with glottal inverse filtering are the following. a) The vocal tract transfer function must be estimated from the speech signal. But, the speech signal is not the impulse response of the vocal tract. It is, in a first approximation, the convolution of the impulse response and the glottis signal which includes effects of the interaction between voice source and vocal tract. b) During connected speech, the transfer function varies with time whereas conventional estimation techniques (e.g. linear predictive analysis) posit that the speech signal is stationary during the analysis interval. c) Even during the emission of sustained vowels, the vocal tract transfer function cannot be assumed to be stationary because the vibrating vocal folds rhythmically connect and disconnect subglottal and tract cavities. As a result, both eigenfrequencies and bandwidths change within a glottal cycle. As a consequence, glottal signals are difficult to estimate reliably and, more often than not, attempts at inverse filtering are confined to sustained vowels.

Here, we propose to carry out inverse filtering in the following way. First, closed phases of the glottal cycle are detected via a compound speech signal model that consists of two linear auto-regressive models. Fitting of the model is carried out by means of the overall prediction error and the energy difference between the "open" and "closed" phase components of the speech signal. Second, formant frequencies and bandwidths of the "closed" phase components are estimated. Indeed, a conventional solution of problem (c) is to estimate the tract transfer function throughout the closed phases of the glottal cycles [1]. A consequence is that effects of the interaction between vocal tract and glottal source are included in the voice source signal [2]. Third, after eliminating real poles and complex poles whose bandwidths are larger than 500 Hz, inversed filtering is carried out by means of a cascade of second-order cells, that is one cell per pair of complex conjugate poles.

MODEL

The compound model of the speech signal we have proposed earlier is the following [3] [4] [5] [6].

$$y(n) = a_0 + \sum_{i=1}^{N} a_i y(n-i), \quad w(n-d) < r \tag{1}$$

$$y(n) = b_0 + \sum_{i=1}^{M} b_i y(n-i), \quad w(n-d) \geq r \tag{2}$$

Signal $y(n)$ is represented by means of two linear auto-regressive models (1) & (2). n is the time index, r a threshold and d a delay. a_i and b_i are the coefficients of linear sub-models (1) & (2) and N and M their orders. $w(n)$ is an auxiliary signal that must have a single vertex per cycle and the same period as signal $y(n)$. However, signals $w(n)$ and $y(n)$ need not be aligned.

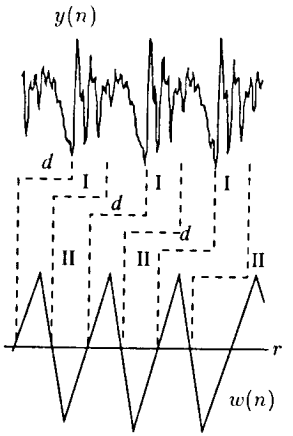


Figure 1: Speech signal $y(n)$ and synthetic triangular auxiliary signal $w(n)$.

METHOD

In the framework of glottal inverse filtering, we made use of two synthetic auxiliary signals. The first was a rectangular and the second an isosceles triangular waveform. The difference between rectangular and triangular auxiliary signals was that for the latter the length of cutouts (I) could be varied between 0 (the summit of the triangle) and period T (its base) by means of threshold r. The function of delay d was to position these lengths, cut out by means of threshold r, with respect to the time coordinates of the intersections of threshold r with signal $w(n)$ (Fig.1). Delay d and threshold r were either determined by means of an optimizer or a systematic search. When auxiliary signal $w(n)$ was a rectangular waveform, the cutout lengths were constants equal to the crenellation width and the only variable whose optimal value had to be searched for was delay d. The crenellation width of rectangular auxiliary signal $w(n)$ was fixed at 40 % of glottis cycle length T. Therefore, the lengths of cutouts (I & II) were respectively equal to $w(n)$'s crenellation width and the remaining 60 % of the fundamental period. Delay d was varied between 0 and T. For each choice of delay d, sub-models (1) & (2) were fitted to their respective cutouts (I) & (II) by means of singular value decomposition and the normalized overall prediction error was calculated. Sub-model orders were respectively equal to 9 and 8 and the sampling frequency was equal to 8 kHz. The best break-up of the speech signal into "open" and "closed" phase cutouts was the one that gave rise to a local minimum of the prediction error and to the biggest difference between the energies of signal components (I) & (II). In other words, the "closed" phase components of the speech signal were assigned to those cutouts that gave rise to a local minimum of the overall prediction error and to a maximum of the signal energy. For this choice of d, formant frequencies and bandwidths of the "high-energy" cutouts were determined and inverse filtering was carried out by means of a cascade of second-order cells.

The stages of the segmenting, fitting and inverse filtering algorithm were as follows :

(i) Asynchronous positioning of an analysis window of a length of 26 msec ;
(ii) Estimation of glottis cycle length T by means of the smoothed speech signal ;
(iii) Initialization of delay d, d=0 ;
(iv) Segmentation of the windowed speech signal (cf. Figure 1) by means of a rectangular auxiliary signal ;
(v) Least mean square fitting of sub-model (1) to cutouts (I) and of sub-model (2) to cutouts (II) ;
(vi) Calculation of the total normalized prediction error. Normalization was by the cutout lengths ;
(vii) Incrementation of d. If $d <$ T then step (iv) otherwise step (viii) ;
(viii) Calculation of the energy differences between components (I) and (II) for delays d that gave rise to the five smallest prediction errors ;
(ix) Selection of the segmentation (on the base of delay d) that gave rise to a maximum energy difference during step (viii) ;
(x) Reestimation by means of a covariance multi-interval method of linear predictive coefficients a_i of components (I) arrived at step (ix) ;
(xi) Computation of formant frequencies and bandwidths by means of the predictive coefficient polynomial ;
(xii) Discarding of real poles and of complex pole pairs giving rise to bandwidths larger than 500 Hz ;
(xiii) Inverse filtering by means of a cascade of second-order cells, i.e. a cell per pair of complex conjugate poles remaining after step (xii) ;
(xiv) Segmentation by means of period T and delay d of the glottal waveform so arrived at ;

RESULTS AND DISCUSSION

Figures 2 and 3 show glottal waveforms, of sustained vowels or vowel transitions, obtained by means of the inverse filtering method previously explained. The displayed waveforms of four speakers were arrived at wholly automatically. It is seen that they have traits that are typical of waveforms that have been obtained in the framework of other studies. Part of the observed inter- and intra-speaker variability is generally believed to be a consequence of the fickle ability of the linear predictive model to represent the speech signal adequately. But, variability may also have been a consequence of the occasional inability of steps (iii) to (viii) to segment identically from one analysis window to the next. Also, we have tried out, on the same speech signals, auxiliary signals of rectangular and triangular shape. The ability of the triangular waveform to give rise to variable cutout lengths did however not appear to be an advantage over the rectangular waveform whose cutout lengths were fixed.

It is planned to post-process waveforms so as to get rid of pulse estimates that are outliers and handle intra-speaker variability via vector quantization which chooses a representative set of glottal pulses. Indeed, here the purpose of inverse filtering is not to provide entire glottal waveforms. Instead, the objective is to arrive at a set of speaker-typical glottal pulses so that the discrimination performance of acoustic features, related either to the glottal pulse or the vocal tract transfer function, can be compared in the framework of a speaker recognition task. The purpose is to find an answer to the question of whether speaker identity is made up of acoustic cues that bear on the voice source, or the vocal tract or a combination of the two.

References

[1] A. K. Krishnamurthy and D. G. Childers. Two-channel speech analysis. *IEEE Trans. Acoust., Speech and Signal Processing*, ASSP-34(4):730-743, 1986.

[2] T.V. Ananthapadmanabha and G. Fant. Calculation of true glottal flow and its components. *Speech Communication*, 1:167-184, 1982.

[3] J. Schoentgen and Z. Azami. Pitch-synchronous formant extraction by means of a compound auto-regressive model. *Proceedings Eurospeech'93*, pages 401-404, 1993.

[4] Z. Azami and J. Schoentgen. Extraction des formants en synchronie avec l'excitation à l'aide d'un modèle auto-régressif composé. *Proceedings XXèmes Journées d'Etude sur la Parole*, pages 235-240, Trégastel, 1994.

[5] J. Schoentgen. Self-excited threshold auto-regressive models of the glottal pulse and the speech signal. *Proceedings of the ICSLP*, pages 1063-1066, 1994.

[6] J. Schoentgen. Dynamic models of the glottal pulse. In *Levels in Speech Communication, Relations and interactions, Sorin et al. (Eds), Elsevier, Amsterdam*, pages 249-266, 1995.

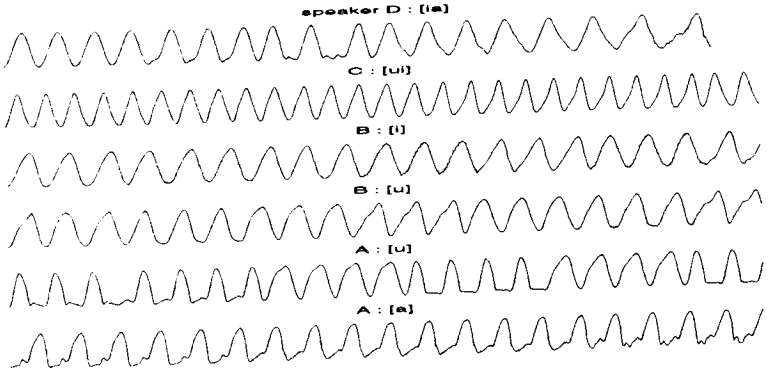


Figure 2:

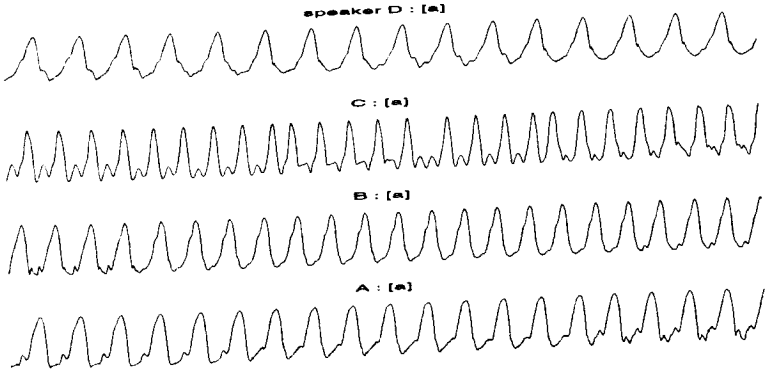


Figure 3:

STABILITY AND BIFURCATIONS OF THE TWO-MASS MODEL OSCILLATION: ANALYSIS OF FLUID MECHANICS EFFECTS AND ACOUSTICAL LOADING

R. Laboissière and X. Pelorson

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

E-mail: **rafael@icp.grenet.fr**

ABSTRACT

We present some extensions to the results found by Lucero (1993) concerning the analysis of the large-amplitude oscillation of the vocal folds using the two-mass model. We focus on two points which were not considered in that work: the introduction of a more realistic model of the fluid mechanics aspects of the glottal flow, and the effects of the acoustical loading of the vocal tract. A numerical technique is presented for finding the equilibrium points and analysing their stability for generic aerodynamic and mechanical models of the vocal folds, including as well a representation of the acoustical impedance of the vocal tract. Our results confirm the interest of an analysis of stability of equilibrium points to obtain the oscillation regions of the vocal-folds, but also indicates to the need of better aerodynamic and acoustical models.

INTRODUCTION

Over the years, several researchers have been trying to quantify vocal fold vibration. One of the main question one is interested in answering is: Given a mechanical, aerodynamical, and acoustical model of the vocal folds and the vocal tract, under which conditions of the control parameters (e.g. lung pressure and stiffness of the laryngeal muscles) will the vocal folds oscillate? As even the simplest models of the vocal tract (e.g. Ishizaka and Flanagan 1972) are described by non-linear differential equations for both the mechanical and the aerodynamical parts, direct analytical analysis are difficult to be carried out. The difficulties are expected to increase as more realistic models of the larynx will be developed. This are the main reasons according to which previous works have been focused on small-amplitude analysis of vocal fold vibration (Titze 1988). The drawback of this kind of technique is the linearization of the equations of motion, making the conclusions hardly extensible to the large-amplitude oscillation behaviour.

More recently, some non-linear techniques have been applied to the study of vocal fold vibration. They range over a wide variety of mathematical tools. Awrejcewicz (1990) uses characteristic multipliers to change 'bifurcation' parameters in order to discover new periodic solutions via Hopf bifurcation. Weakly nonlinear analyses are done by Jensen (1990) to investigate the instability of the flow in a collapsed tube. Empirical orthogonal eigenfunctions are extracted from biomechanical simulations of the vocal folds by Berry et al. (1993); those authors show that chaotic oscillation can arise as a result of desynchronization of the low-order modes. Although those works represent a real progress with respect to the the former small-amplitude analysis, they lack the cleverness of fully analytical techniques.

In this respect, Lucero (1993) presented an analytically-based analysis of the large-amplitude oscillation of the vocal folds using the two-mass model. This technique consists, at first, in finding the equilibrium points of the dynamical equations of motion. As a second step, an analysis of stability is carried out, essentially by linearizing the system about those equilibrium points and by examining the sign of the real part of its characteristic equation. Although the results obtained were quite promising, the referred work was based on an oversimplified model of both the fluid mechanics aspects of the glottal flow and the geometry of the vocal folds. Furthermore, the coupling between the vocal fold oscillation and the acoustical loading of the vocal tract, as well as the effects of viscous resistances, were neglected.

The goal of the present study is twofold: first, we will redo the analysis of Lucero (1993) showing that some of his conclusions are due to the introduction of a 'spurious' element of the fluid mechanics. Second, we will apply a numerical version of the analysis of stability of the equilibrium point using a more realistic model of the glottal flow and including the effect of vocal tract loading.

ANALYSIS OF EQUILIBRIUM POSITIONS FOR THE TWO-MASS MODEL

We will proceed to a verification of the results of Lucero (1993) by eliminating the loss due to sharp edges (flow separation in vena-contracta effect; for more details see Pelorson et al. 1994). We will use the same notations as in Lucero (1993) and we ask the reader to refer to that paper for the meaning of the mathematical symbols. In the case of an open glottis ($x_1 > -x_{10}$ and $x_2 > -x_{20}$, where x_i and x_{i0} are the position and the rest positions for the masses 1 and 2), the driving force on the mass 1 is given by $F_1 = l_g d_1 P_S f_p$, where l_g is the width of the glottis, d_1 the length of mass 1 and P_S the sub-glottal pressure. The term f_p depends on the position of the masses and on a factor κ (see Ishizaka and Flanagan 1972) for sharp edges ($\kappa = 0.37$):

$$f_p = \frac{(x_1 + x_{10})^2 - (x_2 + x_{20})^2}{(x_1 + x_{10})^2 + \kappa(x_2 + x_{20})^2}.$$

As the contraction at the entrance of the vocal folds is smooth, we believe that there is no reason for the vena-contracta effect, κ having to be set to zero. In this case, equation (8) of Lucero (1993), obtained by setting the derivatives of the equations of motion to zero, becomes $(y_{1e} - 1) = H(1 - y_{2e}^2/(\beta^2 y_{1e}^2))$, where $\beta = x_{10}/x_{20}$, and H is a constant that depends on several parameters of the model, including the mass stiffnesses (k_1 and k_2). $y_{i0} = 1 + x_i/x_{i0}$ are the normalized mass displacements. The final solutions for $\beta = 1$ (rectangular prephonatory glottis) are (i) $y_{1e} = y_{2e} = 1$ (rest positions), and (ii) the solutions of the following equation

$$y_{1e}^2 + (1 - \alpha)(1 + \alpha)H y_{1e} - (1 - \alpha)^2 H = 0.$$

As H and $\alpha = k_c/(k_2 + k_c)$ are always positive, it is straightforward to prove that the roots of the above equations are always real and one of them is always negative. This invalidates the result of Lucero (1993), where there was possibility for the existence of three simultaneous equilibria. The main conclusion is that there will be always two equilibrium positions for any value of the command parameters of the model (the stiffnesses of the masses, related to α and the subglottal pressure).

NUMERICAL ANALYSIS OF STABILITY OF EQUILIBRIA

Lucero (1993) did an analysis of stability of the equilibrium points and found an analytical formulation for obtaining the bifurcation points and the regions of stability for the command parameters space. We extend the technique to a more realistic two-mass model (Pelorson et al. 1994) including the effects of moving flow separation point and viscosity losses, thanks to a more elaborated model of the changing geometry of the vocal folds. The acoustical loading at the outlet of the two-mass model is taken into account by modelling the input acoustical impedance of the vocal tract as a linear filter (plane wave propagation in the vocal tract is assumed). Hence the dynamical equations of motion for the two-mass model together with the dynamical

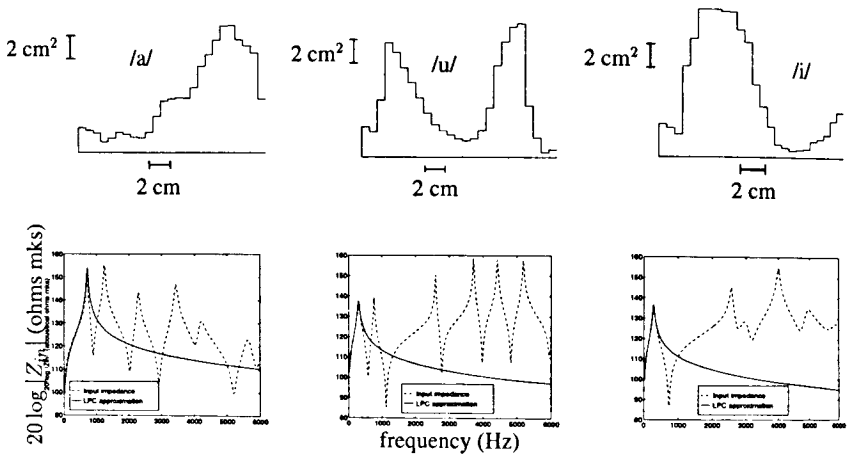


Figure 1: Area functions (top panel) and input impedances (bottom panel) for three French vowels: /a/ (left), /u/ (middle), and /i/ (right).

equation describing the input impedance filter compose the global equations of motion of the system.

More specifically, we approximated the effect of vocal-tract input impedance by a filter which takes into account only the first formant. Fig. 1 shows the area functions for three French vowels (/a/, /i/, and /u/) and the associated input impedances as a function of frequency $Z_{in}(f)$ computed from the area functions using the plane wave propagation hypothesis (dashed lines in Fig. 1). The poles and zeros of the vocal-tract impedance were computed from the impedance spectrum by a LPC approximation (solid lines in Fig. 1). We plan to include formants of higher order in a future work. By now, we are interested just in the effect of the first formant on vocal fold vibration and we believe that they will be more marked than the effect of the other formants.

The acoustical loading is modelled then by the pressure at the input of the vocal tract P_{CV}, which is the result of filtering the glottal flow U_g through the linear filter $Z_{in}(f)$. P_{CV} varies thus with time, perturbing the pressure difference across the glottis $P_S - P_{CV}$ (as a first approximation, P_S is considered independent of the glottal flow U_g in the present study). The whole model can be described by a set of augmented differential equations. The technique we used for finding equilibrium points and for analysing their stability consists essentially in linearizing the model about the equilibrium points using a perturbation analysis.

The system of non-linear differential equations can be compactly described using the notation:

$$\dot{u} = F(u),$$

where u are the state variables of both the mechanical and acoustical parts. About any equilibrium point $\overline{u}$, it is possible to linearize the system (see Guckenheimer and Holmes 1986 for details):

$$\dot{\xi} = D\xi$$

where $u = \overline{u} + \xi$ and D is the Jacobian of the function F about $\overline{u}$. Stability of the linearized system on state variables ξ depends on the eigenvalues of D. Specifically, in order to have a stable the system about the equilibrium point $\overline{u}$, the real part of all eigenvalues of D have to be negative.

We carried out this analysis varying two parameters of the model: P_S and k_1 (the stiffness of mass 1). The other free parameters were kept constant to typical values given in the literature (see Pelorson et al. 1994). Starting from the rest position for the masses, we found the equilibrium positions, which corresponded always to a convergent configuration

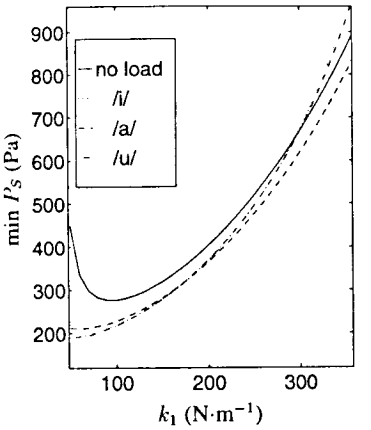


Figure 2: Minimum subglottal pressure for vocal fold oscillation as a function of the stiffness of mass 1.

for the vocal folds. For different values of k_1, we searched by a bisection algorithm the value of P_S in the boundary between the oscillating and non-oscillating regions. The results are shown in Fig. 2. It is possible to see that the region of stability is increased due to the acoustical loading of the vocal tract. The effect is more accentuated for the vowel /a/ and almost the same for vowels /i/ and /u/.

CONCLUSIONS

We extended the technique presented by Lucero (1993) for analysing the stability of equilibria of the vocal folds using a more realistic model of the flow through the glottis and including the effects of acoustical loading. We considered the analysis of the equilibrium points done by Lucero (1993) and we studied influence of some aerodynamical effects. We proposed a numerical technique for obtaining the stability of equilibria, being able to determine regions of larynx/lungs command space for which the vocal folds will oscillate. The main virtue of the proposed technique is the ability to determine the bifurcation boundaries in the control space (including vocal tract configuration) without having to run temporal simulations. Our results shows the importance of the aeroacoustics in such an analysis. We agree with Lucero (1993) that the next logical step would be the study of more realistic models. However we emphasize that in parallel to improvements on the mechanics, much has to be done concerning the aerodynamical and acoustical descriptions.

ACKNOWLEDGEMENTS

This research was supported by the European Community ESPRIT - BR grant # 6975 Speech Maps.

REFERENCES

Awrejcewicz, J. (1990). Numerical investigations of the constant and periodic motions of the human vocal cords including stability and bifurcation phenomena. *Dynamics and Stability of Systems 5*(1), 11–28.

Berry, D. A., H. H, I. R. Titze, and K. Krischer (1993). Interpretation of biomechanical simulations of normal and chaotic vocal fold oscillations with empirical eigenfunctions. Status and Progress Report 5, NCVS.

Guckenheimer, J. and P. Holmes (1986). *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields.* Heidelberg: Springer Verlag.

Ishizaka, K. and J. L. Flanagan (1972). Synthesis of voices sounds from a two-mass model of the vocal cords. *Bell Syst. Tech. J. 51*, 1233–1268.

Jensen, O. E. (1990). Instabilities of flow in a collapsed tube. *J. Fluid Mech 220*, 623–659.

Lucero, J. C. (1993). Dynamics of the two-mass model of the vocal folds: Equilibria, bifurcations, and oscillation region. *J. Acoust. Soc. Am. 94*(6), 3104–3111.

Pelorson, X., A. Hirschberg, R. van Hassel, A. P. J. Wilnands, and Y. Auregan (1994). Theoretical and experimental study of quasisteady-flow separation within the glottis during phonation. Application to a modified two-mass model. *J. Acoust. Soc. Am. 96*(6), 3416–3431.

Titze, I. R. (1988). The physics of small-amplitude oscillation of the vocal foldds. *J. Acoust. Soc. Am. 83*(4), 1536–1552.

CHANGES IN SPEECH PRODUCTION FOLLOWING HEARING LOSS DUE TO BILATERAL ACOUSTIC NEUROMAS

Joseph Perkell, Joyce Manzella, Jane Wozniak, Melanie Matthies, Harlan Lane, Mario Svirsky, Peter Guiod, Lorraine Delhorne, Priscilla Short, Mia MacCollin and Clay Mitchell

Massachusetts Institute of Technology and Massachusetts General Hospital

ABSTRACT

This is a report of speech production changes in a patient who lost hearing due to bilateral acoustic neuromas and received an auditory brainstem implant to provide some "auditory" stimulation. Speech production and perception and neurological status were measured multiple times before and after onset of hearing loss. "Postural" parameters, such as average vowel SPL, duration, and F0 changed with hearing status, whereas phonemic parameters, such as fricative spectra and VOT were more stable.

INTRODUCTION

Auditory feedback is clearly essential for speech acquisition. It is questionable, however, whether auditory feedback is necessary for speech production in adulthood, since the speech of people deafened as adults can remain intelligible for decades. However, their speech often develops abnormalities, indicating some role for auditory input in adult speech motor control [1-4]. Studies of changes in speech production when postlingually-deafened patients receive cochlear implants have led us to hypothesize that auditory feedback has at least two functions in adult speech motor control: (1) maintenance of the **phonemic settings** of a robust *internal model* (established during acquisition) of the relations between speech motor commands and the sound output and (2) monitoring the transmission channel to help make situation-dependent adjustments in **postural settings** of parameters that underlie *average sound level, rate, F0, low-frequency spectral slope* and *vowel formants,* which influence *clarity* and *intelligibility.* By inference, phonemic settings should be less labile than postural settings. Since phonemic settings and postural settings affect the same articulators, there can be *interactions* between them, but in some cases their changes can be observed separately [1-4]. The purpose of this study is to further investigate these hypotheses by studying speech changes in a patient who loses, rather than gains hearing.

METHODS

Bilateral acoustic neuroma (Neurofibromatosis 2, or NF2), is a rare hereditary disease characterized by benign tumors of the central nervous system, which tend to arise bilaterally on the eighth nerves and may lead to hearing loss, first on one side and then on the other (often from surgery that is required to remove the tumors to prevent more serious consequences). The symptomatology and severity can vary widely among patients, but a significant proportion have their most severe symptoms confined to bilateral hearing loss [5].

Subjects for this research are adult NF2 patients who are speakers of English, with good hearing in one ear and zero or near zero speech discrimination scores in the other ear; and little or no oro-sensory, speech-motor, or other speech-language problem. We have recorded 44 NF2 patients who met most or all of these criteria.

The subject of this report is the first patient (a 30-year-old female) who met the criteria and, during the course of the research, suffered profound hearing loss in her remaining good ear; we will refer to her as NFA (for NF2 Female subject A). During her surgery for tumor removal, the auditory nerve had to be severed, and the electrode array of an auditory brainstem implant was placed.

The auditory brainstem implant (ABI) has been developed (and NFA was implanted) at the House Ear Institute, Los Angeles, CA. It consists of an electrode array placed on the cochlear nucleus, trans-cutaneous electro-magnetic signal transmission and an external microphone and signal processor. The electrode array has seven active electrodes and one reference electrode, forming seven channels that are stimulated with an F0 F1 F2 F5 strategy, intended to provide spectral, amplitude and temporal information, including voicing. NFA's ABI processor was activated several weeks after implant surgery.

Recording sessions were conducted at -20, -10, -1, 11, 35, 60, 76, and 83 weeks relative to the time of the surgery that produced NFA's onset of hearing loss (OHL). Pre-OHL testing was done with NFA wearing her CROS hearing aid, and post-OHL testing was done with NFA using her ABI.

Assessments and complications

Each two-day session typically included: one or two recordings of speech acoustic and physiological parameters, a neurological exam, a set of speech perception tests, and, to monitor for motor changes, tests of non-speech oral-motor capabilities and a videotaping of the subject's face while reading a passage. Each post-OHL session also included an "on-off" experiment, described below.

Speech perception tests consisted of combinations of auditory alone, visual alone and auditory-visual presentations of: 12 consonants in a /Cac/ utterance, 8 vowels in /bVt/, 10 vowels in /bVd/, monosyllabic words (NU-6), suprasegmentals (SPAC) and sentences (CUNY).

Speech production measures were made of: SPL, F0, duration, H1-H2 (low-frequency spectral slope), F1 and F2 of the vowels /i, ɪ, ɛ, æ, ɑ, ʌ, ɔ, u/ spoken in /bVt/ in a carrier phrase; VOT for /p, b, t, d, k, g/ in /Cɑd/ in a carrier phrase, spectral properties of the sibilants /s/ and /ʃ/ in /Sɑd/ in a carrier phrase; average airflow rate, and intersyllable regulation of F0 and SPL in readings of the Rainbow Passage. This set of materials (or a subset) was repeated five times for each recording. (Aerodynamic and acoustic parameters of voice production were also measured, but are not covered in this report.)

The "on-off experiment" involved having NFA turn off the speech processor of her ABI for 24 hours, then recording five five-minute blocks of 10 repetitions of a subset of speech materials in which her speech processor was: off (1), on (2), on (3), off (4) and off (5) [4].

Motor losses were induced by two surgical procedures. The tumor removal resulted in damage to the left facial nerve which caused a readily-apparent left facial palsy. At week 72 (prior to the last two reported recordings), the left hypoglossal nerve was anastomosed to the facial nerve in an attempt to restore some left facial function. This procedure resulted in a tongue motor deficit. The tongue deficit was not obvious, but it was confirmed by the non-speech motor test. Clearly, the deficits influence the interpretation of much of the production data. In addition, NFA had an upper respiratory tract infection during the recording one week before surgery, which might also influence some results.

RESULTS

Speech perception

NFA had good aided hearing pre-OHL. For example, auditory-alone consonant scores were close to 90%. Post-OHL, those scores were consistently poor (about 17% correct). It appears that by week 83 NFA was getting some benefit from her ABI (mainly indicated by improvement in consonant scores from visual-alone to auditory-visual). NFA had good visual-alone speechreading scores (about 74% correct) which remained consistent pre- to post-OHL.

Results for the suprasegmental materials were generally also good pre-OHL and dropped dramatically post-OHL. Scores for these tests post-OHL were better in the auditory-visual than the visual-alone condition. Thus the ABI seems eventually to have provided some additional cues to speechreading.

These results were consistent with clinical reports that NFA does not discriminate well among the different channels of her ABI. Presumably, then, the ABI provides her with little spectral information, but does convey some F0, loudness and voiced/unvoiced information, which she was beginning to use by week 83.

Speech production

In general, the left facial palsy could have a post-OHL effect on many supraglottal parameters; however, some of those parameters should be more affected than others. For example, bilabial consonants are obviously influenced and

velars shouldn't be. Parameters that reflect laryngeal and respiratory function should be uninfluenced. The anastomosis surgery at 72 weeks should have only influenced subsequent tongue articulations. Nevertheless, it is possible that NFA developed compensatory strategies using structures that were not directly affected by the surgeries.

Postural parameters

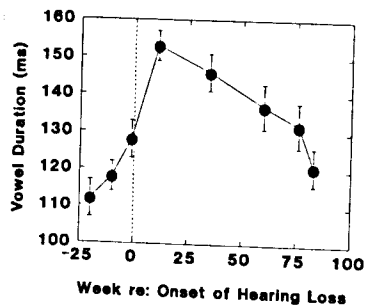


Figure 1: Mean duration (ms) for the eight vowels vs. weeks re time of OHL.

Figure 1 shows mean duration (ms) for the eight vowels, versus weeks re time of OHL, which is represented by the vertical line. Each point represents the average values for five repetitions of the eight vowels; the error bars show +/- one standard error. There is a slight, increasing trend pre-OHL, a further large (25 ms) increase between weeks -1 and 11, and then a gradual return to near pre-OHL values. Roughly-analogous patterns were shown by average vowel SPL, F0, average airflow (from lung volumetric measurements during the Rainbow Passage) and vowel H1-H2 (the amplitude difference between the first two harmonics in the acoustic spectrum, a measure that correlates with the degree of glottal abduction). The patterns give a general impression of an initial post-OHL change to more "deaf-like" speech [3], with a gradual return to pre-OHL values, as NFA was presumably beginning to use cues from the ABI. However, the pre-OHL trends and overall variability of the data introduce uncertainty about the effect of the hearing change on the speech parameters.

Figure 2 shows average vowel duration vs condition (processor on or off)

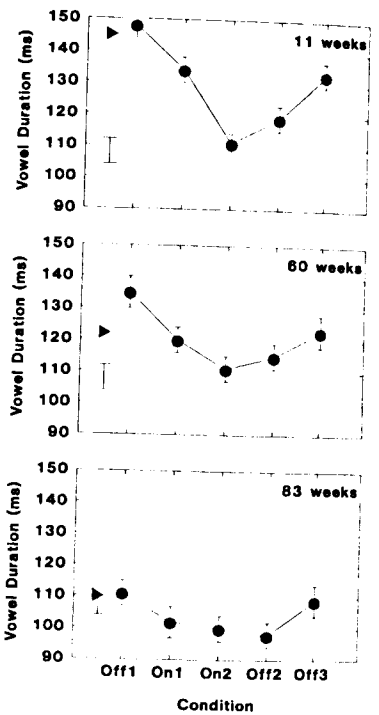


Figure 2: Average vowel duration vs condition (processor on or off) at 11, 60 and 83 weeks post-OHL.

for experiments at 11, 60 and 83 weeks post-OHL. Each point represents the average of 10 repetitions of the vowels /i, ɪ, ɛ, ɑ, u/. The triangle indicates the mean value of this 5-vowel subset from the longitudinal recording made in the same session, and the vertical bar indicates the range between the mean values (of this subset of 5 vowels) in the first two pre-OHL recordings. Among the three plots, the overall values of the on-off data correspond approximately to the respective longitudinal values. In each plot, duration is high in the initial processor off condition; when the processor is turned on, duration drops, then continues to drop; and when the processor is turned off again, duration rises. The magnitude of the effect corresponds to the distance between the respective current longitudinal value and the pre-OHL range. Thus, there is a clear effect of the auditory stimulation: with it NFA's speech is faster. This result helps to

counteract uncertainty about interpreting the longitudinal data because of the pre-OHL trends and variability.

Segmental parameters

Longitudinal plots of values of spectral median and symmetry [2] on the sibilants /s/ and /ʃ/ showed no meaningful change until week 60. Then, after the anastomosis surgery, values for the two sounds began to converge, indicating a blurring of the contrast that could not be corrected without the aid of auditory spectral information. The relative stability of the sibilants for 60 weeks post-OHL is consistent with the hypothesis that phonemic settings are robust.

Values of voiced and voiceless VOT (corrected for syllable duration – [1]) did not change longitudinally, in spite of changes in the related parameters of SPL, F0 and H1-H2. The stability of VOT could be due to the use of temporal information delivered by the ABI. On the other hand, voiced and voiceless VOTs are well separated in some speakers decades after onset of deafness in adulthood [1], so we might not expect large VOT changes to begin with.

Intersyllable regulation of F0 and SPL

Measures of syllable-to-syllable fluctuations in SPL and F0 (normalized for overall levels) in readings of the Rainbow Passage were compared between the first two pre-OHL sessions and the two post-OHL sessions at 35 and 60 weeks. The amounts of fluctuation in both SPL and F0 were significantly higher in the post-OHL data.

DISCUSSION AND CONCLUSIONS

The results presented support our hypotheses about differences between postural and phonemic settings, and they are consistent with the following interpretation. Soon after experiencing a serious loss of hearing and introduction of a novel and relatively undifferentiated kind of "auditory" stimulation, NFA's speech became more like that of a deaf person: slower, louder, and with an abnormal (for her) F0 [3]. As indicated by the on-off results, at all times, the postural parameters were sensitive to hearing status, i.e., relatively labile; however, it took NFA about a year to learn how to use the relatively crude auditory input to re-adjust her postural settings to the levels she had been using when she had useful natural hearing. Throughout this dramatic change in hearing and recalibration of postural settings, the two measured phonemic settings remained stable, indicating their robust nature. So far, inter-syllable regulation of F0 and SPL, measures of control variability, seem to have the lability of postural settings, but more data are needed.

We caution that we have chosen examples that illustrate our points. Although we have not found clear counter-examples, the data are very complicated and variable, and not all results are as easy to interpret. Only a fraction of the available results can be reported here, and new recordings and analyses are being added to NFA's picture. Finally, we are beginning to gather similar data on additional subjects; some of those data may contain fewer confounds and thus may be easier to interpret.

ACKNOWLEDGMENTS

We are very grateful to the Acoustic Neuroma Association, Neurofibromatosis, Inc.; to all of our subjects – each one of them an extraordinary person; and most especially, to NFA. This work was supported by N.I.D.C.D.

REFERENCES

[1] Lane, H., Wozniak, J. and Perkell, J.S. (1994). Changes in voice-onset time in speakers with cochlear implants, *J. Acoust. Soc. Am.*, vol. 96, 56-64.

[2] Matthies, M.L., Svirsky, M.A., Lane, H., and Perkell, J.S. (1994). A preliminary study of the effects of cochlear implants on the production of sibilants, *J. Acoust. Soc. Am.*, vol. 96, 1367-1373.

[3] Perkell, J.S., Lane, H., Svirsky, M.A. and Webster, J. (1992). Speech of cochlear implant patients: A longitudinal study of vowel production, *J. Acoust. Soc. Am.*, vol. 91, 2961-2979.

[4] Svirsky, M.A., Lane, H., Perkell, J.S., and Webster, J. (1992). Speech of cochlear implant patients: Results of a short-term auditory deprivation study, *J. Acoust. Soc. Am.*, vol. 92, 1284-1300.

[5] Parry, D.M., Eldridge, R., Kaiser-Kupfer, M.I., Bouzas, E.A., Pikus, A. and Patronas, N. (1994). Neurofibromatosis 2 (NF2): Clinical Characteristics of 63 affected individuals and clinical evidence of heterogeneity, *Am. J. Med. Genetics*, vol. 52, 450-461.

THE EFFECTS OF COCHLEAR IMPLANTS ON SPEECH PRODUCTION IN POSTLINGUALLY ACQUIRED DEAFNESS

R. Cowie, E. Douglas-Cowie, M. Sawey & G. Mulhern
School of English / School of Psychology, Queen's University, Belfast, UK

ABSTRACT

Acoustic analyses were made of speech recordings from 75 deafened subjects before and after cochlear implantion, and 51 controls, using the ASSESS system. Pre-implant speech was abnormal in timing, intensity range, pitch height and change, frication, and spectral balance. Implantation reduced some anomalies, left some unchanged, and aggravated others. Many effects were sex-related.

INTRODUCTION

This paper reports part of a large-scale evaluation of cochlear implantation in the UK, co-ordinated by the MRC Institute of Hearing Research. We studied effects of implantation on speech production.

Our evaluation involved three phases. Two were auditory. Phoneticians rated speakers on a range of dimensions: and naive listeners' impressions were studied. This paper describes the third phase, which used the ASSESS system to obtain objective acoustic measurements.

ASSESS develops previous work on properties which can be measured automatically and which appear to reflect disorders of speech [1],[2]. It is based on standard descriptors - spectrum, intensity contours, and pitch contours. It forms a rich description by breaking these into significant units. Preprocessing finds inflections in the contours (points where volume or pitch stops rising and starts falling, or vice versa). Contours can then be described as a series of rises and falls. Blocks which correspond at least roughly to natural units are also found - silences, sound blocks, tunes, and fricative bursts. Sound blocks are defined by the way intensity rises after a silence, peaks, and falls to the next silence. Tunes are defined by the way pitch moves between two silences long enough to be considered pauses. Fricative bursts are defined by energy in the upper spectrum.

ASSESS generates a systematic statistical summary of these elements and higher order attributes derived from them. A fuller description is given in [3].

METHOD

The study considered 51 normal hearing and 75 deafened subjects. All of the latter were recorded pre-implant and 9 months after, and 29 were also recorded 18-24 months after. The reading material analysed was the "Rainbow passage".

After processing through ASSESS data were inspected graphically and a few gross 'outliers' were removed - usually about two or three per passage.

Absolute level measurements were unavailable, and intensity measures were normalised by setting median intensity at the start of each passage to 60dB. This is reasonable given that in auditory ratings controls and pre-implant patients scored almost identically on average volume, and post-implant speakers' ratings showed a significant but small trend towards lowered volume.

RESULTS AND DISCUSSION

The main statistic used was analysis of variance. Independent variables were speakers' sex and hearing status - pre-implant, 9 months post, 24 months post, and control. Hearing status was treated as a between groups variable. This is conservative - if anything it tends to underestimate effects of implantation.

Timing

Table 1 shows that deafened speakers spoke more slowly than controls. The effect is significant (F3, 213 =8.8, p<.0001). Implantation does not reduce the problem: if anything it worsens it.

Table 1: Reading time excluding pauses (in seconds)

	total duration		n of pauses	
	fem	male	fem	male
pre implant	36.5	35.2	62.5	48.2
9 mths post	36.8	35.9	63.9	53.9
24 mths post	37.9	36.8	68.5	60.2
controls	29.4	30.0	53.6	46.2

The effect is not due to pausing. However the number of silences is high in deafened speakers, and significantly higher after implant (F2,170= 4.1, p=.018). No significant change was found in the duration of silences.

Deafened speakers show too many discontinuities in general - not only silences, but also inflections in the intensity and pitch contours. Table 2 shows two relevant measures, numbers of rises in the two contours. With pitch, the overall contrast including the controls is highly significant (F 3, 211=9.5, p=.0001), but implants do not affect the anomaly significantly (F 2,170=0.4, p=.67). With intensity as with silences, the effect worsens significantly post implant (F 2,170=4.3, p=.016).

Table 2: Numbers of rises

	Intensity		Pitch	
	female	male	female	male
pre implant	93.7	87.1	57.7	54.7
9 mths post	97.3	91.2	59.6	56.2
24 mths post	102	98.1	55.8	60.5
controls	81.3	79.8	46.7	45.1

Some aspects of timing do improve with implant, though. Rises and falls in intensity tend to last too long in deafened speakers, as is seen in median durations of rises and falls for each speaker (Table 3). Improvement after implantation seems marginal when the measures are analysed separately, but analysing them together shows a robust effect (F 2,173= 4.1, p=0.018). Improvement is essentially complete 9 months post implant.

Table 3: median durations of rises and falls in intensity (in milliseconds)

	Rises		Falls	
	female	male	female	male
pre implant	79.1	79.2	90.1	86.6
9 mths post	75.9	78.0	82.3	85.6
24 mths post	75.0	78.9	82.9	87.5
controls	74.9	75.0	81.1	80.0

Pitch shows a partially similar trend (Table 4). For the deafened as a whole, median pitch falls are too long. Fall length reduces with implantation (F 2,170 =3.5, p=0.03). But sex complicates the trend. Pre-implant females already have shorter pitch falls than control females, but the reduction in fall length occurs for both sexes. This is an improvement for the males, but the effect on females is that 24 months post implant, their pitch fall is considerably too short.

Table 4: Median pitch fall duration (ms)

	female	male
pre implant	95.5	106.6
9 months post	91.9	91.0
24 months post	82.4	95.9
controls	101.8	92.2

These findings emphasise the need to be wary of global statements about timing. Anomalous shortening may occur because deafened people with implants have a rather undiscriminating sense that they should liven up their speech.

Intensity

The clearest intensity effects involve spread measures, particularly interquartile range (IQR), which spans the middle 50% of observations. IQR is too high in pre-implant patients and falls following implantation (Table 5). The fall is significant with F 2,172= 6.8, p=.001. It may continue after 9 mths post implant.

Table 5: Intensity IQR (in dB)

	female	male
pre implant	11.56	12.94
9 mths post	10.67	11.17
24 mths post	9.98	11.05
controls	9.96	9.30

Table 6 clarifies the effect by showing the limits of speakers' usual range, the 10% point (below which intensity falls less than 10% of the time) and the 90% point (analogously defined).

The 90% point is strikingly stable, but pre-implant subjects have a low 10% point - i.e.they overuse rather low levels. Post implantation the 10% points rise significantly (F 2,172= 6.8, p=.001) - i.e. implants narrow intensity range by raising the lower limit.

Table 6: Intensity extremes (in dB)

	10% point		90% point	
	female	male	female	male
pre implant	48.4	46.9	67.2	67.0
9 mths post	49.9	48.7	67.5	67.1
24 mths post	50.9	48.8	67.7	66.5
controls	49.6	50.6	66.6	66.9

The large rises and falls which begin and end sound blocks are distinctive. Table 7 shows that they are much longer than rises and falls in general, as would be expected. They are also a case of change which continues after 9 months post implant. Considering patient performance on rises and falls together shows an effect of time (F 2, 171=4.6, p=.011). Post hoc tests show that the significant contrast ($p<0.01$) is between pre implant and 24 months post.

As with pitch falls, the changes are appropriate for males. But they leave 24 month post implant females with shorter rises and falls than control females.

Table 7: Durations of rises and falls which begin and end blocks (in ms)

	opening rises		closing falls	
	female	male	female	male
pre implant	151	153	187	202
9 mths post	143	152	171	193
24 mths post	140	144	162	182
controls	151	135	171	166

Pitch

There are no strong, straightforward pitch effects, partly because of occasional extreme values, but measures which bypass these extremes show effects of hearing loss and of implantation.

One such measure comes from the midpoints of quadratic curves which are fitted to tunes. The interaction between sex and hearing status falls short of significance in the full analysis (F 3, 212 =2.4, p=.067) but reaches it in the analysis which considers only pre-implant and control subjects (F 1, 117 =5.7, p=.019). As table 8 shows, female pitch is never far from normal, but male pitch is high pre-implant and remains so.

Table 8: Fitted midpoints of tunes (in Hz)

	female	male
pre implant	192.6	140.5
9 mths post	192.19	132.7
24 mths post	186.2	144.1
controls	198.0	115.5

Extreme pitch changes also show a sex-related pattern, as shown in Table 9. Significant sex*hearing status interactions occur with all these measures - 10% points for rises (F 3, 213=3.9, p=.01) and falls (F 3, 213 =4.0, p=.008) and 90% points for both (rises F3,213=3.2, p=.023, falls F3, 213 =2.9, p=.035).

Table 9: Extremes of pitch change per rise or fall (in Hz)

		10% point		90% point	
		female	male	female	male
pre implant	R	1.39	1.49	22.6	23.0
	F	1.65	1.80	32.0	33.0
9 mths post	R	1.36	1.29	21.8	18.6
	F	1.66	1.45	32.1	23.1
24 mths post	R	1.38	1.32	21.2	19.4
	F	1.51	1.50	30.2	24.5
controls	R	1.51	1.22	29.1	16.3
	F	1.80	1.40	37.8	21.5

As with mid pitch, sex differences are reduced pre-implant. Both extremes are high in pre-implant males and low in pre-implant females. Implants reduce change, taking males towards control norms and females away from them.

Pitch variability, both within and between individuals, is strongly reduced by implantation. Within individuals, variability shows in the movements which open and close tunes. Table 10 shows the standard deviation of the slopes of these movements. This reflects the extent to which patients vary the pitch movements which begin and end tunes. Analysis of variance considering the patient groups on both measures shows a significant effect of time (F 2,170=3.1, p=.047). Post hoc tests show that the only significant difference is between pre-implant and 24 mth groups.

Table 10 : variability of initial and final pitch movements in a tune (Hz/sec)

	initial pitch movements		final pitch movements	
	female	male	female	male
pre implant	95.7	86.0	93.4	84.3
9 mths post	87.4	73.3	96.7	68.0
24 mths post	77.9	72.1	77.2	66.1
controls	99.9	76.3	99.9	58.7

Again, reductions in variability mark a move towards normality for the males and away from it for females.

In several measures variance within the pre-implant group is abnormally high because some individuals lie beyond the normal range. Pre-implant males show too wide a range of pitch variability,

which narrows post implant. Females show no consistent change in variability. Females pre-implant show an abnormal range of movements at the beginnings and ends of tunes: there is marked narrowing post-implant. The male pattern is probably similar, but less consistent.

Males also show an abnormally wide range of values for properties involving tunes' mean height and shape. The range of mean heights is wide before implant and remains so. The shape measures reflect two patterns which are uncommon among controls: tunes which start low then rise steadily in pitch, and tunes which drop pitch in the middle. At 24 months post implant, half of the males showed at least one of these patterns.

Frication and the spectrum

Pre-implant patients show under-frication on all measures - number of bursts, average duration of a burst, and level of fricative energy in a burst. There are significant effects of hearing status on all three variables (respectively F3, 210 =3.4, p=.019, F3, 209=4.1, p=.007, and F3, 208=5.0, p=.002).

Implantation has an effect. When only patients are considered, all three variables show effects of hearing status which are significant or nearly so (F2,169 =3.2, p=.042, F2,168 =3.0 p=.053, and F2, 167 =3.0, p=.054 respectively). Energy level changes towards the control pattern, but for number and duration of bursts, initial change is in the wrong direction.

Subspectra for fricative bursts and peaks in the intensity contour (the best simple approximation to vowel centres) are summarised by the slope the profile of energy against frequency, and the mean, marking the spectrum's centre of gravity.

Table 11 gives slopes and means for fricative spectra. Hearing status affects both (slope F3,127 =4.6, p=.004, mean F3, 127=4.1, p=.008). Essentially patients show too little energy in the upper spectrum, before and after implant.

Table 11: shapes of fricative spectra

	slopes (dB/8ve)		means (Hz)	
	female	male	female	male
pre implant	-.175	-.156	861	866
9 mths post	-.050	-.147	909	873
24 mths post	-.111	-.079	879	892
controls	-.009	+.012	920	924

Table 12: shapes of intensity peak spectra

	slopes (dB/8ave)		means (Hz)	
	female	male	female	male
pre implant	-0.94	-0.97	630	608
9 mths post	-1.11	-1.00	588	605
24 mths post	-0.99	-0.94	626	619
controls	-0.90	-0.83	636	663

Table 12 shows that deafened speakers also lack energy in the upper intensity peak spectrum. Again hearing status has significant effects (slope F3, 127=3.8, p=.011, mean F3,127 =2.8, p=.043), but implantation does not. The parallel with frication suggests that speakers may have a general problem with the upper spectrum rather than frication as such.

There is evidence that deafened speakers fail to distinguish fricatives spectrally [4]. ASSESS provides a related measure, variation in the centre of fricative energy in bursts. Hearing status has an effect in the expected direction (F3, 202=4.1, p=.008). Implantation has no significant effect.

CONCLUSION

Objective measures show that speech production changes after implantation, but not always in the right direction. This may not be surprising given the level of input that current devices provide. Speech production may be a sensitive monitor of improvements in implant technology.

REFERENCES

[1] Cowie, R. & Douglas-Cowie, E. (1992), *Postlingually acquired deafness: speech deterioration and the wider consequences*, Berlin: Mouton de Gruyter.

[2] Andreasen, N., Alphert, M. & Merrill, J. (1981), "Acoustic analysis: an objective measure of affective flattening", *Arch. Gen. Psychiatry*, vol 38, 281-285.

[3] Cowie, R., Sawey, M. & Douglas-Cowie, E. (1995), "A new speech analysis system: ASSESS (Automatic Statistical Summary of Elementary Speech Structures)", *Proc 13th ICPhS*, Stockholm.

[4] Lane, H. and Webster, J. (1991), "Speech deterioration in postlingually deafened adults", *J. Acoust.Soc.Am.* vol. 89, pp. 859-866.

SIVO-II: A SPEECH ANALYSING HEARING AID FOR PROFOUNDLY HEARING IMPAIRED PEOPLE

A Faulkner[1], J R Walliker[1], F Coninx[2], M Dahlquist[3], C Beijk[2], E Fresnel-Elbaz[4], K J Smith[1], J Wei[1], A Bosman[5], G F Smoorenburg[5], and A J Fourcin[1] for the STRIDE Consortium

Department of Phonetics and Linguistics, University College London [1]. Institute voor Doven at St Michielsgestel[2]. Department of Speech Communication and Music Acoustics, KTH[3], Fondation Rothschild (Paris)[4], University Hospital Utrecht [5]

ABSTRACT

The SiVo-II speech analysing hearing aid has been developed and assessed as part of the STRIDE project [1]. The project aimed to assess the potential of a speech pattern analysing mode of processing in a hearing aid for the profoundly hearing impaired. Trials in this user group in four European countries indicated that the speech analytic approach has significant advantages, especially for speech perception in noise

INTRODUCTION

Speech analysis has been proposed as an important component of signal processing in hearing aids for the profoundly impaired [2]. This approach has several potential advantages. It facilitates the matching selected speech information to residual hearing ability, and can make use of noise-resistant speech analysis. It could, moreover, lead to a commercial product that is highly cost-effective compared to alternatives such as cochlear implants

THE SIVO-II AID

The SiVo-II [3] is a wearable body-worn unit that allows speech information processing algorithms to be tested by profoundly hearing impaired subjects in daily life. It makes use of a TMS320C50 fixed-point DSP. The SiVo-II aid processes speech to extract voice frequency and amplitude information and match this to residual low frequency hearing. The signal presented by the aid is a sinusoid whose frequency and amplitude are controlled so as to preserve voice pitch and loudness information. Mapping of both frequency and intensity range are employed to ensure that the signal is always audible without discomfort.

To achieve optimal matching to residual hearing, the SiVo aid has been designed to act as its own audiometer. The audible intensity range at each frequency is determined using the user's standard transducer and ear-mould, ensuring correct calibration. This range is then directly used to control the intensity range of the aid's output.

Noise resistant fundamental frequency extraction

Fundamental frequency extraction is carried out by a Multi-Layer Perceptron (MLP) artificial neural network. The MLP algorithm has been trained to produce from the acoustic speech signal an output pulse corresponding to each instant of larynx closure, and hence, to give a cycle-by-cycle estimate of voice fundamental frequency. Training is achieved by adjusting the coefficients of the MLP to maximise the correspondence of its output to a target signal derived from an electrolaryngograph. It has proved possible to train the MLP to operate effectively with speech to noise ratios in the range of 5 to 10 dB, which are known to cause severe difficulty to profoundly hearing impaired users of conventional hearing aids.

In noise, the MLP method has been found to perform as well as other methods that are capable of implementation in a wearable processor [4,5]. For the classification of speech as voiced or voiceless, the MLP method was superior to all others. Since voicing information is a primary source of lipreading support, the MLP method was judged appropriate.

POLY-LANGUAGE ASSESSMENT TOOLS

One objective of the STRIDE project has been to provide assessment methods that are comparable between different languages and test centres. These included the following two tests that were employed in the user trials:

Vowel-Consonant-Vowel Tests

Vowel-consonant-vowel (vCv) tests of audio-visual consonant perception have been defined as a segmental basis for quantitative comparison across languages. Tests have been prepared using 10 consonants (**p b m f v t d s n l**) that are common to Dutch, English, French and Swedish.

Prosodic Test

In these same four languages, perceptual stress/accent can be cued by a major change in fundamental frequency on the contextually important word, and this provides the basis for a common approach to prosodic assessment. Tests have been designed using lists of three mono-syllabic word utterances with sentence stress on one of the three words cued by a falling pitch pattern..

USER TRIALS

Four clinical centres took part, each in a different country and using a different language, with a total population over the two phases of 34 participating hearing aid users. The current report concentrates on the phase II trials, in which 22 adult users took part. The selection criterion was a limited ability to make use of conventional hearing aids to aid lipreading in the quiet, and a profound post-lingual sensori-neural hearing loss. The age, duration of deafness and hearing losses of the selected group are shown in table 1.

Subjects received wearable SiVo-II aids and had conventional aids refitted where clinicians judged that the existing aid was not optimal. Speech perceptual assessments were carried out before and after a training period. Training was provided over a period of typically 6 to 8 weeks, during which users attended for three or four training sessions of approximately 2 hours duration. During each training session, matched training was given in the use of both the SiVo-II and the conventional hearing aid (CHA).

Table 1. Summary of hearing-impaired user group characteristics.

Country		FR	SW	NL	UK	All
	n	5	2	7	8	22
Age	*mean*	59	56	53	48	53
	min	35	33	35	24	24
	max.	70	78	85	71	85
Years of Deafness	*mean*	42	33	23	23	28
	min	30	10	6	6	6
	max.	60	56	:50	53	60
250 Hz HL (dB ISO)	*mean*	87	83	78	86	83
	min	75	70	15	70	15
	max.	95	95	105	100	105
0.5, 1 & 2 kHz HL (dB ISO)	*mean*	106	99	104	110	106
	min	90	93	93	93	90
	max.	125	105	110	123	125

Three tests of receptive ability were administered: vCv consonant lipreading with and without support from an aid; stress pattern recognition; and "connected discourse tracking" (CDT) using texts designed for language learners and of similar complexity across languages. CDT was performed during the training sessions, of which it formed an integral part. Specially designed questionnaires provided a more global evaluation.

The results show rather marked variation from one country to another, and over individual users within each country. While the results combined across the users in all four countries show the SiVo-II aid to be performing only at the same level as a conventional hearing aid, the UK subject group as a whole, and individual users in other countries, showed a clear advantage from the SiVo aid.

Objective measures of speech receptive benefit

Where data come from all four field trial centres, the dominant factor in each case is that of country. In the CDT and stress test, the aid effect depends significantly on the country.

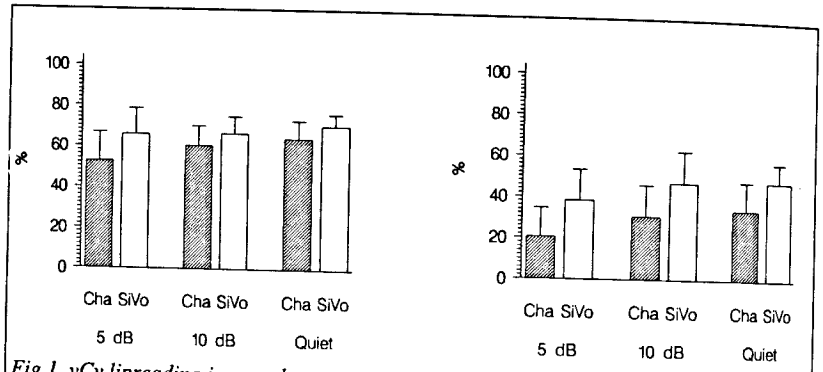


Fig 1. vCv lipreading in speech spectrum shaped noise. The figure shows the percentage correct score (left panel) and percentage of voicing information received (right panel) when using the SiVo-II aid (white bars) and the reference conventional hearing aid (hatched bars). Performance is shown for speech in quiet and speech-to-noise ratios of 5 and 10 dB. The error bars show the upper 95% confidence limit.

One unexpected factor in the results from different countries is now known to have been the sound reproducing loudspeakers used in Sweden and the Netherlands. The SiVo-II analysis algorithms are designed to operate with live speech, and require the preservation of the temporal characteristics of the speech waveform for correct operation. After results were collected, it was found that the SiVo analysis cannot operate properly with the acoustic signal from the loudspeakers used at IvD and KTH. As a result, the Swedish and Dutch data from tests using recorded speech (vCv and stress tests) at IvD and KTH cannot be taken to be representative of the benefit that would be expected using live speech.

vCv in noise (Fig 1)

The noise-resistance of the SiVo processing was expected to be a major factor in objectively measured benefit. vCv tests in noise were made in the UK. Data were collected at different background noise levels, and from a group of 11 SiVo-II users, three of whom were not in the STRIDE field trial group, but met the criteria for inclusion. There was a significant advantage from the SiVo-II aid compared to the CHA [$F(1,10) = 9.84$, $p = 0.011$]. There was also a significant effect of noise level [$F(2,17) = 4.63$, $p=0.025$]. At the poorer signal-to-noise ratio (5 dB) the scores obtained with the CHA were equivalent to scores obtained by the same subjects in unaided lipreading, that is, there was no benefit from the CHA. At this signal-to-noise ratio, however, the SiVo-II aid is still providing a useful degree of lipreading support. This advantage comes from the preservation of voicing information in the output from the SiVo aid at levels of noise that prevent the perception of this information from simple amplification of the noise+speech signal.

VCV in quiet

In quiet, there was no significant different between the SiVo-II and CHA across the group as a whole or in any of the four countries. Both aids significantly assisted lipreading. There was considerable individual variation, with individual users in each country showing greater benefit from the SiVo-II aid. The data from the Swedish and Dutch subjects are likely to have been affected by the interaction between the SiVo speech analysis and loudspeaker performance.

CDT

The UK group scored significantly higher in CDT with the SiVo-II than the CHA [$F(1,7) = 5.64$, $p<0.05$]. The other user groups showed no significant difference between the two aids.

Stress placement

Once more the UK group scored significantly higher with the SiVo-II than the CHA [$F(1,7) = 8.33$, $p<0.025$]. The other user groups showed no significant difference between the two aids. The main sources of variation in the overall data were country [$F(3,17) = 20.82$, $p <0.005$] and an aid*country interaction [$F(3,17) = 8.3$, $p<0.01$]. There were also effects due to voicing [$F(1,16)=6.3$, $p<0.025$], where phrases using voiceless consonants gave higher scores higher than all voiced phrases, and an aid*voicing interaction [$F(1,16) = 6.6$, $p<0.025$].

Subjective Evaluation

After training, users completed a questionnaire. Overall, users were slightly more satisfied with their conventional hearing aid (CHA) than with the SiVo. This outcome is largely due to the Dutch users, who on the whole preferred the CHA. The Swedish and French users preferred the SiVo-II, and UK subjects showed no clear preference. Considering the difference in convenience of use between a post-aural conventional aid and the bulky SiVo-II, this is an encouraging outcome.

Users were asked whether they would wish to use a SiVo-based hearing aid, either in its current form, in a smaller package that also included an amplification mode. Fifteen of the 22 users taking part in the trial expressed the wish to use a smaller aid that offered both types of processing.

CONCLUSIONS AND FURTHER DEVELOPMENTS

The STRIDE user trial has shown that the approach has clear benefits for some users. The speech analytic approach appears to be especially useful for providing lipreading support in noise. From the data gathered in the project, it is estimated that up to 160,000 persons may be able to benefit from an aid that incorporates both speech pattern extraction and amplification.

Improvements in the performance and design of speech pattern extracting aids are required before an acceptable product could be developed. The sensitivity of the SiVo analysis algorithms to loudspeaker performance also requires to reduced, so that television and other domestic sound reproduction equipment can be used more effectively by users of such aids.

The new EC TIDE project OSCAR is now in progress, and will address these issues. A new version of the aid, SiVo-III, has been developed, which also encodes voiceless excitation information. This has been shown to significantly add to the lipreading support available from fundamental frequency and amplitude information [6]. The OSCAR project is also examining the potential of tactile and combined auditory/tactile aids using speech analytic processing.

ACKNOWLEDGEMENT

Supported by EC TIDE project TP133/206 (STRIDE) and UK MRC grant G9020214. Other partners in the STRIDE project were: Laryngograph Ltd, Oticon A/S, and CNRS Parole et Langage at Aix-en-Provence, CCA Wagram (Paris).

REFERENCES

[1] STRIDE (1995) "Final report of TIDE project TP133/206 STRIDE", ref.: STRIDE/1995/2, Dept Phonetics and Linguistics, University College London

[2] Fourcin, A. J. (1990). "Prospects for speech pattern element aids," Acta Otolaryngol. (Stockh), Suppl. 469. 257-267.

[3] Walliker, J., Daley, J., Smith, K., Faulkner, A. and Fourcin, A. (1993) "Speech analytic hearing aids for the profoundly deaf: Technical design aspects and user field trials results" In: B. Granstrom, S. Hunnicutt and K-E. Spens (Eds) *Speech and language technology for disabled persons,* ESCA/KTH, Stockholm, pp 35-43

[4] Bosman AJ and Smoorenburg, GF (1994) "TIDE project 206 Internal Report. "Evaluation of Three Pitch Tracking Algorithms", ref.: STRIDE/1994/2. Dept Experimental Audiology, University Hospital Utrecht,

[5] Wei J, Howells D, Fourcin AJ, and Faulkner A. (1993) "Larynx period and frication detection methods in speech pattern hearing aids" Proc. Eurospeech '93, 3rd Eur. Conf. on Sp. Comm. and Tech. Vol. 3. ESCA, pp 2037-2049.

[6] Rosen, S., Faulkner, A., Reeve, K. and Smith, K. (1995) Voicing, fundamental frequency, amplitude envelope and voicelessness as cues to consonant identity. Proc. ICPhS, Stockholm, 1995

VOICE PRODUCTION AS A FUNCTION OF ANALYTIC PERCEPTION WITH A SPEECH PATTERN ELEMENT HEARING AID

○C TOFFIN, *K-E SPENS, □K SMITH, □R POWELL, ○P LENTE, □A FOURCIN, □A FAULKNER, *M DAHLQVIST, ○E FRESNEL-ELBAZ, #F CONINX, #C BEIJK, *E AGELFORS, □E ABBERTON
*KTH, SW □UCL, UK ○FONDATION ROTHSCHILD, FR #ST MICHELGESTEL, NL

ABSTRACT

The work described concerns an aspect of the way that real-time auditory feedback of a speaker's voice pitch information can have an appreciable influence on his or her control of vocal fold vibration in continuous speech. Larynx frequency range, vocal fold vibrational regularity and even detailed aspects of voice quality may, for a few profoundly deaf people, be markedly affected. The results discussed arise from the use of a phonetically motivated hearing aid, SiVo (Sine Voice), which has been used in controlled field trials in four countries with a total of 22 profoundly hearing impaired patients. The SiVo aid responds only to the voiced segments of speech and is designed to provide, for each separate input larynx period, a sine wave output which is matched to its user's residual hearing ability. In this first phase of work, the basic noise resisting neural network processing has been trained only on the use of targets produced by English speakers.

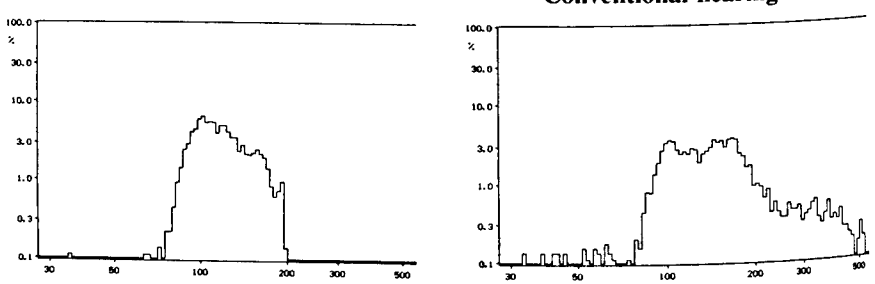
The training of the hearing impaired patients has been equally based on the balanced use of their SiVo and conventional aids and has involved their perception of intervocalic consonantal contrasts, single segment question / statement intonations, and SVO stress placement. No interactive speech production training based on the use of visually presented pattern element displays has been used for this phase of the work.

BACKGROUND

A suite of analytic programs has been applied to the quantitative assessment of the speech and laryngographic recordings obtained at the start and the end of trials in the different centres when the patients were using: their conventional hearing aids, the SiVo aid, and no aid at all. Special reference is made here for three patients, to the influence of these different conditions on: voice quality and vibratory regularity, and on the control of intensity and overall timing. In parallel with the speech based measurements, psycho-acoustic tests to assess temporal discrimination and frequency acuity were made in addition to standard pure tone based audiometry; these and speech receptive assessment results for the whole project are discussed in greater depth separately (Faulkner et al, these Proceedings).

For some of the production measurements major improvements associated with the use of the SiVo aid were found which were reversible even in a single recording session simply by changing back to a conventional hearing aid. An especially striking example of the influence of SiVo auditory monitoring on the speaker's larynx frequency range, Fx, is given below.

SiVo aid **Conventional hearing aid**



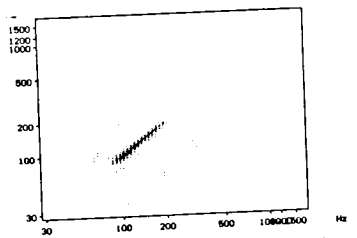
Larynx Frequency Distributions for MR

SiVo

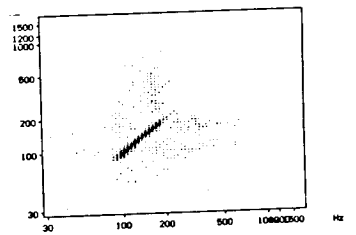
Larynx period detection and sine wave presentation at the most comfortable level for the user at all larynx frequencies, using in the ear audiometry and memory storage are provided within the SiVo aid itself. The recordings were made in the quiet and, in consequence, there was no benefit from the noise resistant features of this analysis.

C H A

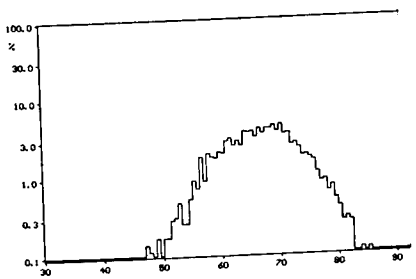
The Conventional Hearing Aid used by each patient in the study provided the best output possible for the individual user. In all cases in the work, the conventional hearing aid was far more familiar in use and sound quality, and equal amounts of training effort were devoted to both aids. All the results are a function of the aid used in the session.



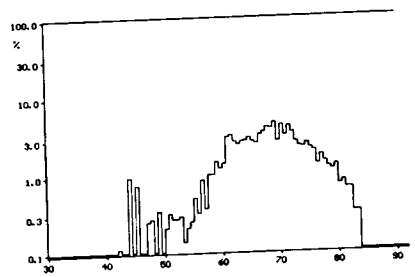
Larynx frequency crossplot using SiVo



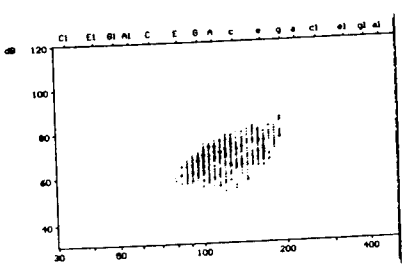
Larynx frequency crossplot with CHA



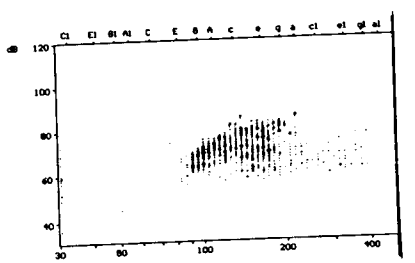
Speech Amplitude Distributions using:--



SiVo and CHA self monitoring



Sivo based second order Phonetogram



CHA based second order Phonetogram

PHYSICAL ANALYSES

The larynx frequency analyses shown on the first page, have been derived from period by period measurements. They are based on the synchronous speech and electro-laryngograph recordings which were made routinely during the clinical sessions. Both spontaneous and read passages were used, of at least two minutes duration. Only the read passages have been used here since this has made it possible for the speaker to aim for the same prosodic structures in the two monitoring conditions - SiVo and CHA - and for detailed segmental waveform comparisons to be made in subsequent analysis.

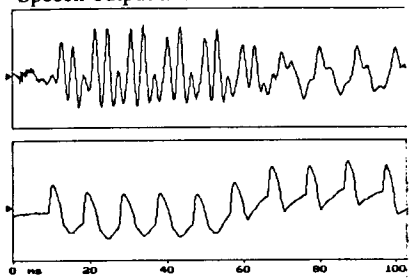
Speech excitation analyses are often based on windowed averages of more than one cycle of laryngeal, vocal fold, vibration. Larynx period, excitation epoch, based measurements for larynx frequency, Fx, give a more detailed foundation for the examination of speech waveforms. They also correspond to important aspects of many perceptually significant speech events. For example, the cycle-to-cycle period irregularities and breathy consonantal onsets of normal speech require this level of description for their detection and understanding; and in speech pathology temporal detail is useful in both assessment and training. In audiology, the ability of the hearing mechanism to base the sensation of pitch on the timing of acoustic events gives normal listeners the ability to hear creaky voice contrasts. For the very profoundly hearing impaired, the lack of peripheral frequency selectivity leaves essentially only temporal processing as the basis for their perception of voice pitch.

This is the case for the French patient MR, who lost his hearing (encephalitis after the essential stages of speech and language acquisition at 5y: 125Hz,65dB; 250Hz,80dB; 500Hz,105dB). For him, the conventional hearing aid can often give ill defined voice pitch periodicity information as a result of the varying harmonic structure of the acoustic input. SiVo analysis is designed to overcome this difficulty by the use of vocal fold closure detection training in the definition of its internal analysis algorithm and the provision of a sine wave acoustic output which, since it has only one harmonic component, does not change the nature of peripheral auditory temporal resonse with changes in input voice pitch.

The rather gross differences between the Fx distributions on the first page is an evident result of the lack of precise monitoring control afforded to MR by the conventional aid. The further analyses of his spoken outputs in the two monitoring conditions (for exactly the same recordings) give the basis for a more detailed understanding of what is happening. The larynx frequency crossplots simply show the distribution of successive pairs of Fx values, derived from the detection of successive epochs of excitation, throughout the whole speech sample. It is evident that it is not only the range of vocal fold frequencies which has been disturbed by the use of the CHA for monitoring but more importantly the temporal organisation at the quite detailed level of period to period closure. The speech amplitude distributions (probability plotted against dB) are based on the vocal fold synchronous determination of peak amplitude in each Fx period. They are remarkable for their similarity - apart from the low amplitude differences due to CHA induced creak. Overall loudness monitoring has not been changed by the switch from one aid to another. This is to be expected if it is essentially pitch perception which is not adequately supported by the conventional hearing aid.

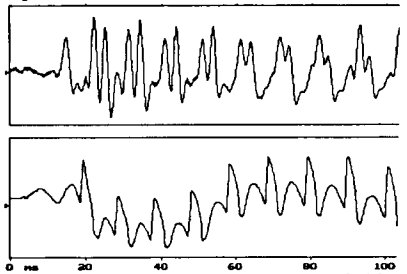
The final figures on the preceding page link the physical correlates of pitch and loudness through a development of ordinary phonetogram analysis. Here, once more, vocal fold closure detection has made it possible to obtain linked "instantaneous" measurements of Fx and excitation amplitude. In addition, the phonetogram analysis has only taken note of those pairs of successive vocal fold vibrations which have fallen into the same, quarter tone, analysis bin. In this way, the stable core of phonatory activity is shown and minor irregularities are eliminated. The SiVo based phonetogram is normal in shape for both frequency and amplitude. The CHA phonetogram is disorganised in the joint amplitude - frequency occurrences.

Speech output and Lx control with SiVo



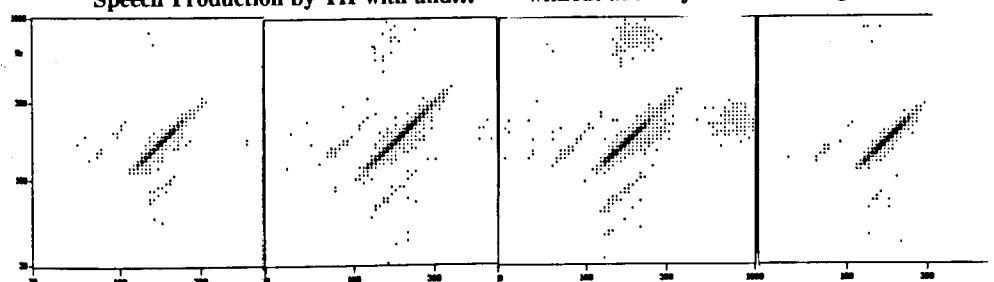
The use of the same read passage for all monitoring conditions has made quite detailed waveform level comparisons possible. The pair of waveforms above show the plots for speech, Sp above, and the laryngograph signal, Lx. The use of the SiVo aid to monitor his speech activity has not only enabled MR to control the broad levels of his phonation but also to produce essentially normal vocal fold closure sequences throughout his read passage. The sample shown is at the beginning of the word "Séguin". (The Sp & Lx waveforms are as initially recorded without time alignment. Note the correspondence in Lx baselines).

Speech output and Lx control with CHA



MR has made many recordings of this passage during the course of the work and he has got into the habit of using the same patterns of prosodic control. This makes it possible to understand what would otherwise be an astonishing difference between the two sets of waveforms above. The conventional aid is not able to provide him with an adequate sensation of voice pitch and in order parially to overcome this lack, he has a tendency to produce a breathy excitation spectrum - which has proportionately more energy at the low end of the spectrum. The Sp & Lx waveforms show this fairly clearly.

Speech Production by TH with and... without auditory feedback using SiVo



A complementary aspect of self monitoring speech with different methods of acoustic feedback comes from the comparison of the effect of the use of the prosthesis with no provision at all. The set of Fx crossplots shown above has been made from a sequence of recordings in which the (UK) user of a SiVo aid has, going from left to right: first used the aid to assist in reading a standard passage; then immediately after read again with no aid; then after an hour read again with no aid; and finally used the SiVo aid again for monitoring. At the beginning and end, using the SiVo for monitoring, the means of the Fx distributions are identical as are also the irregularity measures which indicate the departure of closure periods from the diagonal. In the absence of auditory control, there is both more variability between productions and more variability within. Mean & modal Fx values are different and, more obviously above, irregularity measures differ. This is not the case necessarily with the use of a CHA since here the same errors of control are repeatable.

We are grateful to our other colleagues; & to Laryngograph Ltd for its analyses.

McGURK EFFECT IN GERMAN AND HUNGARIAN LISTENERS

H. Grassegger
Institute of Linguistics, Section of Phonetics, Graz, Austria

ABSTRACT

The goal of this study is to determine how bimodal speech with conflicting auditory and visual information is processed by German (more exactly: Austrian) and Hungarian subjects. This was tested by bimodal presentation of the syllables /ba/, /da/, /ga/, /pa/, /ta/, /ka/, /ma/, /na/, /fa/, /sa/. The results, analysed by confusion matrices for each visual stimulus, showed that the McGurk was less strong and widespread in German than in Hungarian.

INTRODUCTION

The well-known McGurk effect phenomenon demonstrates that visual information on place of articulation influences phonetic perception. Unlike normal audio-visual congruent information which helps auditory perception, lip-read information with audio-visual discrepancy on place of articulation (i.e. whether the place is labial or non-labial) misleads and biases auditory perception.

Although this visual biasing effect on speech perception has been replicated in many studies for English speaking subjects, it has hardly been examined for other languages, with a few exceptions, amongst these Japanese [1], Spanish [2] and a single study with German speaking subjects who identified English bimodal CV syllables [3].

In the present study German (more exactly: Austrian) and Hungarian subjects were tested as to the perceptual influence of bimodal speech with conflicting auditory and visual information. As the phonological inventories of these two languages differ with regard to the consonant categories used in the test syllables (see below) the outcome of bimodal speech perception was expected to be influenced by these differences as well.

TEST SYLLABLES

Ten syllables with consonants occurring in both languages were used: /ba/, /da/, /ga/, /pa/, /ta/, /ka/, /ma/, /na/, /fa/, /sa/. For recording the audio and video signals a male Austrian talker pronounced each syllable once while his face was videotaped with a camera located in front of him and the audio signal recorded seperately to provide highest quality recording. For presentation a random order sequence of 100 audio-visual stimuli was produced resulting from the combination of the 10 audio stimuli dubbed on each visual stimulus. To ensure precise synchronization between audio and video signals the dubbing timing was adjusted by a 25msec frame unit. Each audio-visual stimulus was embedded in a 7sec unit consisting of a 4sec talking face preceded by a 3sec black screen with the respective stimulus number inserted. With a viewing distance of about 1 m visual stimuli were presented on a color monitor showing the speaker's face in approximately life-size. Audio stimuli where presented through the two built-in loudspeakers at each side of the screen.

SUBJECTS

Ten native speakers of German and Hungarian participated in the experiment. All subjects had normal hearing and normal or corrected visions. The age of the German subjects ranged from 20 to 25, the age of the Hungarian subjects from 14 to 16.

PROCEDURE

Subjects were presented an audio-visual stimulus every 7 seconds and were asked to look at and to listen to each utterance. The subjects' task was to write down what they *heard* not what they saw. To make the subjects attend to the visual stimuli they were instructed to report any noticed perceptual discrepancy. It was also suggested that some people might hear syllables not existing in their mother tongue's phonological system, like /bga/ or /pta/.

As the tests for German and Hungarian subjects were carried out in Graz and Budapest respectively some equipment differences in the presentation of the stimuli was unavoidable. Care was taken to make possibly affecting factors consistent.

EXPERIMENTAL DESIGN

The subjects were required to follow the above instructions in five repetitions of trials for the audio-visual condition, thus yielding 50 observations (10 subjects x 5 repetitions) for each AV-stimulus in German and Hungarian.

To measure *auditory* intelligibility five repetitions of the ten audio stimuli were similarily randomized and administered to the subjects only once, thus yielding also in 50 observations (10 subjects x 5 items) for each audio-stimulus in both languages. To avoid the influence of hearing the audio-alone stimuli on the McGurk effect this test was done after the audio-visual session.

RESULTS

The results were analyzed by producing confusion matrices for each visual stimulus and one confusion matrix for the audio-alone condition.

Audio-alone condition

In the audio-alone task almost all of the auditory stimuli were identified as what the speaker intended to pronounce, by the German as well as by the Hungarian listeners. Some minor deviations in the identification of stimuli, amongst these /m/ twice heard as /n/ by Hungarian subjects and three times by German subjects, were not able to explain respective fused responses in the audio-visual condition as a consequence of reduced auditory intelligibility. This was confirmed by performing a chi-square test that compared the frequencies of the fused responses in both conditions (df=1, N=100).

There was only one exception to the almost perfect intelligibility of the auditory stimuli: /b/, which yielded 46% /v/-responses with German and 38% with Hungarian listeners. The chi-square test for the respective responses in the audio-visual condition consequently revealed most of the deviant /v/-responses for /b/ as not significantly different from the audio-alone results and thus not visually biased.

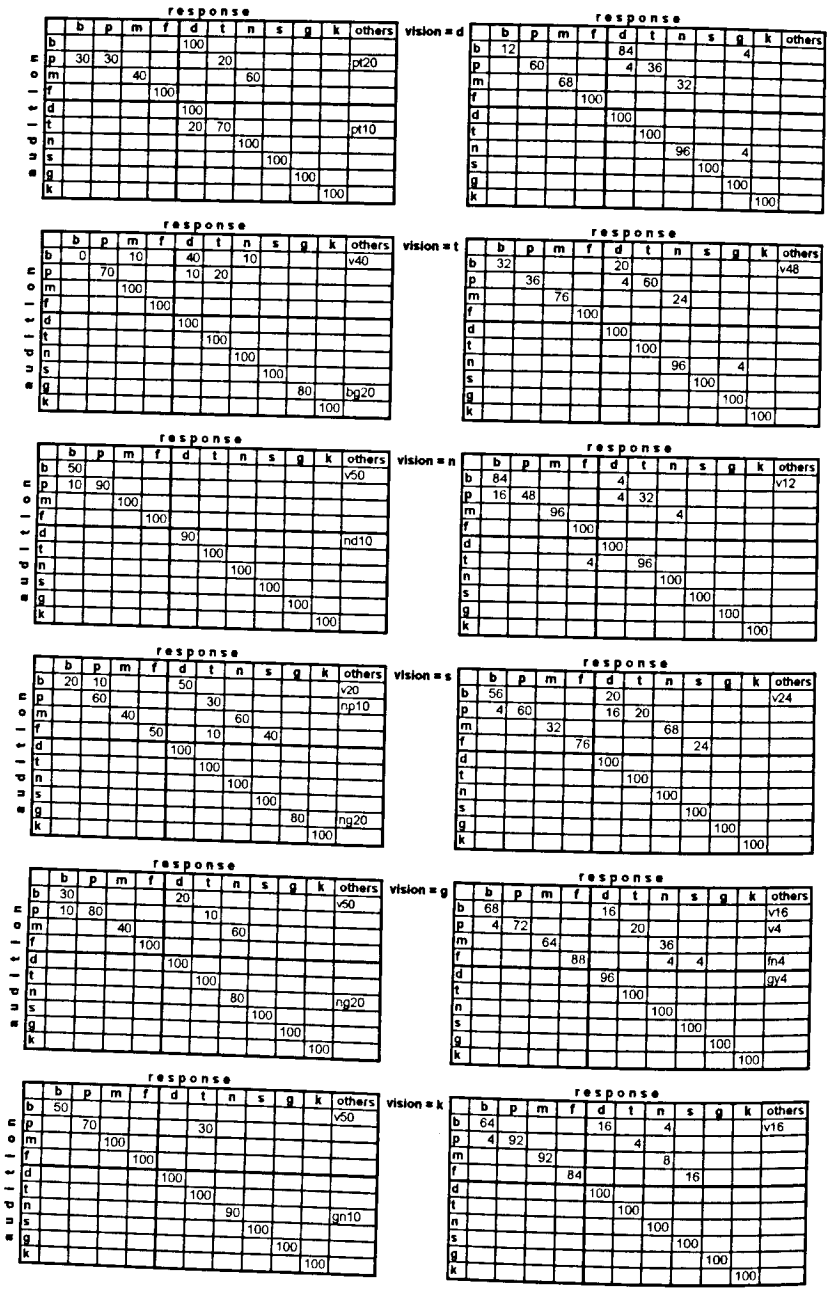
Visual labials

For visually presented labials, i.e. visual /b, p, m, f/, the confusion matrices for both languages show high rates in the diagonal cells, indicating that most of the auditory stimuli were perceived correctly and visual biasing effects were fairly weak. There are only two exceptions.

The first one is auditory /b/, which - evidently due to the above mentioned poor intelligibility - even with visual labials was most frequently heard as /v/, more so by German than by Hungarian listeners. Visual /f/ most effectively supports the obviously inherent labio-dental information of the intended auditory /b/: with visual /f/ Hungarian listeners judged /b/ only 40% of the time as /b/, 4% as /p/ and 56% (!) as /v/; in the same visual condition German listeners never (!) recognized auditory /b/, fused responses being /m/ with 10% and /v/ with 90% (in this latter case significantly different from the audio-alone condition, thus showing high visual biasing effect).

The second exception is auditory /n/, which yielded (its complete auditory

Table 1. Confusion matrices for the stimuli with visual non-labials, indicated in % in 50 observations for German (left column) and Hungarian (right column).



the time with visual /b/ and 60% (!) with visual /m/ for German subjects, 16% with visual /p/ and 12% with visual /m/ for Hungarian subjects.

Visual non-labials

For the visual non-labials, i.e. visual /d, t, n, s, g, k/, it was mainly with the auditory labials that visual effects occurred. This can easily be interpreted from the confusion matrices in Table 1, where in the diagonal cells of the lower right section (i.e. the *non*-labial section) 100%-values indicating absence of visual biasing effects predominate.

For auditory labials the influence of visual non-labials is least prominent with /f/. Fused responses in both languages only occur, when auditory /f/ is combined with visual /s/, yielding erroneous /s/-responses (40% with German, 24% with Hungarian subjects) or /t/-responses (10% with German subjects). With Hungarian subjects the perception of /f/ is - to a lesser degree - also biased by visual /g/ and /k/.

Visual biasing effects for auditory /b/ are evidently enhanced by its rather poor intelligibility (see above), but nevertheless it is noteworthy that combined with visual non-labials there is a significant amount of erroneous /d/-responses (even 100% with visual /d/ for German subjects) in both languages. The same holds true for auditory /p/, which in spite of its full intelligibility in the audio-alone test shows a remarkable frequency of /t/-responses.

The perception of auditory /m/ is visually biased by visual /d/, /s/ and /g/ in both languages, whereas with visual /t/, /n/ and /k/ visual biasing effects only occur with Hungarian subjects. In all cases, however, the erroneous responses are restricted to /n/, thus retaining the auditory information on the manner of articulation.

As already mentioned above the outcome of bimodal speech perception was expected to be influenced by the difference of the phonological inventories of the two languages tested. As Hungarian with a voiced and voiceless palatal plosive and with a palatal nasal has - as far as the categories of the test syllables are concerned - a more complete series of consonants than was offered by the test stimuli, it is most striking that Hungarian subjects only once produced a fused response *gy* (which orthographically stands for the voiced palatal plosive) 4% of the time when auditory /d/ was combined with visual /g/ (see "others" in the response column).

CONCLUSION

This study showed that the McGurk effect occurs in both languages investigated, slightly more easily to induce in Hungarian than in German. However, the visual biasing effects seem to be not symmetrically distributed for labial and non-labial articulation. In either language visual labials do not highly influence the perception of auditory non-labials except for the (dental) nasal, whereas visual non-labials produce a fairly strong visual biasing effect on auditory labial stimuli.

Comparison of the results of the audio-visual condition with the audio-alone condition indicated that the McGurk effect was more easily induced with poorer intelligibility (as in our case for /b/), but was not eliminable for stimuli of 100%-auditory intelligibility.

REFERENCES

[1] Sekiyama, K. & Tohkura, Y. (1993), "Inter-language differences in the influence of visual cues in speech perception", *Journal of Phonetics* 21, pp. 427-444.

[2] Massaro, D. et al. (1993), "Bimodal speech perception: an examination across languages", *Journal of Phonetics* 21, pp. 445-478.

[3] Mills, A. E. & Theim, R. (1980), "Auditory-visual fusions and illusions in speech perception", *Linguistische Berichte* 68, pp. 85-108

THE MCGURK EFFECT IN JAPANESE AND AMERICAN PERCEIVERS

K. Sekiyama, L. D. Braida**, K. Nishino*, M. Hayashi***, and M. M. Tuyo***
** Kanazawa University, Kanazawa, Japan*
*** Massachusetts Institute of Technology, Cambridge, MA, USA*
**** Osaka City University, Osaka, Japan*

ABSTRACT

This study examined our previous finding that Japanese perceivers are less subject to the McGurk effect than Americans. Stimuli were created using eight speakers. Although the results replicated the group difference, Japanese subjects showed a strong McGurk effect in some cases. The strong visual effect in the Japanese was related to auditory ambiguity and the visual robustness of the stimuli whereas the McGurk effect in Americans was stable in various cases.

PURPOSE

The McGurk effect demonstrates that visual lip movements influence perception of auditory speech even when the two sources of information are in conflict [1]. For example, when auditory /ma/ is dubbed onto visual lip movements of /ta/, this auditory-visual speech will be often perceived as "na." Whereas we found this effect to be robust for American perceivers, it was much weaker for Japanese perceivers when stimuli were Japanese syllables [2, 3]. This finding suggests that linguistic and/or cultural framework affects the manner of integration. This study examined this inter-language difference further with new stimuli, because our previous finding was based on stimuli from only one Japanese speaker. Experiment 1 was to see if the group difference is replicated for various Japanese speakers. In Experiment 2, the examination was done with more forceful visual stimuli of both Japanese and American speakers.

METHOD

Subjects

The subjects were native speakers of Japanese and American English with normal hearing and normal or corrected vision. In Experiment 1, the subjects were 24 students at Kanazawa University and 24 students at Massachusetts Institute of Technology (MIT). In Experiment 2, the subjects were 16 students at Osaka City University and 16 newly recruited students at MIT. All the subjects were age under 30 and had no experience of living in a foreign country.

Stimuli

Eight syllables were used: /ba/, /pa/, /ma/, /da/, /ta/, /na/, /ga/, and /ka/.

In Experiment 1, the speakers were four native speakers of Japanese. Recorded sound and videotaped lip movements from one speaker were combined and sixteen pairs were created so that each syllable had an auditory-visual discrepant pair (as shown in Table 1) as well as an auditory-visual identical one (audio /pa/, video /pa/).

Table 1. Combinations of auditory (A) and visual (V) syllables in auditory-visual discrepant pairs. R shows a typical response when the A and V syllables are perceptually integrated.

	A	V	R
Labial	b	g	d
	p	k	t
	m	n	n
Nonlabial	d	b	bd
	t	p	pt
	n	m	mn
	g	b	bg
	k	p	pk

In Experiment 2, among 30 who were videotaped, we chose two Japanese and two American speakers whose utterances were the clearest to lipread. The auditory-visual stimuli were created using the same syllable combinations as in Experiment 1. Considering that the duration of Japanese vowels is much shorter than that of English ones, the Japanese speakers were instructed to pronounce vowels longer than usual so that their duration is comparable with those of English stimuli. This resulted in slower articulations than usual, which seemed to make the visual stimuli easier to lipread.

Procedure

In both experiments, the stimuli were presented in three conditions: auditory-visual (AV), visual (V), and auditory (A) conditions. The subjects were asked to write what they heard (AV and A conditions), or what they thought the speaker was saying (V condition), as an open choice response.

RESULTS

Figure 1 shows the magnitude of the McGurk effect (visual effect) for each speaker. The visual effect here refers to the influence of discrepant visual cues. It

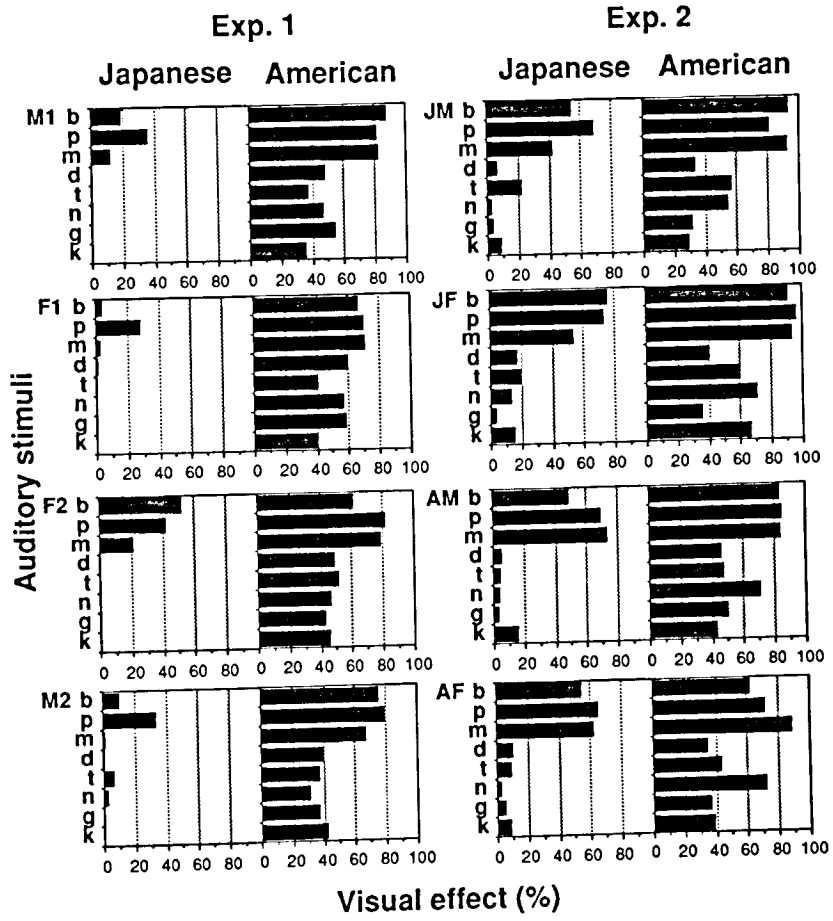


Figure 1. The magnitude of the McGurk (visual) effect in the Japanese and American subjects for each auditory syllable of each speaker. In Experiment 1, all the speakers, were Japanese, two males (M1, M2), and two females (F1, F2). In Experiment 2, there were wo Japanese (JM: male, JF: female) and two American (AM: male, AF: female) speakers. The results are shown only for auditory-visual discrepant pairs.

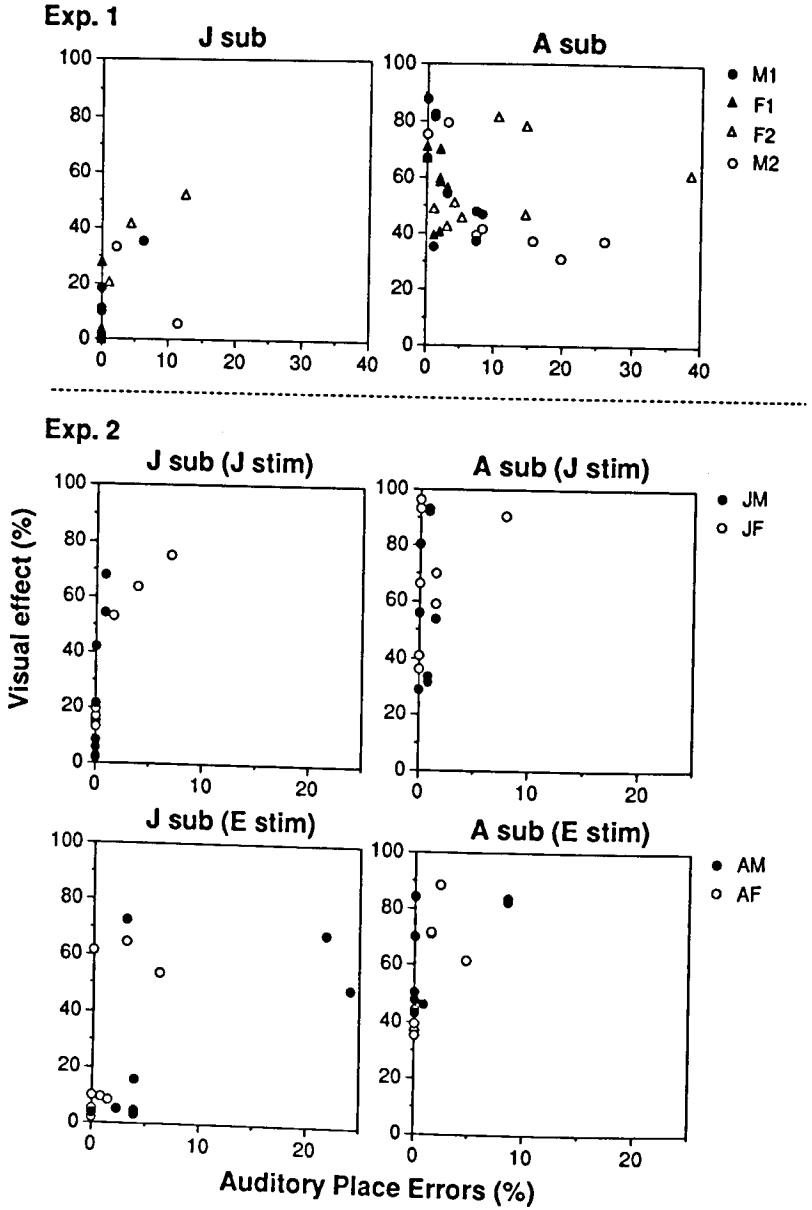


Figure 2. The relationship between the magnitude of the McGurk effect (visual effect) and auditory ambiguity (auditory place errors, i.e., errors in terms of place of articulation when presented with no visual cues). Each point corresponds to one of the eight auditory syllables pronounced by one speaker.

was defined as percent of auditory errors in terms of *place of articulation* (labial vs. nonlabial) caused by discrepant visual cues. For example, if auditory /ba/ is perceived as "da" in 60 % of the cases when combined with visual /ga/ and this "da" response occurs in 5% fo the cases when auditory /ba/ is presented with no visual cues, then the visual effect is 60 - 5 = 55%.

The results of Experiment 1 replicated our previous finding, showing a much weaker McGurk effect for the Japanese subjects than for the American subjects. Agreeing with our previous results, the McGurk effect was stronger when auditory stimuli were labial.

In Experiment 2, although the group difference was replicated, the Japanese subjects showed a stronger McGurk effect than before for auditory labials, together with large individual differences. Unlike our previous study [3], there were no difference between the Japanese and English stimuli.

DISCUSSION

Effect of speaking rate

Although the two experiments replicated the result that the McGurk effect is stronger in the American subjects than in the Japanese subjects, the McGurk effect for the Japanese subjects perceiving Japanese speech was stronger in Experiment 2 than in Experiment 1 and our previous study. We believe this difference to be due to the forcefulness of visual stimuli in Experiment 2, where we choose speakers easy to lipread as well as instructed the Japanese speakers to *lengthen* the vowels. We did not instruct the Japanese speakers to do so in Experiment 1 or our earlier study [3]. We instructed the American speaker in our earlier study to *shorten* the vowels. These facts suggest that the slower articulations in Experiment 2 led to more forceful visual stimuli, which increased the magnitude of the visual effect for the Japanese subjects.

Effect of auditory ambiguity

Figure 2 shows the relationship between the magnitude of the McGurk effect and auditory ambiguity. We hypothesized that if an auditory token has ambiguous quality, it is very susceptible to the visual effect, whereas an auditory token of unequivocal quality will not be influenced by visual cues so easily. In Figure 2, auditory ambiguity is indicated by auditory place errors.

The results supported this auditory ambiguity hypothesis only for the Japanese subjects perceiving Japanese speech. Look at the upper two panels in the left column. In these cases, it is found that the magnitude of the visual effect is limited when the percent of auditory place errors is zero: The visual effect was less than 30% in Experiment 1 and less than 40% in Experiment 2. When there are some auditory place errors, the magnitude of the visual effect is an increasing function of the ambiguity. Thus, if the percent of auditory place errors could be smaller than zero, then the two indexes might show a linear correlation.

In contrast, the results for the American subjects show strong visual effects even when the percent of auditory place errors was zero.

These results show that the magnitude of the McGurk effect in the Japanese subjects tends to vary depending on quality of auditory and visual speech, while the McGurk effect in American subjects are stable under various conditions.

ACKNOWLEDGMENT

This study was supported by grants from HFSP, NIH, and the Japanese Ministry of Education.

REFERENCES

[1] McGurk, H. and MacDonald, J. (1976) "Hearing lips and seeing voices," Nature, Vol. 264, Pp. 746-748.

[2] Sekiyama, K. and Tohkura, Y. (1991) "McGurk effect in non-English listeners," Journal of the Acoustical Society of America, Vol. 90, Pp. 1797-1805.

[3] Sekiyama, K. (1994) "Differences in auditory-visual speech perception between Japanese and Americans," Journal of the Acoustical Society of Japan (E), Vol. 15, Pp. 143-158.

CONFIGURATIONAL vs. TEMPORAL COHERENCE IN AUDIOVISUAL SPEECH PERCEPTION

M.-A. Cathiard, M.T. Lallouache°, T. Mohamadi° and C. Abry°*
**Laboratoire de Psychologie Expérimentale, Université Mendès-France, BP 47 F-38 040 Grenoble Cedex 9 /°ICP URA CNRS n°368 INPG/ENSERG Université Stendhal BP 25*

ABSTRACT

This contribution evaluates the natural *coherence* of the building up of audiovisual information in the flow of speech that is provided on the face. Our conclusion is that the bimodal coherence of speech is above all configurational, and not essentially temporal, as it is the case in another intersensory system, such as auditory-visual localization in the ventriloquist effect [1, 2].

1. INTRODUCTION

The aim of this contribution is to evaluate the natural coherence of the building up of visual *and* audio information in the flow of speech, using a gating and desynchronization procedure.

It has been previously shown that, when optic information is naturally in advance on acoustics – as in the case of a visible gesture like rounding –, featural/gestural anticipation can be identified only by eye several tens of milliseconds before any perceivable sound [3].

This bimodal information being in a sense naturally "desynchronized", an obvious experimental manipulation was to reduce the delay of audition on vision, i.e. to "resynchronize" the audio signal in order to test the boundaries of such a bimodal temporal organization.

2. STIMULI

We used [i#y] and [i#i] (control) vocalic transitions embedded in a carrier sentence of the type: "T'as dit : UHU use?" ("Did you say : UHU ["Indian" name] wear out?"). Indian names were used in order to maximize pure vowel-to-vowel modulation of the output of the vocal-tract, without intervening consonantal gestures. A French male talker was filmed at 25 frames/second. Stimuli were chosen with a 160 ms acoustic pause between vowels. For each digitized frame, articulatory parameters were automatically extracted by image processing [4].

For [i#y] transitions, upper lip protrusion P1 starts at the end of [i] (Fig. 1), together with the constriction, i.e. lip area S begins to decrease. The domain we determined, for these two main components of the rounding gesture, stretches from 4 images before the pause, in order to allow a sufficient range in desynchronization, to 1 image after, when *both* components have reached their maximum.

3. VISUAL TEST

The visual information was explored by gating (40 ms steps): 10 gates of 1000 ms duration allowed to display in all cases the beginning of the carrier sentence "T'as dit...". This visual [i]/[y] identification test (with 10 repetitions for each gate in random order) was performed by 10 naive French subjects, with no deficit in vision and audition. The visual boundary in the [i#y] transition was measured on the mean curve of all subjects using Probit analysis.

This boundary takes place 140 ms before the acoustic onset of [y], and less than 40 ms (a one image step) are enough to switch from 80% [i] to 80 % [y] (Fig. 1). Anticipation is earlier (140 *vs.* 100) and category switching is steeper (40 *vs.* 80) than the one we obtained with another 160 ms paused signal, whose articulatory profile was actually slower, especially in the constriction building up [5]. Anyway, this 140 ms anticipation remains shorter than the maximum case we ever evidenced, in fact within a very long 460 ms pause: 210 ms [5].

Thus we confirm our previous findings on the natural advance of the eye relative to the ear.

4. AUDIO TEST

So what about the building up of audio information ?

An audio [i]/[y] identification test, including the beginning of the stimuli "T'as dit...", stopped 2, 6 and 10 ms after the onset of [#y] and [#i], was performed by the same 10 subjects (10 repetitions by gate). Such a range from 2 to 10 ms has proven to be sufficient to scan properly the building up of featural acoustic information. Mean identification scores were: 58%, 95%, 99% for [y] (Fig. 2); and 100%, 99% et 99% for [i]. This supports the claim that often only one pulse (8 ms in the case of the vowel onsets of our talker) is sufficient to fairly identify the vowel (for French, see [6]).

The building up of visual information (40-80 ms) is thus slower than the audio one (10 ms). But this is fairly compensated by the visible anticipation on the sound, which is naturally displayed in speech (up to 200 ms) due to such a pervasive phenomenon as coarticulation.

5. AUDIOVISUAL TEST

But what are the audio/visual boundaries of this bimodal coherence ?

The same visual stimuli were presented with the sound in synchrony or in advance. For each gate, for which we measured the time course of visual information, we tested the building up of the audio using the 3 steps previously determined (2, 6 et 10 ms), in order to obtain a desynchronization range from 0 to -360 ms, by 40 ms steps. The same 10 subjects where tested on the 10 steps for the 3 vowel onset durations.

Individual curves obtained for each acoustic duration show clearly different patterns. Since averaging was not representative, we grouped them according to similarity of their response profiles. Individual curves are either clearly S-shaped, or they show a first phase, before the visual [i#y] boundary, which is less regular and/or close to chance level (Figs. 3a-f). On the base of the scores for 10 ms vowel onsets (corresponding to high audio performances), we obtained two groups of 5 subjects. The first group (Figs. 3a-c) has, in the phase before the visual boundary, identification [y] scores below 20%; the second group having scores above 20% (Figs. 3d-f).

If we first consider scores in the synchronous condition (plotted on gate n°10) for both groups, we see that, independently of vowel onset duration, individual [y] scores are generally at or above 90% (with one exception). Mean scores for 2, 6 et 10 ms durations are respectively: 96%, 98% et 99%. Comparing audiovisual results obtained in the worst condition, 2 ms (96%), with the audio alone condition (58%), vision benefit reveals largely sufficient to disambiguate a poor audio signal. We thus rejoin results in a more classical condition, namely speech in noise (for French, see [7]).

When desynchronization occurs, for these 2 ms vowels, we see that rounding information – an anticipating one in the original – can bring to them a visual benefit up to -160 ms. One must recall that for this value, i.e. up to image n°6 (see Fig. 1), visual information alone reached 85% [y] responses, whereas just 40 ms before it scored 12%. In other terms, we were able to test step by step what phase of the anticipatory gesture could enhance ambiguous audio information. It comprises in fact all the phase "sheltered" by the gesture: *after the visual boundary.*

Let's consider now desynchronization effect beyond this visual boundary, i. e. for the phase corresponding articulatorily to an [i]. For Group 1, we see that the duration of vowel [y] onsets – comparatively to identification scores in the audio condition – does not seem to influence subjects' behaviour. In fact, what is properly characteristic of this group is its high sensitivity to visible articulatory information. The curves we obtain for the three conditions display a clear S-shape, which looks strongly like the ones (mean and individual) obtained for vision alone: the identification boundary is located, for the three audiovisual conditions in the vicinity of the visual boundary. This similarity of the curves in the visual and audiovisual conditions indicates that, when desynchronization delivers images in advance of the sound – in this case an articulatory information specific of an [i] (in a desynchronization ranging from -200 to -360 ms, for this transition) –, then subjects identify the oncoming of an [i] vowel, in spite of the fact that they receive an audio information largely sufficient to recognize an [y]. Things are going on as if in the case of conflicting information – [i] being visible et [y] audible –, visible information was guiding perception.

Subjects from Group 2 are sensitive also to conflicting information. Whereas audio

is clearly identified by them as [y] – at least for 6 and 10 ms durations –, a visible [i] pushes their scores towards chance level.

To summarize: (i) desynchronization has the largest effect only when the visual boundary is crossed; (ii) beyond this boundary, no subject is insensitive to visual information, i.e. clearly no subject displays a steady 100% [y] along all desynchronization values. Moreover in (ii) the proportion of those who answer [i] for audio [y] is very close to the results found in a rounding judgment task for the same conflicting French vowels [8]. However, up to the present experiment, no such "McGurk effect" had been successfully obtained by a desynchronization procedure for vowels [9].

6. CONCLUSION

The natural delay of audio, relative to the visual signal, in speech coarticulatory anticipation, can be reduced without affecting intelligibility, as long as the configurational visual cues are in accordance with the sound. This hypothesis of a primacy of configurational over temporal coherence could be used to explain other results on desynchronization (reviewed in [10, 3, 11]) for detection tasks [12] as well as for intelligibility ones [13, 14].

Acknowledgements: This work was supported in part by a grant in Cognitive Science from the French Ministry of Research and Technology, a France Telecom-Université Stendhal contract n°927B032 and an Esprit Basic Research project n°6975 *Speech Maps*.

REFERENCES

[1] Radeau, M. (1994). Auditory-visual spatial interaction and modularity. *Current Psychology of Cognition, 13(1)*, 3-51.

[2] Abry, C., Cathiard, M.-A., Robert-Ribès, J., & Schwartz, J.-L. (1994). The coherence of speech in audio-visual integration. *Current Psychology of Cognition, 13(1)*, 52-59.

[3] Cathiard, M.-A. (1994). *La perception visuelle de l'anticipation des gestes vocaliques : cohérence des événements audibles et visibles dans le flux de la parole.* Thèse de Psychologie Cognitive, Université Grenoble 2.

[4] Lallouache, M.-T. (1991). *Un poste "visage-parole" couleur. Acquisition et traitement automatique des contours des lèvres.* Thèse de l'ENSERG, Spécialité : Signal Image Parole, Grenoble.

[5] Cathiard, M.-A., & Lallouache, T. (1992). L'apport de la cinématique dans la perception visuelle de l'anticipation et de la rétention labiales. *Actes des 19èmes Journées d'Études sur la Parole*, Bruxelles, 19-22 Mai, 25-30.

[6] Serniclaes, W., & Wajskop, M. (1972). L'identification vocalique en fonction de la fréquence fondamentale et de la durée de présentation. *Revue de Phonétique appliquée, 22*, 39-50.

[7] Benoît, C., Mohamadi, T., & Kandel, S. (1994). Effects of phonetic context on audio-visual intelligibility of French. *Journal of Speech and Hearing Research, 37*, 1195-1203.

[8] Lisker, L. & Rossi, M. (1992). Auditory and visual cueing of the [± rounded] feature of vowels. *Language and Speech, 35(4)*, 391-417.

[9] Massaro, D.W., & Cohen, M.M. (1993). Perceiving asynchronous bimodal speech in consonant-vowel and vowel syllables. *Speech Communication, 13*, 127-134.

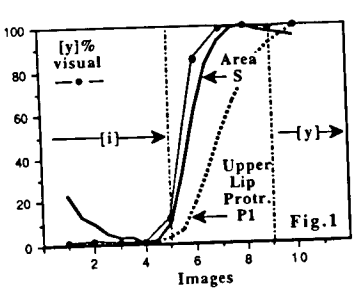
[10] Summerfield, Q. (1992). Lipreading and audio-visual speech perception. In V. Bruce, A. Cowey, A.W. Ellis & D.I. Perrett (Eds.), *Processing the facial image* (Proceedings of a Royal Society Discussion Meeting, 9-10 July), Clarendon Press, Oxford, pp. 71-78.

[11] Cathiard, M.-A., & Tiberghien, G. (1994). Le visage de la parole : une cohérence bimodale temporelle ou configurationnelle? *Psychologie Française, 39 (4)*, 357-374.

[12] McGrath, M., & Summerfield, Q. (1985). Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. *The Journal of the Acoustical Society of America, 77(2)*, 678-685.

[13] Smeele, P.M.T., & Sittig, A.C. (1991). The contribution of vision to speech perception. In *Proceedings of the 2nd European Conference on Speech Communication and Technology*, Genova, Italy, 24-26 September, vol. 3, 1495-1497.

[14] Smeele, P.M.T. (1994). *Perceiving speech : Integrating auditory and visual speech.* Doctoral dissertation, Delft University.



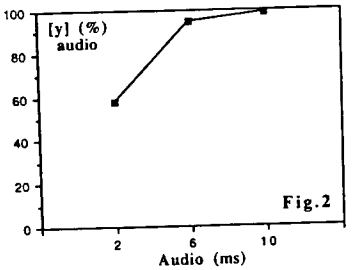
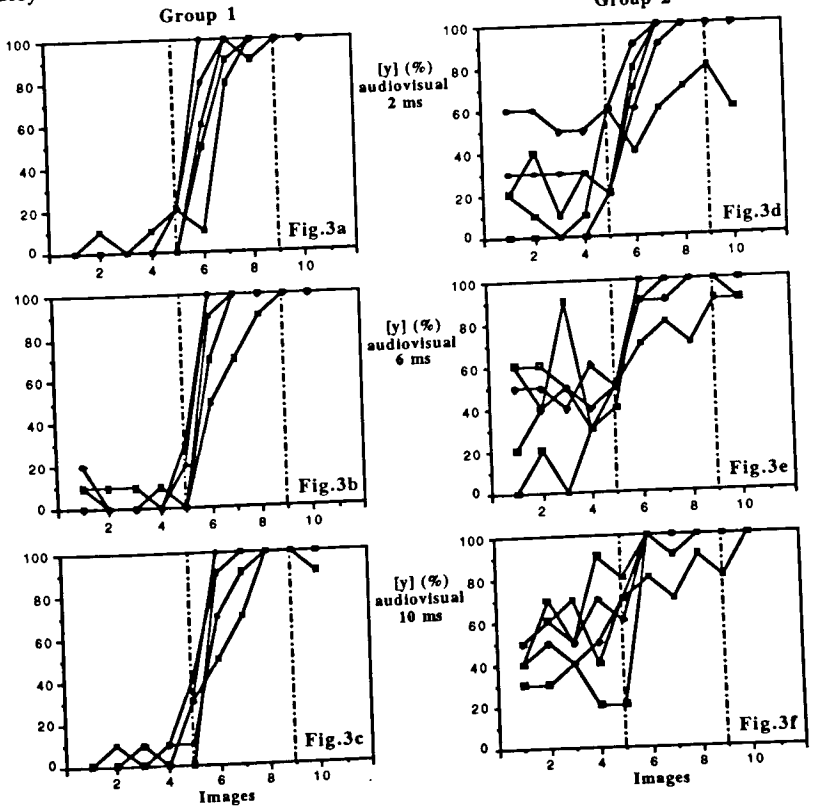


Fig.1.– Mean visual identification results for [tadi#y] superimposed on normalized articulatory parameters -S and +P1 (scores are plotted on the last image of the gate). **Fig.2.**– Mean auditory scores for [tadi#y] with short [y] onsets. **Figs.3a-f.**– Individual audiovisual scores for group 1 (left) and group 2 (right). N°10: scores in the synchronized condition with (from top to bottom) 2, 6 and 10 ms [y] onsets after image n°9. For scores on n°9 the sound is aligned on n° 8, i.e. with a -40ms desynchronization; etc.; with a maximum -360 ms on n°1.



SPEECH INTELLIGIBILITY OF SYNTHETIC LIPS AND JAW

T. Guiard-Marigny (1), C. Benoît (1) and D.J. Ostry (2)
(1) Institut de la Communication Parlée, INPG-Université Stendhal, Grenoble, France
(2) Psychology Department, McGill University, Montreal, Canada.

ABSTRACT

Audio-visual speech intelligibility was tested using high-quality 3D models of the lips and jaw. The models were animated on the basis of six parameters obtained from the actual movements of a speaker's face and synchronized with the original audio utterances. Eighteen French nonsense utterances were presented to twenty subjects at five levels of added noise. Intelligibility was best when the lip and jaw animations were presented along with the acoustic speech signal.

INTRODUCTION

Even though the auditory modality is dominant in speech perception, it has been shown that seeing the speaker's face increases intelligibility, especially in a background noise [12, 3, 4, 13, 2]. Synthetic faces are thus expected to enhance the intelligibility of speech synthesizers which is still far lower than that of humans.

It has been shown in English [9], and then in French [8] that the human lips carry more than a half of the visual information provided by the whole natural face. Moreover, vision of the teeth increases the intelligibility of a message: the teeth help disambiguate sounds differing in jaw position like "bib" versus. "bab" [9].

In this paper we evaluate through a perception test the contribution to speech intelligibility of a lip model alone and of the same lip model superimposed upon a synthetic jaw and upper skull.

THE 3D LIP MODEL

The 3D model of the lips used in this study was developed on the basis of a geometrical analysis of the natural lip movements of a French speaker [5]. The model is controlled with five parameters which can be measured directly from the recorded lip movements of a real speaker's face. A specially designed workstation [7] is used to obtain accurate measures of the parameters from a videotape. The measurement procedure produces an output file which contains the five parameters measured at 20 ms intervals; this file is used as a command file to our model. The digitized voice of the natural speaker is synchronized with the visual display.

THE JAW MODEL

Apart from the lips, the most visible articulator is the jaw and with it, the chin and the teeth. Since the jaw is a rigid skeletal structure, the animation process is easier than with the lips. Like all rigid objects, jaw motions have six degrees of freedom. Thus, its position relative to the skull can be defined with three orientation angles (yaw, pitch, roll) and three positions (horizontal, vertical, lateral).

The synthetic jaw which was used for our model was developed at McGill University [6] in order to visualize jaw motion kinematics that are recorded with an optoelectronic measurement system. It comprises a 3D digitized upper skull and jaw along with their corresponding teeth. The jaw model is animated using empirically recorded jaw orientation angles and jaw positions [6]. The visual display of the synthetic upper skull and jaw was synchronized with the corresponding natural audio signal.

ANIMATION OF THE MODELS

The lip and jaw models were integrated in a single display. The lip model was directly superimposed on the 3D skull and jaw. For tests of the model, lip movements were obtained using the video analysis technique described above. Jaw movements were obtained in a similar manner from the motion of the chin of the speaker using image processing techniques like those developed for lip movement. It should be noted that while it would have been desirable to use the optoelectronic measurement system at McGill to obtain jaw motions, this technique requires the use of an acrylic and metal dental

appliance which makes it difficult to measure lip movement.

Jaw motions in speech are controlled primarily in three degrees of freedom [10], namely the pitch angle, the vertical position and the horizontal position. The positions of two points on the jaw are sufficient to reconstruct these three motions in the sagittal plane. However, since the jaw is not directly visible and the overlying skin moves relative to the jaw, the points needed to reconstruct sagittal plane jaw motion cannot be obtained with non-intrusive methods. Nevertheless, it can be seen from the data reported in [11] that the basic parameters of jaw motion are often strongly correlated in running speech. To a first approximation, the three basic jaw motions can thus be predicted from the displacement of a single point on the jaw. Since the teeth are not always visible, we have decided to obtain this single point by tracking a dot on the chin. Of course, in so doing, a discrepancy cannot be avoided between the actual jaw motion and that of the reference point on the chin.

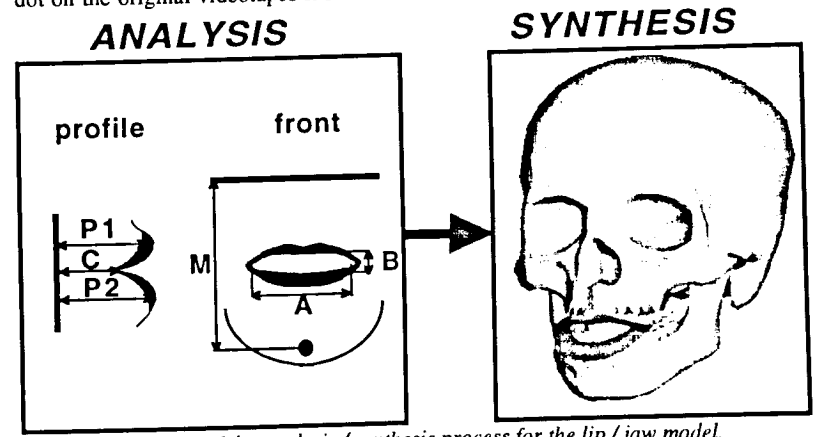
For purposes of our first tests of the lip / jaw synthesizer we have used an audio-visual corpus which has already been used extensively at ICP in order to make geometric measurements [1, 5] and to evaluate the contribution of vision to speech intelligibility [2, 8]. Since the speaker's chin was made up with a single dot on the original videotapes it seemed sufficient for the initial evaluation. A schematic of the analysis and synthesis process used to obtain the animation is presented in Figure 1.

INTELLIGIBILITY OF THE MODELS

Following two previous experiments [2, 8], the audio-visual intelligibility of the lip model and of the superimposed models of the lips and jaw (called the lip / jaw model) have been tested at five levels of acoustic degradation.

Preparation of the Stimuli

The speech material consisted of the natural acoustic utterances of a French speaker synchronized with three kinds of display: no video, synthetic lips, synthetic lips and jaw. The corpus consisted of VCVCV nonsense utterances. The vowels tested were /a/, /i/ and /y/. The consonants were /b/, /v/, /z/, /ʒ/, /ʁ/ or /l/. The test words were embedded in a carrier sentence of the form "C'est pas VCVCVz ?". Eighteen different sentences were first digitized and then acoustically degraded by addition of white noise, at five signal to noise levels, in 6 dB steps. Thus overall, there were 90 different acoustic stimuli. A pseudo-random order was used for presentation of the stimuli to subjects. Ten additional stimuli preceded the actual test in order to help subjects adapt to the test conditions.



Figue 1. Schematic of the analysis / synthesis process for the lip / jaw model.

The same sequence of acoustic stimuli was used in all three experimental conditions. The visual stimuli for the two synthetic models were recorded frame by frame on a videotape. The models were synthesized at a 25 ips rate with the virtual camera located at a 25° angle from the sagittal plane. The audio stimuli were subsequently synchronized with the visual display.

Twenty normal French listeners took part in the experiment. They were seated at a 1m distance from a 15" color monitor equiped with a loudspeaker. The order of presentation of the three sub-tests was balanced across the subjects. The subjects were required to identify both the vowel and the consonant in each utterance.

Global intelligibility

A test word was considered correct only if both the vowel and the consonant were correctly identified. The intelligibility scores obtained with the audio alone and with the lip model in this experiment were comparable to those reported in [8]. The data showed that the lip model restored approximately a third of the missing information when the acoustic signal was degraded. Moreover, we obtained a noticeable gain in speech intelligibility when the synthetic jaw was added to the synthetic lips, as shown on Figure 2.

Confusions

When the visual display of lip movement was added, the identification of /b/ was improved. However, /b/ was often given as the response to /v/ stimuli. The identification of the other consonants was also improved except for /ʒ/. For the vowels, /y/ was almost always correctly identified but /i/ and /a/ were still confused.

When synchronized with the lip model, the jaw model generally enhances intelligibility. The number of cases in which subjects were unable to respond at all was reduced by a factor of two. The vowel /i/ was confused less with the vowel /a/, mostly in consonantal contexts that close or tend to close the lips (/b/ or /v/). In addition, there were less confusions between /ʒ/ and /ʁ/, regardless of the vocalic context. Moreover, /b/ was no longer confused with /v/, especially in the context of the vowel /i/. The visibility of the teeth presumably accounts for this disambiguation.

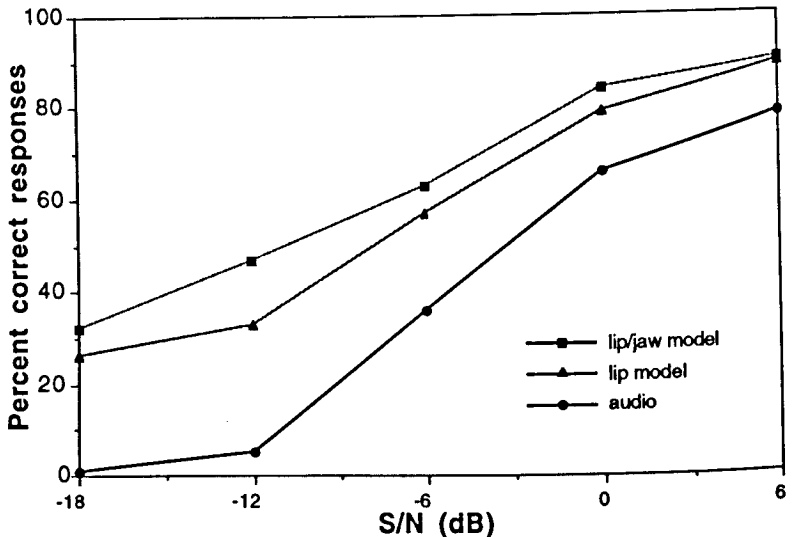


Figure 2. Audio-visual intelligibility of the lip model and of the lip / jaw model compared with the intelligibility of the auditory signal alone. The data are shown for various levels of noise.

On the other hand, with the combined lip and jaw models, /l/ and /v/ were more often confused with /ʒ/. However, this only occurred in rounded vocalic contexts , such as, /y/. Vision of the jaw also led to a greater number of confusions between /i/ and /a/ in a /z/ context. This was mostly due to the individual utterances /izizi/ and /azaza/ selected as stimuli for this experiment.

CONCLUSION

We obtained a noticeable gain in speech intelligibility when a synthetic jaw was displayed along with synthetic lips. However, compared to the intelligibility scores obtained in [8] with a synthetic face, the gain is small. This is likely due to the unnatural display of the lips superimposed on the skeletal skull and the jaw. Nevertheless, we can speculate that a semi-transparent skin overlaid on a display of the intrinsic articulators of the vocal tract may further improve intelligibility. The intelligibility scores in such an "augmented reality" of visible speech should be tested in the near future.

ACKNOWLEDGEMENT

This research was supported by the CNRS and by a grant from the ESPRIT-BRA progamme ("MIAMI" project No 8579), and by NIH Grant DC-00594 from the National Institute on Deafness and Other Communication Disorders.

REFERENCES

[1] Benoît, C., Lallouache, M.T., Mohamadi, T. & Abry, C. (1990), *A set of French visemes for visual speech synthesis*, Talking Machines, Bailly & Benoît, Eds, Elsevier B.V, Amsterdam, pp. 485-504.

[2] Benoît, C., Mohamadi, T. & Kandell, S.D. (1994), *Effects of phonetic context on audio-visual intelligibility in French*, Journal of Speech & Hearing Research, vol. 37, pp. 1195-1203.

[3] Erber, N.P. (1969), *Interaction of audition and vision in the recognition of oral speech stimulii*, Journal of Speech & Hearing Research, vol. 12, pp. 423-425.

[4] Erber, N.P. (1975), *Auditory-visual perception of speech*, Journal of Speech & Hearing Disorders, vol.40, pp. 481-492.

[5] Guiard-Marigny, T., Adjoudani, A. & Benoît, C.(1994), *A 3D model of the lips for visual speech synthesis*, Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis, New Paltz, New York, USA, pp. 49-52.

[6] Guiard-Marigny, T. & Ostry, D.J (1995), *Three-dimensional visualization of human jaw motion in speech*, 129th Meeting of the Acoustical Society of America, Washington, USA, to appear.

[7] Lallouache, M.T. (1991), *Un poste "visage-parole" couleur. Acquisition et traitement automatique des contours des lèvres*, Unpublished Thesis Dissertation INP, Grenoble, France, 214 pp.

[8] Le Goff, B, Guiard-Marigny, T., & Benoît, C. (1995), *Real-time analysis-synthesis and intelligibility of talking faces*, Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis, New Paltz, New York, USA, pp. 53-56.

[9] McGrath M. (1985), *An examination of cues for visual and audio-visual speech perception using natural and computer-generated faces*, Ph.D. Thesis, Univ. of Nottingham, UK.

[10] Ostry, D.J. & Bateson, E.V. (1994), *Jaw motions in speech are controlled in (at least) three degrees of freedom*, Proceedings of the International Conference on Spoken Language Processing, vol. 1, pp.41-44.

[11] Ostry, D.J. & Munhall, K.G. (1994), *Control of jaw orientation and position in mastication and speech*, Journal of Neurophysiology, vol 71, No 4.

[12] Sumby, W.H. & Pollack, I. (1954), *Visual contribution to speech intelligibility in noise*, Journal of the Acoustical Society of America, vol. 26, pp. 212-215.

[13] Summerfield, Q. (1979), *Use of visual information for phonetic perception*, Phonetica, vol. 36, pp. 314-331.

A LONG-SHORT VOWEL DICHOTOMY IN FLUENT ENGLISH?

Leigh Lisker and Arthur S. Abramson
Haskins Laboratories, New Haven, Connecticut, U. S. A.

ABSTRACT

American English vowels are usually put into two categories: long vs short or tense vs lax, the supporting data coming from non-spontaneous speech. Duration data from free conversation are consistent with this view, in that /ɪʊɛʌ/ are shorter than /ioeuaɔæ/. But /ɪʊɛʌ/ occupy a range of durations that is not in any sense discontinuous with the range of durations manifested by the remaining vowels of the language.

INTRODUCTION

The vowels of English have long been said to fall into two categories, involving a dimension [±long] and/or [±tense] and/or [±ATR] (advanced tongue root). There are both phonological and phonetic motivations for making this division. Among the latter is the view that in the particular pairs /ɪ-i/ /ʊ-u/ /ɛ-e/ a salient difference between the first and second vowels is durational, which might perhaps be "explained" by a difference in either tensity or in the position of the root of the tongue. However all the vowels of the language are usually assigned to one or the other category, even if not all those of one of the categories are unequivocally paired with particular members of the other. Of course, all the pairings proposed involve clear differences in timbre. Here we will put to one side any consideration of "tensity" or tongue-root position, focussing on length, an auditory attribute that ostensibly corresponds to the measurable duration of the "vocalic stretch" in an acoustic signal. Although the phonological literature does not show complete agreement on the membership of the two categories, there is considerable overlap among the various classifications that have been proposed. Thus everyone reports /ɪʊɛʌ/ as short or lax, while /iueo/ are long or tense. The status of /ɑɔæ/ is less clear. So for Goldsmith [1] /ɪɛæʌɑʊɔ/ are short, all others long. But others view /ɑ/ and /æ/ as being long. From a presumably more phonetic perspective, however, /æ/ has been called lax (and therefore short?) as against the tense /ɑ/ [3]. Confounding the issue is that vowel height is also a determinant of vowel duration, low vowels being longer than high ones [4]. For certain views expressed on the matter it is unclear whether the long-short (or tense-lax) classification is phonetically based or is rather motivated by phonological (including phonotactic) considerations. The best known phonetic study of the matter [2] supports the view that /ɪʊɛʌ/ are relatively short, but does not clearly suggest a short-long dichotomy (Fig.1).

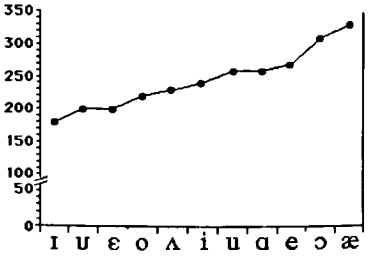


Figure 1. Rearrangement of the mean vowel durations for five speakers reported in Peterson and Lehiste (1960).

Thus we find no greater difference between the "short" /ʌ/ and long /i/ than between many other adjacent or near-adjacent pairs not considered to differ in either length or tensity, e.g. /ɪʊ/ /ʊʌ/ /ɛʌ/ /ue/ /eɔ/.

The present study was undertaken to gather evidence from spontaneous speech on the robustness of relative duration as an acoustic correlate of the generally accepted vowel length distinction in American English.

PROCEDURE

In two separate sessions, of about ten minutes each, a pair of native speakers of American English, one female and one male, were recorded while engaged in informal spontaneous conversation. Four speakers in all participated. The recorded speech, digitized at 22 Kh, was subjected to FFT analysis by means of the Signalyze' program. Signal intervals corresponding to stressed vowels between obstruent consonants were selected on the basis of auditory, waveform and spectrographic criteria, and their durations were defined as the interval from release of the prevocalic constriction to the onset of the following constriction. Excluded from consideration were vowel tokens judged auditorily to have been produced with either "contrastive stress" or "drawling." as well as the three diphthongs /aʷ ɑʲ ɔʲ/ and the retroflex vowel [ɚ]. The measurement data thus assembled were classed by phonological category, and the categories grouped into "short" and "long" sets. For the present purpose we elected to apply two classifications, one where /ɪʊɛʌ/ are short and all the others long, and a second in which /ɪʊɛ/ represent the short vowel set and /iue/ are their long counterparts. Given the nature of the speech samples, the numbers of tokens for the different phonological categories and durational sets were expectably quite unequal.

Although a division of the English vowels into long and short sets appears to be solidly enough based to call for no more data collection based on unspontaneous speech, a set of "control" data was gathered from a single speaker producing forms of the type /bVt/ and /bVd/ in a carrier Please pronounce — once again. The target "words" in this context were regularly produced with voiceless /b/ and non-flapped allophones of the final alveolar consonants, and were subjected to the same recording, signal processing and analysis applied to our samples of spontaneous speech.

FINDINGS

As we have seen, the vowel duration data presented in [2] do not provide the strongest possible evidence for a dichotomous separation into short and long sets. Thus even the single speaker data in [2], with no scope for interspeaker variability, show /ʊ/ and /ʌ/ differing by 30 ms, while for /ʌ/ vs /i/ the difference amounts to all of 3 ms. Our own single-speaker control data (Fig. 2) are much more consistent with a short-long dichotomy, since the durational differerence between the adjacent pair /ʌ/-/i/ shows a significance level (unpaired t-test, df 38, t = -6.3. $p \leq .0005$) matched only by that for /ɪ/-/ʊ/.

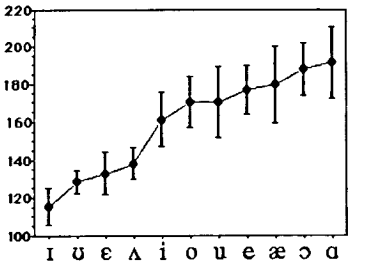


Figure 2 . Means ± one standard deviation of vowels (20 tokens each) produced in a fixed carrier sentence by one speaker.

Turning now to the spontaneous speech data, we find mean durations and standard deviations for the eleven vowel categories measured as shown in Fig. 3.

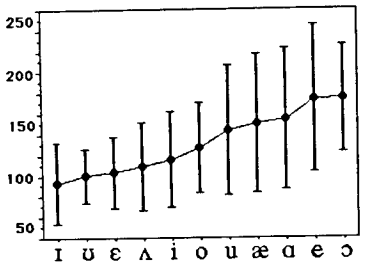


Figure 3. Mean vowel durations(ms) ± one s.d. for our four speakers

This picture is closer to the one in Fig. 1 than in Fig. 2, in that while /ɪʊɛʌ/ are shorter than any of the others, the division between the "short" /ʌ/ and the "long" /i/ is no greater than those of /o/–/u/ or /ɑ/–/e/. It would seem, then, that at the "superficial" phonetic level, i.e. where physical measurement may be relevant, there is no clear basis for a sharp cleavage between long and short vowels in spontaneous English. Of course, if we group /ɪʊɛʌ/ as "short" and all the others as long, then we shall certainly find mean group durations that are significantly different, as the individual speaker data in Table 1 indicate.

Table 1.Mean durations in ms: means, standard deviations, and significance levels per unpaired t-*tests.*

Spkr:	DS	DL	MC	JH
/ʊɛʌ/				
M	108	98	99	92
SD	46	35	30	16
n	45	101	44	25
/iueɔæɑ/				
M	181	134	143	125
SD	67	52	45	40
n	43	81	59	25
df	87	174	101	48
t	-6.0	-4.9	-5.6	-3.8
$p <$	.001	.001	001	.001

When these grouped data are summed across the four speakers we see just what we should expect (Fig. 4). The mean for the four shortest vowels is of course smaller than the other, but the two means are separated by an amount that almost exactly equals the average of the standard deviations of the two groups.

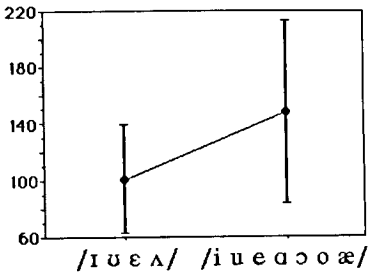


Figure 4. Means ± one s.d. for the summed data of Table 1

If our spontaneous speech data fail to show any sharp division between short and long vowels, they certainly do not encourage the supposition that this failure is the result of a faulty length assignment of the (to some) questionable vowels /æɑɔ/, since moving one or more of these to the "short" category would hopelessly weaken any case for a long-short (or tense-lax) distinction based on a phonetic length difference.

Table 2 .Mean durations in ms: means, standard deviations, and significance levels per unpaired t-*tests.*

Spkr:	DS	DL	MC	JH
/ɪʊɛ/				
M	112	93	95	92
SD	48	29	21	16
n	33	83	22	22
/iue/				
M	203	129	140	114
SD	67	53	47	42
n	14	28	34	9
df	45	109	54	29
t	-5.3	-4.5	-4.3	-2.2
$p <$	.001	.001	.001	.03

To do justice to the literature on vowel length in English, we should point out that many studies restrict attention to three vowel pairs: /ɪ/–/i/, /ʊ/–/u/, /ɛ/–/e/. The data for the shorter vs longer vowels of this restricted subset (Table2) show that for each speaker the two vowel sets differ significantly in their mean durations. At the same time it may be noted that JH produced his long vowels with durations scarcely greater than those of DS's short vowels. When the data are pooled across speakers their means and standard deviations are as shown in Fig.5. It is evident that while there is a difference between the means of pooled short and long vowels, a large proportion of the short vowels lie well within the range of values characteristic of their long counterparts.

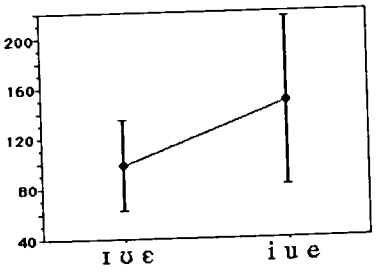


Figure 5. Means ± 1 s.d. for pooled data of Table 2.

CONCLUSION

There is every reason to believe that the vowels of English (and very likely all other languages as well) show regular differences in duration, at least according to the commonly accepted definitions of vowel onset and offset. It is less certain that the American English fall into two distinct subsets, short vs long. A partitioning of these vowels into short and long categories, insofar as it is phonetically based, rests on data derived from speech carefully selected to eliminate a variety of factors affecting speech timing generally,– speaker variability; overall speech tempo; contextual factors, e.g. immediate phonetic context, place within word, phrase or utterance. If speech data gathered under controlled conditions **do** indicate a short-long dichotomy (as per the data of Fig.2), and if phonetic/phonological classification appeals to this sort of data, then it appears that in spontaneous speech there are factors at work (controlled for in laboratory speech) powerful enough to blur, if not entirely obliterate, durational differences that might be considered to be inherent properties of the vowels themselves.

ACKNOWLEDGEMENT

This work was supported by NIH Grant HD 01994 to the Haskins Laboratories.

REFERENCES

[1] Goldsmith, J.A. (1990), *Autosegmental and metrical phonology*, Oxford: Basil Blackwell.

[2] Peterson, G.E. and Lehiste, I. (1960), "Duration of syllable nuclei in English are linguistic sounds made of ?", *Journal of the Acoustical Society of America*, vol. 32, pp. 693-703.

[3] Ladefoged, P. (1982), *A Course in Phonetics*, (2nd edition). New York: Harcourt Brace Jovanovich, Inc.

[4] House, A.S. (1961) "On vowel duration in English", *Journal of the Acoustical Society of America*, vol. 33, pp. 1174-1178.

ACOUSTIC CHARACTERISATION OF SCHWA: A COMPARATIVE STUDY

Sally A R Bates

Department of Linguistics, Edinburgh University

ABSTRACT

This paper presents results from a comparative acoustic study of the contextual variability displayed by the twelve monophthongs in Southern British Standard English. These indicate a hierarchy of vowel robustness which, in general, parallels the hierarchy of inherent vowel duration. Maximal context-dependency for schwa supports the proposal that it has no independent phonetic target but is completely unspecified for tongue position.

INTRODUCTION

The principle aim of the investigation was to assess the magnitude and patterns of contextual variability displayed by schwa, the central or 'reduced' vowel in English, in the light of the proposal that it may be completely unspecified for tongue position [1]. Following their study of schwa tokens in /pVpə'pVpə/ sequences, Browman & Goldstein reject a targetless analysis in favour of a co-production account of schwa's variability. They claim that schwa is characterised by an active gesture but that this gesture is completely overlapped by the gesture for a following full vowel. However, in a more recent study which examines coarticulatory effects on Dutch schwa in 'VCəC and Cə'CV nonsense words, Van Bergem [2] presents evidence in support of a targetless analysis. The present investigation represents an extension of Browman & Goldstein's and Van Bergem's work insofar as it examines schwa in a more comprehensive range of contexts and in meaningful connected speech data. It also provides a comparative framework in which to assess schwa's variability and its targeted/targetless status.

A second objective was to address the question of whether the full vowels vary inherently with respect to the extent to which they are susceptible to coarticulatory effects. Stevens' [3] quantal theory of speech production predicts greater acoustic stability for the point vowels /i, ɑ, u/ than for the non-point vowels. According to Stevens, these vowels are articulated in those regions of the vocal tract where articulatory perturbations have minimal effects on the acoustic output. Recasens [4] proposes that front vowels are inherently more stable than back vowels because they involve a greater degree of mechanical constraint on the tongue-body during their production. These proposals are further explored in a quantitative evaluation of the relative context-dependency shown by the vowels /i, ɪ, ɛ, a, ɜ, ɑ, ʌ, ɒ, ɔ, o, u/. The data comprises over 8000 vowel tokens, including over 2000 schwa tokens. These are taken from 660 phonemically balanced sentences read by one male speaker.

STATISTICAL ANALYSIS AND RESULTS

The overall degree of context-dependency shown by different vowels was assessed in a series of multiple regression analyses. The predictor variables were preceding context (lhs) and following context (rhs), characterised by formant values averaged over the initial and final 10 ms of the vowel, and vowel duration (dur). The dependent variable was either F1 or F2 sampled at the durational vowel midpoint. The results are presented in Tables 1 and 2. The R^2 value represents the total proportion of variance in the dependent variable that is accounted for jointly by the predictor variables. The β values indicate the relative weighting of the unique contributions made by the individual predictors. All R^2 values are significant at $p < .01$. With few exceptions, the β values are also significant at $p < .01$. Those values which are not significant are italicised and emboldened.

Schwa displays the highest levels of context-dependency. The mid-high, lax vowels /ɪ/ and /ʊ/ display a comparable degree of context-dependency along F2 and, in the case of /ɪ/, along F1 also. Following schwa and the mid-high, lax vowels, the tense vowel /u/ and the mid-low, lax vowels /ʌ/ and /ɛ/, in descending order, show the highest R^2 values. The low back vowels /ɑ/ and /ɔ/ show the least overall context-dependency followed, in ascending order, by the low, front vowel /a/, the mid-low, back vowel /ɒ/ and the tense vowels /ɜ/ and /i/. All vowels display a lower absolute amount of context-dependency along F1. However, with the exception of /ʊ/, the rank ordering of vowels from least to most context-dependent is similar for both F2 and F1.

The β values indicate generally greater carryover than anticipatory coarticulation. One-way analyses of variance performed on the same data, also show greater differentiation in vowel midpoint value as a function of preceding compared with following consonantal place of articulation for all vowels except /ə/ and /ɪ/. Schwa and /ɪ/ both show a slightly greater degree of differentiation in midpoint value as a function of following consonantal place of articulation (see [5]).

CONCLUSION

The results of the regression analysis provide strong support for the theory that schwa is unspecified for tongue position. Given that 100% prediction accuracy is unlikely owing to the presence of random variability, 92% explained variance arguably denotes maximal context-dependency along F2. The total proportion of explained F1 variance is also high and comparable to the levels of explained F2 variance reported for /ə/ by Van Bergem [2] (between 72% and 79%). This finding provides empirical support for Keating's [6] phonetic underspecification hypothesis. A strong interaction of context effects through schwa, highly linear first and second formant trajectories and a comparable range in onset, midpoint and offset value for the same schwa data reported by Bates [5] also support the characterisation of schwa as a phonetically transparent segment which is interpolated through by the trajectory between adjacent specified segments.

Table 1: Regression results for F2
'lhs' denotes preceding context, 'rhs', following context and 'dur', duration.
β values which fail to attain significance at p < .01 are italicized and emboldened

	R^2	F-value	df.		β value	
				lhs	rhs	dur
i	.5294	246.77	658	.48	.32	.44
ɪ	.8963	3747.83	1301	.57	.50	.14
ɛ	.6973	343.94	448	.52	.53	.18
a	.4155	75.35	318	.53	.43	.09
ə	.9166	8057.98	2199	.57	.52	.02
ɜ	.5192	51.48	143	.61	.36	*.02*
ɑ	.1119	6.69	181	.24	*-.10*	-.21
ʌ	.5785	129.94	284	.71	.40	-.18
ɔ	.3770	50.22	249	.49	*.05*	-.46
ɒ	.4449	91.63	343	.56	.18	-.43
ʊ	.9024	283.53	92	.57	.49	-.16
u	.7569	210.64	203	.52	.53	*.04*

Table 2: Regression results for F1

	R^2	F-value	df.		β value	
				lhs	rhs	dur
i	.3923	141.61	658	.45	.36	-.17
ɪ	.7328	1194.72	1307	.53	.48	.08
ɛ	.4209	109.02	450	.35	.46	.38
a	.3633	61.44	323	.50	.22	.54
ə	.7372	2049.44	2192	.55	.43	.17
ɜ	.3818	29.43	143	.29	.31	.44
ɑ	.1320	8.77	173	.31	.15	.23
ʌ	.4355	72.77	283	.39	.33	.49
ɔ	.0650	5.76	249	.23	*-.11*	*-.02*
ɒ	.3359	57.66	342	.41	.32	.32
ʊ	.3428	16	92	*.13*	.47	.22
u	.4514	55.69	203	.28	.52	*-.14*

Evidence that the full vowels vary inherently with respect to degree of context-dependency is also consistent with Keating's [7] proposal that segments may show varying degrees of underspecification along a given dimension. According to Keating's window model of coarticulation, segments are characterised by the full range of contextual variability they exhibit. Segments with a full or narrow specification for a given feature show less overall variability along the corresponding phonetic dimension(s) than segments which are less narrowly specified. The results reported here indicate a continuum of phonetic underspecification. Broadly speaking, this ranges from the inherently long vowels /ɑ, ɔ, a/ which may be thought of as the most narrowly specified and hence least contextually variable, to the less narrowly specified and more contextually variable short vowels /ɛ, ʌ, ʊ/, to schwa and /ɪ/ which have the shortest intrinsic durations and which, being completely unspecified, show maximal context-dependency.

The comparable level of context-dependency observed for /ɪ/ as for schwa accords with its status as the other reduced vowel in English. The near maximal context-dependency along F2 for /ʊ/ may also be attributed to its lexical distribution. A high proportion of /ʊ/ tokens occur in words which carry relatively little semantic weight such as the modal verb forms "could, would, should" or the prepositions "to, into" and the pronoun "you" in which it alternates with [u]. Phonetic vowel reduction is closer to diachronic fossilisation in these function words than in words which carry a heavier semantic load.

Greater acoustic stability for the more peripheral vowels (i.e. /ɑ, a, i/) and for the back, rounded vowels /ɔ, ɒ/ compared with the more central vowels (i.e. /ɪ, ɛ, ʌ, ʊ/) accords with Stevens' [3] quantal predictions. The high context-dependency observed for /u/ may be attributed to the effects of lip-rounding coupled with its fronted realisation by the present speaker. The results do not support Recasens (1991) proposal that front vowels are inherently more stable than back vowels.

REFERENCES

[1] Browman, C. P. & Goldstein, L. (1992), Targetless schwa: an articulatory analysis. In Ladd, B. and Docherty, G. J. (Eds.), *Labphon II*, 26-67. Cambridge University Press, Cambridge.

[2] Van Bergem, D. (1994), A model of coarticulatory effects on the schwa. *Speech Communication*, vol. 14, 143-162.

[3] Stevens, K. N. (1989), On the quantal nature of speech. *Journal of Phonetics*, vol. 17, 3-45.

[4] Recasens, D. (1991), An electropalatographic and acoustic study of consonant-to-vowel coarticulation. *Journal of Phonetics*, vol. 19, 177-192.

[5] Bates, S. (forthcoming). Towards a definition of schwa: an acoustic investigation of vowel reduction in English. *Phd thesis*. University of Edinburgh.

[6] Keating, P. (1988), Underspecification in phonetics. *Working papers in Phonetics*, vol.69. University of California Linguistics Club.

[7] Keating, P. (1990), The window model of coarticulation. In Kingston, J. and Beckman, M. E. (Eds.), *Papers in Laboratory Phonology I: Between Grammar and Physics of Speech*. Cambridge University press, Cambridge.

VOWEL NORMALIZATION REVISITED: INTEGRATION OF ARTICULATORY, ACOUSTIC, AND PERCEPTUAL MEASUREMENTS

C.-S. Yang and H. Kasuya
Utsunomiya University, Utsunomiya, Japan

ABSTRACT

Vocal tract (VT) area functions were measured from magnetic resonance images (MRI) for five Japanese vowels /i, e, a, o, and u/ across a child, a female adult, and a male adult. Effects of uniform and nonuniform normalization of the area function with respect to the length of three parts of VT, *i.e.* oral, pharyngeal and laryngeal sections, are investigated at articulatory, acoustic, and perceptual levels. Significance of uniform normalization is suggested.

INTRODUCTION

Relationships of formant patterns of vowels among male, female and child speakers are found to be nonuniform [1],[7]. How human beings normalize in the auditory perception a specific class of vowels is a classical but difficult problem. Many efforts have been made to solve the problem [1]-[7]. Kasuya, *et al.* [3] and Fujisaki and Nakamura [4], for example, proposed a coordinate system for the auditory representation of vowel classes based on uniform scaling of the first three formant frequencies in terms of the vocal tract length. Fant attributed the non-uniformity observed in the formant patterns between female and male speakers to that of VT dimensions [1]: ratio of pharynx length to mouth cavity length is greater for males than for females. Nordstrom found that anatomical differences between males and females only explain part of the differences based on the VT shapes predicted from X-ray photographs of midsagittal sections [5].

In this paper, we first measure the area function of three parts of the VT, *i.e.* oral, pharyngeal and laryngeal cavities, from MRI data of the five Japanese vowels using a newly developed image processing method [8] and investigate acoustic and perceptual significance of uniform and nonuniform scaling of the length of the three cavities.

METHOD

Measurement of VT area function

We have developed a method to accurately measure 3 dimensional VT shapes from MRI data, for the acquisition of which a General Electric SIGNA machine (1.5 T) was used [8]. VT data were obtained for five Japanese vowels, /i, e, a, o, and u/, of three subjects, a child, a female, and a male. The VT was divided into three sections as shown in Fig.1: the oral (from the lips to the uvula), pharyngeal(from the uvula to the top of the epiglottis), and laryngeal (from the top of the epiglottis to the glottis) sections. Length of each section was measured along the VT center line which was semiautomatically estimated on the midsagittal section image. Percentage of the length of each VT section to the entire VT length was then calculated for all the vowel data.

Uniform and nonuniform scaling of VT dimensions

Each of the area functions of the child and female subjects was normalized, first by making the entire VT length identical to the male's by using two different methods, *i.e.* uniform and nonuniform scaling, and then by adjusting the areas so that the maximum value becomes identical to that of the male.

In the uniform scaling, an entire VT length was uniformly extended to that of the male following the next equation (see Fig.2(a)):

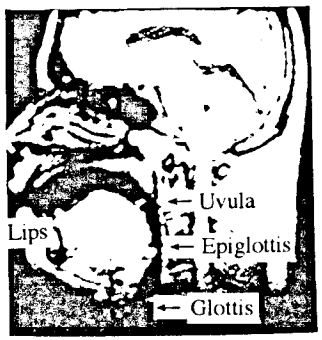


Fig. 1 Midsagittal section of the vocal tract.

$$x' = \alpha_L \times x,$$

where x is an original length from the glottis, α_L is a scale factor and x' is a normalized length.

In the nonuniform scaling, on the other hand, each length of the three VT sections of the child and female subjects was separately extended to that of the male as follows (see Fig.2(b)):

$$x_i' = \alpha_{L_i} \times x_i, \text{ for } i=1, 2, 3,$$

where x_i is an original length measured from the upstream end of the i-th section and α_{L_i} is a scale factor of the i-th section.

After scaling the VT lengths, all the values of the area function were adjusted so as to have the same maximum value as that of the male.

Computation of formant frequencies

The first four formant frequencies and bandwidths were computed from the acoustic transfer function [9] for all the VT area functions including the uniformly and non-uniformly normalized ones.

Perceptual experiments

Perceptual similarity experiments were performed on the phonetic quality of vowel sounds synthesized with the formant frequencies that were obtained from the area functions of the female subject.

Vowel stimuli used were the original vowel sounds that were synthesized from the original area functions of the female (Reference, REF), vowels that were synthesized from the uniformly normalized area functions (Uniform vowel, UV), and those that were synthesized from non-uniformly normalized area functions (Nonuniform vowel, NV). Fundamental frequencies of the REF stimuli were the ones of the vowels spoken by the female and those of the UV and NV stimuli were the same as the ones of the male. Duration of the stimuli were all 600 ms.

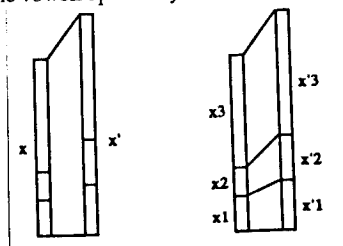


Fig. 2 (a)Uniform and (b)nonuniform normalization of the VT length.

A triad consisting of either REF, UV and NV, or REF, NV and UV was presented to a speech scientist who was trained to make a phonetic judgment of vowel sounds. An interval between the stimuli within the triad was 800 ms and time for the judgment was 3.5 seconds. All the triads of the five vowels were randomly presented to the subject who was required to make a judgment on which is phonetically more similar to the REF.

RESULTS AND DISCUSSIONS

VT area functions

Figure 3 illustrates a percentage of each length of the three VT sections to the whole VT length. The male shows smaller oral cavity length but larger laryngeal section length in percentage than the female and child. It is also seen that the percentage of nonuniformity is different from vowel to vowel.

Scalings of VT length

Using the uniform and nonuniform sacling methods described above, the area functions of the female and child were normalized as shown in Fig. 4, where (a) and (b) are for uniform and nonuniform scalings, respectively. It seems from the figure that the normalized area functions depend little on the type of the scalings in all the vowels.

Formant frequencies

The first four formant frequencies were calculated from the original and normalized area functions [9]. Distributions of the first and second formant frequencies of the female are shown in Fig. 5. Differences in the formant values between the two scaling methods were all less than 5 % which is close to the perceptual difference limen (DL) of the formant frequencies [10]. This was the case for the child. These seggest that nonuniformity of the VT dimensions among the child, female and male speakers is only a secondary factor in the normalization process.

Perception of vowel quality

Results of the perceptual similarity tests of vowel quality between REF and UV or NV stimuli were such that REF stimuli were more similar to UVs than NVs in the vowels /i, a, and o/ , nearly equally similar

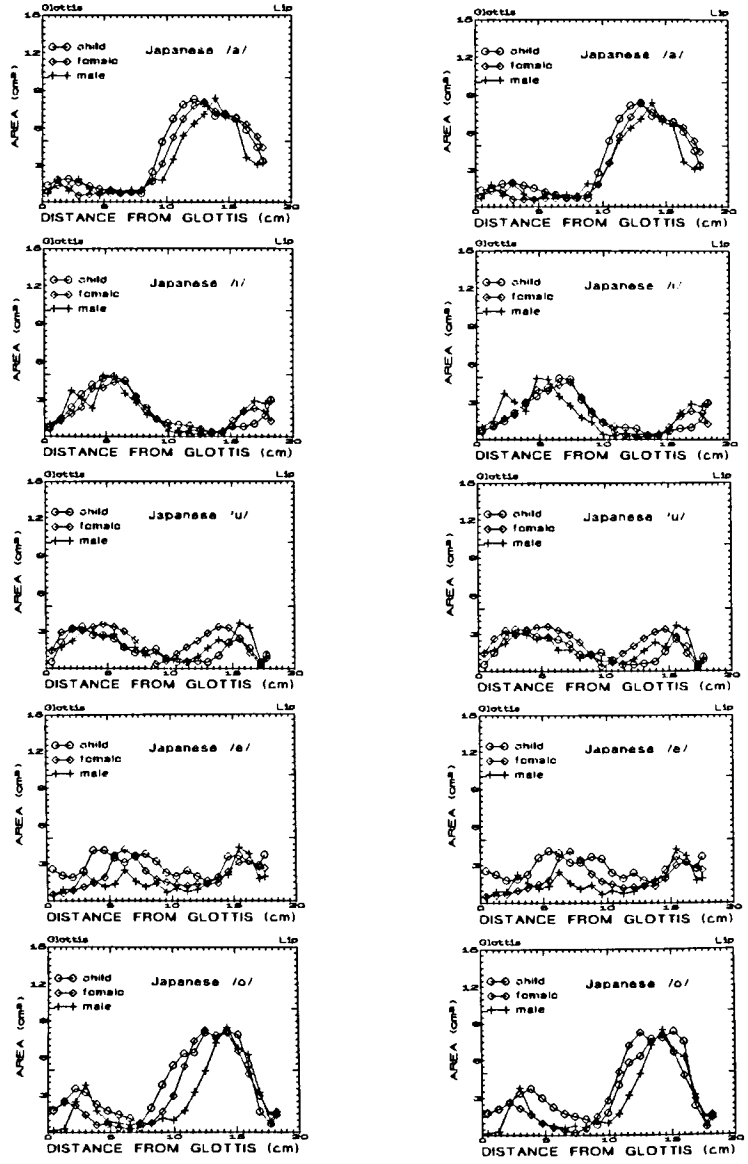


Fig. 4 VT area functions of the child, female and male subjects for the five Japanese vowels, normalized by (a) uniform and (b) nonuniform scaling.

to the two in the vowels /e and u/.

CONCLUSION

Nonuniform scaling of the vocal tract dimensions with respect to the length of oral, pharyngeal and laryngeal cavities effects little on the first three formant frequencies and the vowel sounds of the formants normalized uniformly were perceived phonetically equivalent. These findings support the significance of uniform scaling of the VT length.

REFERENCES

[1] Fant, G. (1973), *Speech sound and features,* The MIT Press.
[2] Kasuya, H., Suzuki, S. and Kido, K. (1968), "Changes in pitch and first three formant frequencies of five Japanese vowels with age and sex of speakers," J. Acoust. Soc. Jpn., vol. 24, pp.355-364 (in Japanese).
[3] Kasuya, H., Suzuki, H., and Kido, K.(1968), "On auditory model of vowel perception," Proc. 6th Int. Congr. on Acoustics, B-3-3, Tokyo, Japan.
[4] Fujisaki, H. and Nakamura, N. (1969), "Normalization and recognition of vowels," Ann. Rep. of the Engr. Res. Rept., vol.28, pp.61-66.
[5] Nordstrom, P.-E. (1977), "Female and infants vocal tract simulated from male area functions," J. of Phonetics, vol.5, pp.81-92.
[6] Kent, R.D. and Forner, L.L. (1979), "Developmental study of vowel formant frequencies in an imitation task," J. Acoust. Soc. Am., vol. 65, pp.208-217.
[7] Wakita, H. (1977), "Normalization of vowels by vocal tract length and its application to vowel identification," IEEE Trans. Acoust., Speech & Signal Process., vol. ASSP-25, pp.183-192.
[8] Yang, C.-S. and Kasuya, H. (1994), "Accurate measurement of vocal tract shapes from magnetic resonance images of child, female and male subjects," Proc. Int. Conf. on Spoken Language Process., vol. 2, pp.623-626.
[9] Sondhi, M. M. and Shroeter, J. (1987), "A hybrid time-frequency domain articulatory speech synthesizer," IEEE Trans. Acoust., Speech & Signal Process., vol. ASSP-35, pp. 955-967.
[10] Flanagan, J.L. (1972), *Speech analysis, synthesis and perception*, 2nd ed., Springer-Verlag, New York.

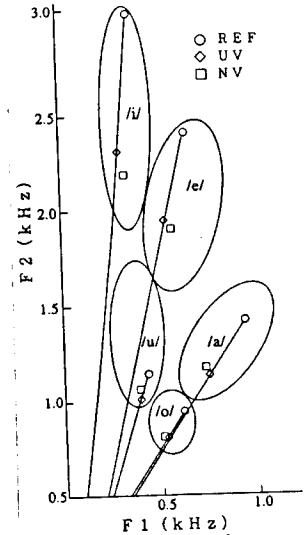


Fig.5 F1 and F2 of the vowels computed from uniform(UV) and nonuniform(NV) scalings.

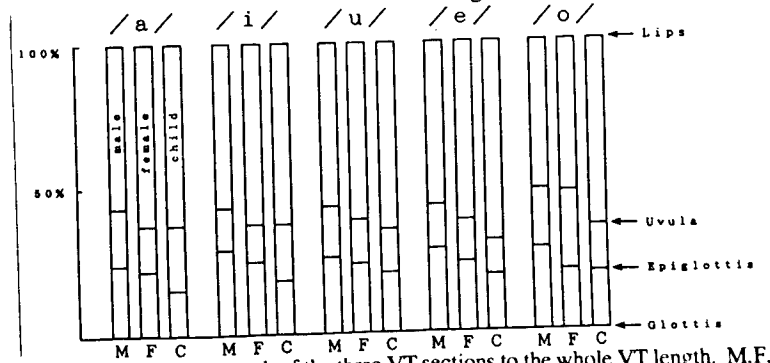


Fig. 3 Percentage of the length of the three VT sections to the whole VT length. M,F, and C are respectively male, female and child subjects.

TEMPORAL ADJUSTMENT OF DEVOICED MORAE IN JAPANESE

Mariko Kondo
Department of Linguistics, University of Edinburgh

ABSTRACT

The effect of voiced and devoiced high vowels on moraic duration in Japanese were measured. Results showed that devoiced morae were significantly shorter than /CV/ morae. However, shorter durations of devoiced morae appeared to be adjusted at the word level, rather than within a mora. The apparent lengthening of moraic consonants was actually caused by the underlying devoiced vowel in the same mora.

INTRODUCTION

Standard Japanese is often cited as a Tmora-timed languageU. However, the theory of TmoraU as the basic unit of Japanese is disputable: the only convincing theoretical support is based on accentuation: lexical accent location is based on the moraic unit, not the syllabic unit.

It has traditionally thought that each mora in Japanese has a similar duration. In reality, many studies have agreed that the duration of morae actually differ, but there is a strong tendency for Japanese to try to equalise the duration of morae. (Hoequist Jr., 1983 [3], Sato, 1993 [6], etc.). Campbell and Sagisaka (1991) [2] did not find equal duration of morae in raw durations, but normalising segmental durations using z-score, they found mora-based segmental elasticity and duration compensation within CV sequences (moraic units) rather than V-C sequences (non-moraic units).

Some studies have questioned the phonetic reality of morae, notably Beckman (1982) [1]. Beckman measured the duration of segments and morae in various segmental combinations, but the mora did not appear to have phonetic reality.

Port et al (1987) [5] demonstrated that the duration of each mora was not necessarily equal, but the number of morae in a word determined the duration of words.

Most studies on temporal features in Japanese have not examined cases of devoiced morae. In Standard Japanese, high vowels /i,u/ are devoiced between voiceless consonants, or between a voiceless consonant and a pause. When a vowel is devoiced, the preceding consonant becomes moraic, constituting a mora on its own without a vowel. Beckman (1982) [1] showed that moraic consonants were not consistently longer than the non-moraic consonants.

Port et al (1987) [5] did not measure the duration of devoiced morae, but measured the duration of whole words with a devoiced mora and found that even when there was a devoiced vowel in a word, its duration was still adjusted and the word duration was fairly constant dependent on the number of morae in the word.

If morae with a devoiced vowel are considerably shorter than CV morae, does durational adjustment still operate at the word level? The results from the above studies have shown some sort of durational adjustment of segments based on mora, but are devoiced morae simply durational exceptions? If there are more than one devoiced mora in a word, does the word duration still maintain the target duration based on the number of morae in a word?.

This paper will investigate two levels of durational adjustments with particular relation to devoiced morae and words with devoiced morae. Principally, the following points will be examined: (1) the duration of devoiced and undevoiced morae, (2) the duration of a whole word with and without devoiced morae, and (3) the effect of the number of devoiced morae on the duration of a word.

EXPERIMENTAL METHODS

Six native speakers of Standard Japanese (2 male and 4 female) pronounced 41 test words containing 71 devoiceable vowels 3 times each in random order (41 test words x 3 times x 6 subjects = 738 tokens) containing 1328

devoiceable vowels. Their pronunciation of devoiceable vowels in the same words was not always consistent. When there was variation in the voicing of the same devoiceable vowel in the same word, the word was segmented and the segment durations were measured. 45 of the devoicing sites had voicing variations, excluding word-final position and pre-pausal position. The duration of moraic consonants were compared with that of corresponding CV morae. One female subject did not show any voicing variation. Therefore the results do not include her data. The comparison of duration was made only in the same mora in the same word uttered by the same speaker. All words which had voicing variation were segmented using the SUN Waves+ package.

RESULTS AND DISCUSSION

Durational ratio between moraic consonants and CV morae

The durational ratios between moraic consonants and corresponding CV morae was calculated. Two sets of measurements were taken for each devoicing site: for example, if a vowel in a word was devoiced in one utterance (p) and voiced in the other two utterances (q) and (r), two ratios (p/q) and (p/r) were calculated; if a vowel was devoiced in two utterances (x) and (y) and voiced in the other utterance (z), two ratios (x/z) and (y/z) were obtained. The period of aspiration after plosives was included as a part of plosives. The results are listed in Table 1.

Table 1 Average ratio by preceding consonants

consonant	No. of samples	mean ratio	SD
plosives	48	85.9 %	13.05
fricatives	16	88.0 %	14.96
Affricates	26	77.7 %	12.42
TOTAL	**90**	**83.93 %**	**13.97**

Statistical analysis by one-way ANOVA showed that the difference in the durational ratio between /CV/ morae and moraic consonants among the three types of preceding consonants was significant [$F(2, 87) = 4.002$, $p < .025$].

There were 18 out of 90 cases [45 sites x 2 comparisons] (20%) where moraic consonants were longer than CV morae: plosive [k] 8 out of 48 cases (16.7 %), affricates [tɕ] and [ts] 8 out of 16 cases (50 %), and fricative [ɕ] and [ɸ] 2 out of 26 cases (7.7 %). The average ratio of moraic consonants against CV counterparts was 83.93% (SD 13.97).

Figure 1 shows the mean duration of moraic consonants and consonants and vowels in CV morae averaged by all types of moraic consonants.

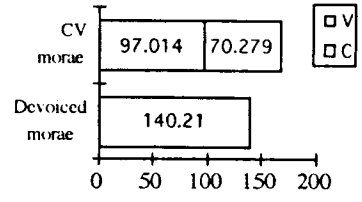


Figure 1 Average durational difference between moraic consonants and consonants and vowels in CV morae of all types of preceding consonants

The T-test (related) results showed that when the high vowels were devoiced, the remaining consonant in the same mora was significantly shorter than the equivalent CV mora, regardless of the type of preceding consonants: plosives [$t(23)=5.78$, $p<.001$], affricates [$t(7)=2.62$, $p<.025$], fricatives [$t(12)=6.62$, $p<.001$], and total [$t(44)=8.49$, $p<.001$].

Secondly, the duration of moraic consonants was also compared with the duration of non-moraic consonants in corresponding CV morae using T-test (related). The result found that the moraic consonants were significantly longer than the non-moraic consonant: plosives [$t(23)=11.93$, $p<.001$], affricates [$t(7)=9.26$, $p<.001$], fricatives [$t(12)=4.74$, $p<.001$], total [$t(44)=13.62$, $p<..001$]. In other words, the moraic consonants were significantly shorter than the equivalent CV morae, but at the same time they were significantly longer than the consonants in corresponding CV morae.

Considering the average ratio between the moraic consonants and the CV morae shown in Table 1, the moraic consonants were 83.93% of CV morae in duration compared to the prevocalic consonants which occupied 57.99% of whole CV duration. Comparing the moraic consonants and the prevocalic consonants, the moraic consonants were on average 44.53% longer than the prevocalic counterparts. In other words, there does appear to be some lengthening of consonants when the following vowels are devoiced and they become moraic, but it does not fully compensate for the reductions caused by the vowel devoicing.

Comparison of closure durations

The closure durations of moraic plosives and the plosive part of moraic affricates were compared with the closure durations of non-moraic plosives and the plosive part of non-moraic affricates. As shown in Figure 2, the average closure duration of plosives and affricates in CV morae was 56.16 ms (SD = 22.70), and that in devoiced morae was 57.16 ms (SD = 24.22). The T-test result (two-tailed) showed that the difference in the durations was not significant [t (31)=-0.51, n.s.]. However, the duration after the release of stop closure and fricative part of affricates in devoiced morae, and the added duration of these frication parts and the following vowel were very different.

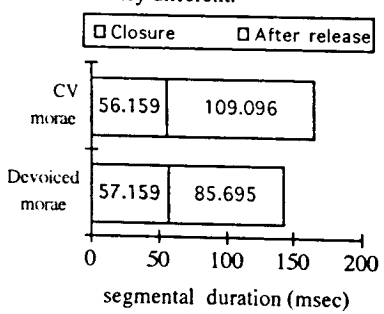


Figure 2 Average closure duration and the duration after release of plosives and plosive part of affricates in CV morae and devoiced morae

The average duration of moraic plosive and affricates, excluding closure duration was 85.70 ms (SD = 24.36), whereas that of CV morae excluding closure duration of the consonants was 109.10 ms (SD = 25.16). The statistical analysis by T-test (one-tailed) found the difference between them was significant [t(31)=7.12, p<.005]. The results suggested that although the durations of whole morae differed significantly, depending whether the vowel was voiced or devoiced, the closure durations of moraic plosives and affricates, and non-moraic counterparts did not show significant difference.

This suggests that the consonant was not actually lengthened, but rather vowel portion is hidden and it only appears as if consonant had been lengthened, as suggested by Jun and Beckman (1993) [4].

Durational adjustment within a mora

Devoiced morae were about 84% of the duration of /CV/ morae, which was a much higher proportion than the result obtained by Beckman (1982) [1]. One possible reason for this high proportion is the duration of the voiced vowels in /CV/ morae. The durational data for /CV/ morae in my experiment included partially voiced vowels which usually had a much shorter duration than their fully voiced counterparts. This might have brought the average duration of /CV/ mora down.

On the other hand, if there is durational adjustment within a mora in order to maintain the duration of each mora more or less similar, the shorter duration of partially voiced vowels should not affect the duration of whole /CV/ morae. The relationship between consonant and vowel durations within a mora was studied.

Although generally there was a tendency of negative correlation, the statistical analysis found the result was not significant [r(64)=-0.196, n.s.]. There seemed to be no durational adjustment between the durations of a consonant and a vowel within the same mora to keep the mora duration equal.

Devoicing and reality of mora-timing

The above results showing shorter durations of devoiced morae meant that the durations of voiced morae would also have to be adjusted to agree with the moraic duration theory of Port et al (1987). Therefore, the durations of whole words were measured and examined with relation to (i) the number of morae in a word, and (ii) the number of devoiced morae in a word.

Since the test words were pronounced in citation, the last mora of a word was usually lengthened, and the duration of the word initial stop closure and the beginning of fricatives were not always clear. Therefore, the durations of whole words were not exact measurements. However, there was a tendency for the duration of a whole word to lengthen as the number of morae in the word increased.

The number of morae in a word varied from 4 to 7, and the number of devoiced morae in a word varied from 0 to 3. Statistical analysis by ANOVA (3-way) found that the word durations were significantly influenced by the subjects [(F(4,81), p<.001] and the number of morae in a word [F(3,81), p<.001], but the effect of the number of devoiced vowels in a word was not significant [F(3,81), n.s.]. There was a significant interaction between subjects and the number of morae [F(10,81), p<.001].

The word durations of each number of morae were analysed using ANOVA (2-way) by the subject and the number of devoiced morae as the factors. The results shown in Table 2 found that for all numbers of morae in a word (4 to 7 morae), the effect of subjects was significant but the effect of the number of devoiced morae was not significant. No significant interaction between the numbers of morae in a word and devoiced morae was found.

Table 2 The ANOVA results of the effects of factors on word duration

No. of morae	Factor	
	Subject	No. of devoiced morae
4	F(3,19), P<.01	F(2,19), n.s.
5	F(4,20), P<.001	F(3,20), n.s.
6	F(3,12), P<.001	F(3,12), n.s.
7	F(4,30), P<.001	F(2,30), n.s.

The statistical results showed that the number of devoiced morae in a word was not an important factor for the whole duration of words. Rather it was the number of morae in a word and individual speakers that significantly influenced the whole duration of words. This may imply that the shorter durations of devoiced morae were adjusted at a word level so that the whole duration of a word does not have to change too much as Port et al (1987) [5] demonstrated.

CONCLUSIONS

Durational measurements of devoiced and /CV/ morae showed that devoiced morae were significantly shorter than /CV/ morae: proposed tendency of equalising mora duration was not tenable in devoiced morae. On the other hand, the number of devoiced morae in a word did not affect the duration of a word. That implies that shorter durations of devoiced morae were adjusted not within a mora but beyond the mora as suggested by Port et al (1987) [5]. Measurement of closure duration of stops suggested that the fairly high proportion of devoiced morae against /CV/ morae was not due to the compensatory lengthening of moraic consonant, but because the devoiced vowel was underlying as proposed by Jun and Beckman (1993) [4]..

REFERENCES

[1] Beckman, M. (1982), "Segment Duration and the 'Mora' in Japanese", *Phonetica* 39: 113-135

[2] Campbell, N. and Sagisaka, Y. (1992), "Moraic and syllable-level effects on speech timing", *Journal of Electronic Information Communication Engineering*, SP 90-107: 35-40

[3] Hoequist Jr, C. (1983), Syllable duration in stress-, syllable- and mora-timed languages, *Phonetica* 40: 203-237

[4] Jun, S.-A. and Beckman, M. (1993), "A gestural-overlap analysis of vowel devoicing in Japanese and Korean", paper presented at the Linguistic Society of America, Los Angeles, 7-10 Jan.

[5] Port, R. F., Dalby, J., and O'Dell, M. (1987), "Evidence for Mora Timing in Japanese", JASA 81: 1574-1585

[6] Sato, Y. (1993), The durations of syllable-final nasals and the mora hypothesis in Japanese, *Phonetica* 50: 44-67

THE RHYTHM RULE IN SPEECH PRODUCTION: THE EFFECT OF INTER-STRESS SYLLABLES

P.F. McCormack, J. C. Ingram~*
**Department of Speech Pathology, Flinders University, South Australia, &*
~Department of English, University of Queensland

ABSTRACT

The adjustment of linguistic stress patterns under the influence of rhythm is well attested, though the effects on speech production have been little investigated. An experiment is reported on the perceptual and acoustic effects on the production of the Rhythm Rule of manipulating the number of syllables between primary stresses.The tendency for stress shift to occur decreased as the number of syllables between primary stresses increased. There were both fundamental frequency and durational changes involved in the perceived shifts.

INTRODUCTION

It is generally acknowledged that shifts in the prominence patterns on some words in connected speech are due to a strong rhythmic constraint to prefer the alternation of stressed and unstressed elements [1] [2]. While *bamboo* spoken in a noun phrase such as *the bamboo one* has the main stress on the last syllable, in *the bamboo chair* there is a perception that the main stress has shifted to the first syllable. The Rhythm Rule formulated by Selkirk [2] involved a formal operation where stress shifts from one syllable of a word on to another in order to avoid "clashing" with an adjoining stress. Gussenhoven [3] proposed an alternative formulation whereby the perceived change in prominence is due to a process of pitch accent deletion rather than of phonetic stress "shift", at least in the pre-nuclear position. Horne [4] investigated Gussenhoven's [3] hypothesis in a single speaker of English, and found that the primary phonetic correlate for the Rhythm Rule was a decrease in the fundamental frequency in the second stressable syllable. These results were consistent with an accent deletion formulation. The primary phonetic cue for "stress shift" was change in fundamental frequency on the second stressable syllable of the "shift" word.

However, Horne [4] only investigated phrases in which the potential shift word was followed by a word with its primary stress on the first syllable (for example, *Dundee tartan*). What has not been investigated as yet is whether the tendency for the Rhythm Rule to occur is also dependent on the number of syllables between the primary stresses in both the "shift" word and the following word. For example, is "stress shift" more likely in *bamboo chair* than it is in *bamboo decoration*? If syllable number does play a role in the rhythm rule, it may also follow that durational changes in the "shift" words also provide a phonetic cue for "stress shift".

AIM

The aim of this experiment is to investigate the perceptual and acoustic effects on the production of the Rhythm Rule of altering the number of syllables between the main stress in the potential stress shift word and that in the following word (fulcrum).

PROCEDURE

Fifteen speakers of Australian English were recorded reading a series of sentences containing noun phrases which comprised of a potential stress shift word followed by words with varying syllable number to their primary stress. The sentences were designed to provide a phonological context where shift and non shift environments could be manipulated. Examples of the 5 contexts used are as follows:

Two contexts where no shift was predicted:

No stress following: *They were japanese ones at the hotel.*

Shift word focused: *They were JAPANESE tourists at the hotel.*

Three contexts where shift was predicted:

One syllable: *They were japanese tourists at the hotel.*

Two syllables: *They were japanese developers at the hotel.*

Three syllables: *They were japanese politicians at the hotel.*

Six potential stress shift words were used: *thirteen, bamboo, sardine, underdone, overnight,* and *japanese.* These words had been identified in a previous experiment [5] as being particularly susceptible to stress shift in speech production.

ANALYSIS

Recorded shift words, embedded in their noun phrase, were digitised at 20.8 kHz using the Soundscope speech signal processing program. The duration of the shift word, the duration of each foot, and the duration of the pause between the shift word and the following word was measured. In order to obtain a measure of variation in the duration of the first foot compared to the second foot, the duration of the first foot as a percentage of the duration of the whole word was calculated (relative duration). The peak fundamental frequency for each foot was also calculated using a peak-picking algorithm within the Soundscope program. In order to obtain some measure of the relative changes in fundamental frequency pattern between the 2 feet over different contexts, the value for the second peak was subtracted from that of the first (fundamental frequency shift).

Three phonetically trained linguists were asked to rate the stress levels in each shift word token as either: 1) the last stressed syllable is more prominent 2) both stressed syllables have equal prominence, or 3) the first stressed syllable is more prominent.

RESULTS

The perceptual results indicated that not only was there a strong perception of shift in the 3 rhythm contexts, but the strength of the shift dropped away as the syllable number between the main stress in the shift word and the main stress in the following word increased. There was a clear pattern to the perceived stress shift judgements across the contexts. The 3 contexts in which shift is predicted demonstrate strong shift values, well above the "equal prominence" value of one. The 2 contexts in which shift is predicted not to occur demonstrate low shift values, well below the value of one. A one way analysis of variance for context against stress shift judgement with an adjusted least significant difference (Bonferroni) test at the .05 level indicated significant differences between the rhythm and non rhythm contexts, and between the Rhythm 3 context (with 3 syllables distance) and the other 2 rhythm contexts ($p = .000$, $F = 178.28$, d.f. = 4, 375). Figure 1 displays the mean perceptual stress shift ratings for each context.

In all subjects there were phonetic changes in the shift words that corresponded to the judgements of stress shift. In the rhythm contexts the relative duration of the first foot was higher than in the non-rhythm contexts. There was also a positive increase in

fundamental frequency shift. Inspection of the data indicated that for 11 of the 15 subjects these results were due not to absolute changes in the first foot of each word but to changes in the second. The absolute duration and fundamental frequency of the second foot decreased in the rhythm contexts. This resulted in the relative prominence between the 2 feet shifting from the second to the first. However, for 4 of the subjects there were changes in the absolute duration and fundamental frequency for the first foot. Figure 2 displays the mean values for percentage durational change for each context, while Figure 3 displays the mean fundamental frequency shift for each context.

One way analyses of variance for relative duration of the first foot against context, and for fundamental frequency shift against context with adjusted least significant difference (Bonferroni) tests at the .05 level indicated significant differences between the rhythm and non rhythm contexts, and between the Rhythm 3 context (with 3 syllables distance) and the other 2 rhythm contexts ($p = .000$, $F = 29.1$, d.f. = 4, 375; $p = .000$, $F = 34.8$, d.f. = 4, 375). A step-wise linear regression analysis indicated that changes in relative duration was the primary acoustic correlate for the judges' perception of stress shift, followed by shifts in peak fundamental frequency.

DISCUSSION

The results provide confirmation that the Rhythm Rule is dependent on the metrical structure of the word following the shift word. In particular, the number of syllables between the primary stresses of the 2 words is critical for the expression of the Rhythm Rule. The suggestion by Gussenhoven [3] and Horne [5] that the phonetic realisation of the Rhythm Rule is one of pitch accent deletion needs qualification. While it certainly involves a decrease in fundamental frequency on the second foot in the shift word for most speakers, there were also systematic durational adjustments dependent on the metrical structure. The regression analysis indicated that these durational adjustments were the primary phonetic cue used by listeners.. Horne's [4] results, suggesting the primacy of fundamental frequency as the cue for "shift" most likely reflect the effects of investigating the Rhythm Rule without taking variations in inter-stress syllable number into account.

The increased numbers of subjects in this study also highlighted that for some subjects there were positive phonetic changes on the first foot in rhythm contexts. For these subjects, a deletion model of the Rhythm Rule does not appear to provide an appropriate description.

REFERENCES

[1] Liberman, M. & Prince, A. (1977) *On stress and linguistic rhythm*, Linguistic Inquiry, 8, 249-336.

[2] Selkirk, E. (1984) *Phonology and syntax: The relationship between sound and structure*. Camb., Mass.: MIT Press.

[3] Gussenhoven, C. (1986) Review of Selkirk 1984. *Journal of Linguistics*, 22, 455-474.

[4] Horne, M. (1990) Empirical evidence for a deletion formulation of the rhythm rule in English. *Linguistics*, 28, 959-981.

[5] McCormack, P. & Ingram, J. (1989) Phonetic evidence for the rhythm rule. Paper presented at *The Australian Language & Speech Conference*, Melbourne: Monash University.

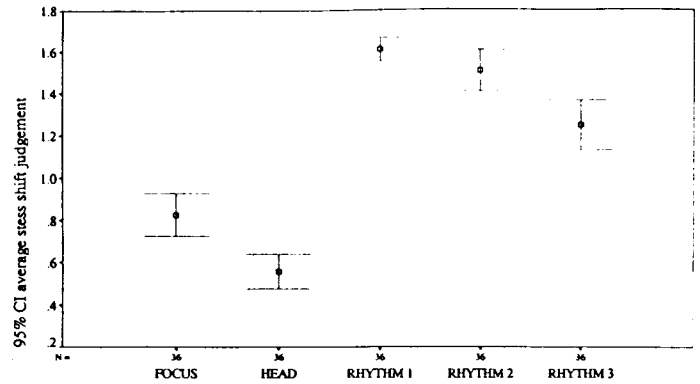


Figure 1. An error plot with 95% confidence interval of the mean stress shift judgement for each of the 5 contexts.

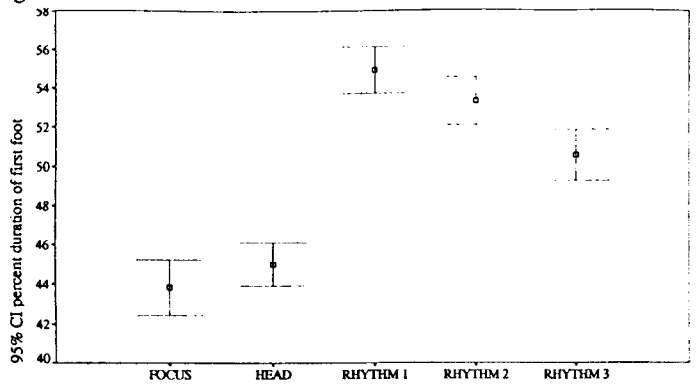


Figure 2. An error plot with 95% confidence interval of the mean relative duration of the first foot in the shift words for each of the 5 contexts (as a percentage).

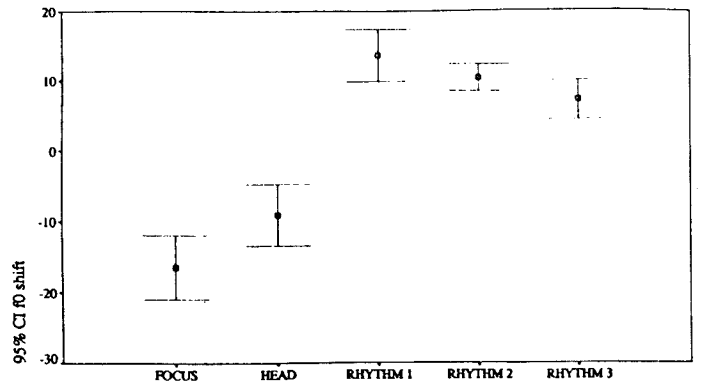


Figure 3. An error plot with 95% confidence interval of the mean relative shift in peak fundamental frequency in the shift words for each of the 5 contexts (in Hertz).

MULTIPLE EFFECTS ON SYLLABLE-INTERNAL TIMING IN NORWEGIAN

Dawn M. Behne and Bente H. Moxness
University of Trondheim, N-7055 Dragvoll, Norway

ABSTRACT

This study examines the concurrent effects of rate, stress, postvocalic voicing and distinctive length on timing in Norwegian C_1VC_2s. Findings suggest that the similar timing patterns associated with postvocalic voicing and phonological length may be distinquished by the timing of C_1. A consistent pattern of results was observed for focal and nonfocal conditions across speaking rates.

INTRODUCTION

The duration of phonetic segments is known to be affected by a variety of factors. When this occurs, the duration and relative timing of other components within a syllable can also be affected. Four factors which are known to affect segment durations in languages are speaking rate, focal stress, postvocalic voicing, and distinctive length. These factors all occur in Norwegian and constitute the basis for this investigation.

Background

Speaking Rate. For many languages, research has shown that segment durations within a syllable are affected by speaking rate. Although speakers may use different strategies to vary speaking rate, both vowels and consonants within syllables produced at a slow rate tend to be longer in duration than those produced at a fast speaking rate (e.g., [1]). Effects of speaking rate on segment durations specific to Norwegian have not been reported, but the general pattern of results observed for other languages is expected.

Stress. The duration of a segments within a syllable can also be affected by stress. Research on the effects of stress on segment duration in Norwegian and other languages has shown that vowels in stressed syllables are typically longer than those in nonstressed syllables and that the duration of consonants within the syllable may be similarly affected (e.g., [2][3]).

Postvocalic Voicing. Vowel duration can also be affected by the voicing of a following consonant. Vowels preceding a voiced consonant are typically longer than those preceding a voiceless consonant (e.g., [4]). This pattern has been observed in Norwegian words such as *takk* [tɑk:] "thanks" versus *tagg* [tɑ:g:] "thorn" [2][5]. Cooccurring with the effect on vowel duration, for many languages, including Norwegian, a postvocalic voiceless consonant has generally been found to be longer than a postvocalic voiced consonant (e.g., [2]).

Distinctive Length. Phonological distinctions can also be realized by means of segment durations within the syllable rhyme. Norwegian has traditionally been described as having a phonological distinction between short and long vowels. Accompanying this vowel length distinction is a difference in postvocalic consonant length. The phonotactics of Norwegian are such that, in a closed syllable, a distinctively long vowel tends to be followed by a short consonant, and a distinctively short vowel typically is followed by a long consonant. For example, the word "takk" [tɑk:] *thanks* has a distinctively short vowel followed by a long consonant compared to "tak" [tɑ:k] *hold* which has a distinctively long vowelfollowed by a short consonant. This quantity distinction of Norwegian vowels [2][6] and consonants[2] is also realized acoustically.

Current Investigation

Previous research suggests that speaking rate and stress have a relatively global affect on the duration of segments within a syllable, whereas postvocalic voicing and distinctive length principally affect segment durations within the rhyme, with an inverse relationship between the duration of a vowel and postvocalic consonant.

Recent findings suggest that in Norwegian [6] effects of postvocalic voicing and distinctive length might not be limited to the rhyme, and that prevocalic consonant duration may also be affected. In Norwegian, the of a prevocalic consonant duration was found to decrease with increased vowel duration due to postvocalic voicing, whereas the duration of a prevocalic consonant increased immediately preceding a phonologically long vowel. These timing patterns have been observed in both focal and non-focal conditions in Norwegian [2] and suggest that the duration of a prevocalic consonant may assist in distinguishing the similar timing patterns within the rhyme associated with postvocalic voicing and distinctive length.

In fluent speech segment durations reflect the concurrent influence of speaking rate, focal stress, postvocalic voicing and distinctive length. The present study extends previous research and investigates whether the timing patterns observed for postvocalic voicing and distinctive length in non-focal and focal conditions are affected by the relatively robust affects of speaking rate.

METHOD

Stimuli

Twelve target words were used in the investigation. All target words were real CVCs containing /i,ɔ,ɑ/ or /i:,ɔ:,ɑ:/ and a postvocalic /k/ or /g/. The initial consonant was either a stop or a fricative.

Brief dialogues were developed for each target word. Each conversation consisted of a question and a response. For each target word the set of conversations was balanced to include the target word as focused and nonfocused in both initial and final sentence position.

Subjects

The subjects were 9 native speakers of Norwegian between 20 and 30 years old with no history of speech or hearing impairment.

Procedure

Recordings were made of each subject producing the full set of conversations with an experimenter in a sound attenuated room. For each conversation the experimenter asked the question and the subject read the response. The full set of conversations was produced by each subject at a self-selected slow, medium and fast speaking rate. Subjects were encouraged to speak as if participating in a natural conversation.

Measurements

Three measurements were made within target C_1VC_2 from subjects' responses in each conversation: (1) frication/closure duration of C_1, (2) vowel duration, and (3) closure duration of C_2. Frication was measured from the beginning to the end of the aperiodic energy. Closure durations were measured from the start of the closure to the beginning of the release. Vowel duration was measured from the onset to the end of periodic energy.

RESULTS

For each of the three measures, a four way analysis of variance was calculated with speaking rate (fast, medium, slow), focus (nonfocal, focal), postvocalic voicing (voiceless C_2, voiced C_2), and distinctive length (short vowel, long vowel) as independent variables. Main effects were observed for all three measures.

Speaking Rate

Effects of speaking rate on segment durations are illustrated across panel columns in Figures 1 and 2. Speaking rate was found to affect the durations of C_1 [$F=109.30$, $p<.0001$], V [$F=241.75$, $p<.0001$], and C_2 [$F=214.24$, $p<.0001$]. For all three segments durations were reliably shorter at the fast rate than at the medium rate [F of $C_1=77.30$, $p<.0001$; F of $V=97.82$, $p<.0001$; F of $C_2=195.74$, $p<.0001$], which in turn were shorter than at the slow rate [F of $C_1=34.79$, $p<.0001$; F of $V=145.79$, $p<.0001$; F of $C_2=38.64$, $p<.0001$]. These findings are consistent with previous research showing that speaking rate has a relatively global affect on segment durations within a syllable, affecting both vowel and consonant durations.

Focal Stress

Main effects of focal stress was also observed for all three segment durations. As a comparison of the panel rows in Figures 1 and 2 illustrates, C_1 [$F=121.79$, $p<.0001$], V [$F=118.63$, $p<.0001$], and C_2 [$F=92.82$, $p<.0001$] durations were longer in the focal condition than the nonfocal condition.

Data were further analyzed to determine whether focus affected segment durations at each of the speaking rates. Reliable differences due to focus were observed for all three segment measures at the fast [F of $C_1=49.18$, $p<.0001$; F of $V=51.32$, $p<.0001$; F of $C_2=21.28$, $p<.0001$], medium [F of $C_1=51.42$ $p<.0001$; F of $V=43.66$, $p<.0001$; F of $C_2=51.78$, $p<.0001$],

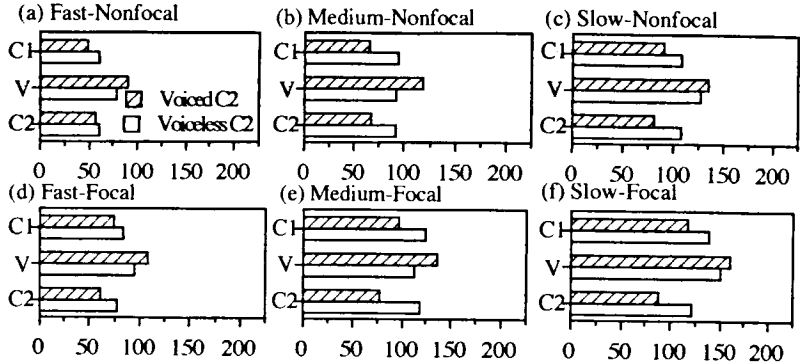


Figure 1. Mean segment durations with postvocalic voiced and voiceless consonants in nonfocal (top row) and focal conditions (bottom row) at fast (left column), medium (center column), and slow (right column) speaking rates.

and slow [F of C_1=29.87, p<.0001; F of C_2=35.99, p<.0001; F of V=22.81, p<.0001] speaking rates. These results support previous findings showing that focal stress tends to have a general effect on segment durations within a syllable [2].

Postvocalic Voicing. Main effects show that postvocalic voicing affected the duration of C_1, V, and C_2. As is demonstrated in Figure 1, the duration of C_1 is shorter when the postvocalic consonant is voiced than when it is voiceless [F=58.44, p<.0001]. Vowel duration is longer before a voiced consonant than before a voiceless consonant [F=69.47, p<.0001]. In addition, C_2 is shorter when it is voiced than when it is voiceless [F=293.93, p<.001]. As the results summarized in Table 1 and the means in Figure 1 show, this same pattern of results was observed for nonfocal and focal conditions at all three speaking rates. However, in some cases differences were not statistically reliable. Most notably, vowel duration was not affected by postvocalic voicing in either the nonfocal or focal condition at the slow speaking rate. Comparable results have been reported for English in conditions when multiple linguistic factors lead to increased segment duration [7], tentatively suggesting a vague upper limit on the duration of segments within a syllable. Similarly, at the fast speaking rate, although the expected pattern of results was obtained, no reliable difference was

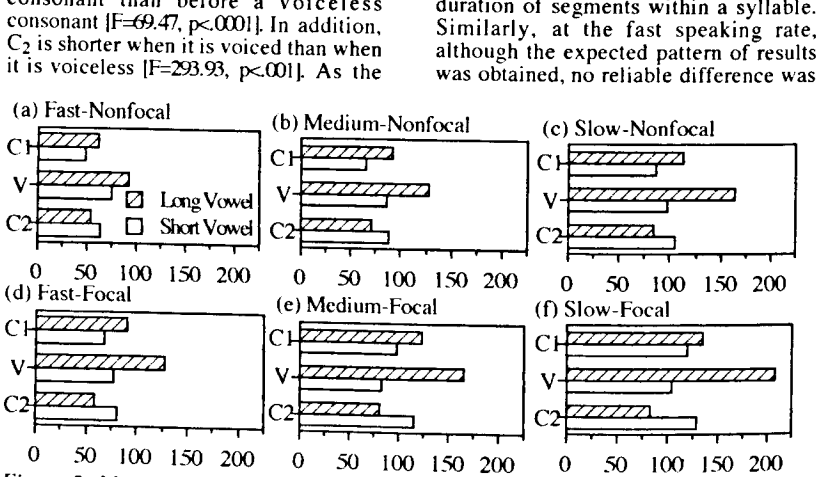


Figure 2. Mean segment durations (in ms) for distinctively short and long vowels in nonfocal (top row) and focal conditions (bottom row) at fast (left column), medium (center column), and slow (right column) speaking rates.

Table 1. F-values and probabilities for C_1, V, and C_2 for postvocalic voicing and distinctive length within the nonfocal and focal conditions at fast, medium and slow speaking rates.

Timing Factors		Postvocalic Voicing			Distinctive Length		
		C_1	V	C_2	C_1	V	C_2
FAST RATE	Nonfocal	F=6.57 p<.0111	F=19.14 p<.0001	F=2.64 n.s.	F=7.09 p<.0084	F=40.25 p<.0001	F=7.095 p<.0083
	Focal	F=3.70 n.s.	F=13.26 p<.0003	F=32.01 p<.0111	F=19.49 p<.0001	F=172.93 p<.0001	F=44.18 p<.0111
MEDIUM RATE	Nonfocal	F=24.85 p<.0001	F=59.27 p<.0001	F=45.50 p<.0001	F=20.63 p<.0001	F=146.86 p<.0001	F=16.36 p<.0001
	Focal	F=16.50 p<.0001	F=26.79 p<.0001	F=123.94 p<.0001	F=14.65 p<.0002	F=349.27 p<.0001	F=95.12 p<.0111
SLOW RATE	Nonfocal	F=5.04 p<.0257	F=1.90 n.s.	F=47.86 p<.0001	F=10.98 p<.0001	F=133.99 p<.0001	F=30.95 p<.0001
	Focal	F=9.44 p<.0024	F=2.32 n.s.	F=93.23 p<.0001	F=4.97 p<.0268	F=283.08 p<.0001	F=175.95 p<.0001

observed for duration of C_1 in the focal condition or C_2 in the nonfocal condition.

Overall, findings for postvocalic voicing show the expected inverse relationship between the vowel and postvocalic consonant durations and. results for C_1 suggest that it may also assist in cuing postvocalic voicing.

Distinctive Vowel Length. Results for distinctive length show that the duration ofC_1 is longer before distinctively long vowels than before distinctively short vowels [F=69.84,p<.0001]. The mean duration of distinctively long vowels is longer than the duration of distinctively short vowels [F=1050.63, p<.001] and C_2 is shorter following distinctively long vowels than following distinctively short vowels [F=287.18, p<.001]. As Figure 2 and the right side of Table 1 illustrate, this timing pattern was reliably observed in nonfocal and focal conditions at the fast, medium, and slow speaking rates. These findings suggest that distinctive length is reflected in the acoustic signal by the duration of the vowel, by the inverse relationship between the V and C_2 duration, and by the duration of C_1.

CONCLUSIONS

The results indicate that speaking rate and focal stress has a global affect on syllable-internal timing. The effects of distinctive vowel length and postvocalic voicing have an inverse effect on the duration of a vowel and postvocalic consonant within the rhyme. However, despite their similar effects on rhyme-internal timing, postvocalic voicing and distinctive vowel length have different effects on the duration of the prevocalic consonant. This pattern was observed in nonfocal and focal conditions across speaking rates. The robust nature of the timing patterns for the prevocalic consonant suggest that it may assist in distinquishing the similiar timing patterns of the rhyme associated with postvocalic voicing and distinctive length.

ACKNOWLEDGEMENTS

This research has been supported by a grant from the Norwegian Marshall Fund to the first author.

REFERENCES

[1] Lindblom, B. (1963),"Spectrographic study of vowel reduction", *JASA*, vol. 35, pp. 1773-1781.

[2] Behne, D. & Moxness, B. (1994). "Concurrent effects of focal stress, postvocalic voicing and distinctive vowel length on syllable-internal timing in Norwegian", *Proceedings of the International Congress on Spoken Language Processing*, Yokohama, Japan, pp. 139-142.

[3] Fry, D (1955),"Duration and intensity as physical correlates of linguistic stress", *JASA*, vol. 27, pp. 155-158.

[4] House, A. & Fairbanks, G. (1953), "The influences of consonant environment upon the secondary acoustical characteristics of vowels." *JASA*, vol. 25, pp. 105-113.

[5] Fintoft, K. (1970), *Acoustic analysis and perception of tonemes in some Norwegian dialects*, Oslo: Universitetsforlagets trykningssentral.

[6] Behne, D. & Moxness, B. (1994), "Syllable- and rhyme-internal timing: Combined effects of postvocalic voicing and distinctive vowel length in Norwegian.", *JASA*, vol. 95 p. 2979.

[7] Behne, D. & Nygaard, L.(1991),"Concurrent effects on duration II: Prevocalic and postvocalic consonants", *Res. on Speech Perc.*, vol 17, 263-284.

Temporal Organisation of Syllable Production in Cantonese Chinese

Eric Zee
City University of Hong Kong

ABSTRACT

The study is an investigation of the temporal organisation of syllable production in Cantonese Chinese. Results show (i) temporal compensation does not take place between C and V in the CV syllables, however, the duration of V is affected by the initial C type; (ii) tone plays a part in determining the temporal pattern of the syllables; and (iii) the reduction of diphthong duration or vowel+nasal sequence is contributed by the reduction of the 1st target vowel not the transition or the 2nd target vowel, or the vowel not the nasal.

INTRODUCTION

The temporal aspect of speech production has been extensively studied [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16]. The purpose of this study is to demonstrate certain regularities that have been found to be characteristics of the temporal structure of syllable production at the local time control level in Cantonese. By local time control, Allen [1] refers to "the specification of segment durations within syllables and possibly syllable durations within rhythmic phrases", and by global time control, he refers to "speech rate or tempo" (p. 222). The paper also shows whether temporal compensation takes place within sequences of successive component segments in the Cantonese monosyllables. Temporal compensation has been used to refer to the cases where the segment durations are inversely correlated [8], [9], [10] as an effort by the speaker to maintain an invariant syllable duration [7] or macrostructural invariance [16]. An inverse or negative correlation between the durations of two successive segments implies that "articulatory events are programmed at some higher level not in terms of single phonemes, but in terms of high-level articulatory units" and temporal compensation occurs when the durations of segments in a sequence are negatively correlated or the production of the segments takes place in terms of phonological units [9].

In Cantonese, four types of syllable structure are distinguished: $(C_1)V$:, $(C_1)D$, $(C_1)V:C_2$, and $(C_1)VC_2$, where C_1 may be an aspirated or unaspirated stop, an aspirated or unaspirated affricate, a fricative, a nasal, a liquid, or a glide; D a diphthong; V a short vowel; V: a long vowel; and C_2 an "-N", i.e., a syllable-final nasal [-m, -n, -ŋ] or "-S", i.e., a syllable-final unreleased stop [-p, -t, -k]. In Cantonese two types of diphthongs, D_1 and D_2, may be distinguished in terms of their internal temporal structure. D_1 refers to the diphthongs, such as [ui, iu, ai, au, ɔi], and D_2 to the diphthongs, such as [ei, ɐi, ɐu, ou, ɵy]. In Cantonese, six long citation tones, ˥, ˧˥, ˧, ˨˩, ˩˧, ˨, or 55, 33, 22, 21, 35, 24, respectively, occur on the open syllables or syllables with a nasal ending, $(C_1)V$:, $(C_1)D$, $(C_1)V:N$, and $(C_1)VN$. Short variants of the tones, ˥, ˧˥, ˧, i.e., 5, 3, 2, realise on the syllables with a stop ending, such as $(C_1)V:S$ and $(C_1)VS$.

The present study investigates the temporal structures in syllables of different types associated with different tones in Cantonese. Due to page limit, the only temporal data to be discussed in this paper are those of syllables, CV:, CV:S, CVS, CD, and CV:N, where V: = [a], V = [ɐ], D = [ai], and N = [ŋ]. More specifically, the paper shows (1) whether temporal compensation takes place between a syllable-initial consonant, C, and the post-consonantal -V:, -V:S, -VS, -D, and V:N in CV:, CV:S, CVS, CD, and CV:N syllables; (2) whether temporal compensation takes place among the first target vowel, the transition, and the second target vowel in D, i.e., the diphthongs; (3) and whether tone plays a part in determining the temporal structure of the syllables under question.

METHOD

In this study, meaningful Cantonese monosyllables of the types $(C_1)V$:, $(C_1)D$, $(C_1)V:C_2$, and $(C_1)VC_2$, associated with different tones, such as ˥, ˧˥, ˧, ˨˩, ˩˧, and ˨ (i.e., 55/5, 33/3, 22/2, 21, 35, and 24, respectively), were used as test syllables: [pa55], [pa33], [pa22], [pa21], [pa35]; [pʰa55], [pʰa33], [pʰa21], [pʰa35]; [tsa55], [tsa33], [tsa22], [tsa35], [tsa21]; [tsʰa55], [tsʰa33], [tsʰa24], [tsʰa21]; [sa55], [sa33], [sa35], [sa21]; [pak5], [pak3], [pak2]; [pʰak5], [pʰak3], [tsak5], [tsak3], [tsak2]; [tsʰak5], [tsʰak3], [tsʰak2]; [sap5], [sap3], [sap2]; [pɐt5], [pɐt2], [pʰɐt5]; [tsɐt5], [tsɐt2], [tsʰɐt5]; [sɐt5], [sɐt2]. Temporal organization of the test syllables is determined by measurements of the durations of the successive component segments contained in the syllables. Using Kay Elemetrics' CSL 4300B speech analysis software on a 486 PC, durations of the successive component segments in the test syllables were measured directly from the speech waveforms. In the cases where the test syllables contain a diphthong or a nasal ending, durations were measured from the speech waveforms as well as from their formant trajectories. Two native Hong Kong Cantonese college students, one male and one female, provided the speech data. The speakers were instructed to utter each of the test monosyllables in a carrier sentence, [ŋɔ jiu tɔk _____ pei lei tʰɐŋ] "I want (to) read _____ for you (to) listen", at normal rate of speech. The test words in the word-lists were randomised. Five repetitions of each of the test syllables were recorded. The recording was performed in a sound-proof booth.

RESULTS

The mean vowel or diphthong durations (n = 5) in CV: or CD syllables, where C is [p], [pʰ], [ts], [tsʰ], or [s], tend to be similar if the tone on the syllable is 55, 33, 22, 35, or 24; and the mean vowel durations (n = 5) in CV: or CD syllables, where C is [pʰ], [ts], [tsʰ], or [s], tend to be similar if the tone on the syllable is 21.

The overall mean vowel [a] duration in all the CV: syllables, where C is [p], [pʰ], [ts], [tsʰ], or [s], is longer (329.16 ms (n = 85; s.d. = 22.18)) if the tone on the syllable is 55, 33, 22, 35, or 24, and shorter (247.00 ms (n = 25; s.d. = 23.63)) if the tone on the syllable is 21. This is also true for CD syllables. The overall mean diphthong [ai] duration in all the CD syllables associated with tone 21 is 274.30 ms (n = 10, s.d.= 13.44) and the overall mean diphthong duration associated with tones 55, 33, 22, and 35 is 348.07 ms (n = 75, s.d. = 24.45). Such a correlation between duration and tone type is not present in CV:S and CVS syllables, due probably to the fact that the tone on these syllables does not reach the level of 1.

Regardless of the tone type on the CV:S or CVS syllables, the mean vowel [a] or [ɐ] durations in these syllables tend to be similar. The overall mean vowel [a] duration in all the CV:S syllables asso-ciated with tones 5, 3, and 2 is 186.61 ms (n = 70, s.d. = 20.07). The overall mean vowel [ɐ] duration in all the CVS syllables associated with tones 5 and 2 is 120.03 ms (n = 40, s.d. = 10.57).

Regardless of the tone type on the syllables, the mean vowel or diphthong durations in CV:, CD, CV:S, and CVS syllables are shorter if C is an aspirated stop or affricate; or longer if C is an unaspirated counterpart or a fricative. The overall mean vowel duration of all the CV: syllables, associated with tones 55, 33, 22, 24, and 35, is 344.57 ms (n = 35, s.d. = 18.49) if C is an unaspirated [p] or [ts]; 312.87 ms (n = 30, s.d. = 20.03) if C is an aspirated [pʰ] or [tsʰ]; and 344.47 ms (n = 15, s.d. = 18.42) if C is [s]. The overall mean diphthong duration of all the CD syllables, associated wtih tones 55, 33, 22, and 35 is 363.66 ms (n = 35, s.d. = 18.59) if C is an unaspirated [p] and [ts]; 328.04 ms (n = 25, s.d. = 19.04) if C is an aspirated [pʰ] or [tsʰ]; and 346.73 ms (n = 15, s.d. = 20.07) if C is [s]. The overall mean vowel duration of all the CV:S syllables, associated with tones 5, 3, and 2, is 202.63 ms (n = 30, s.d. = 14.72) if C is an unaspirated [p] or [ts]; 169.56 ms (n = 25, s.d. = 15.65) if C is an aspirated

[p^h] or [ts^h]; and 182.53 (n = 15, s.d. = 7.51) if C is [s]. The overall mean vowel duration of all the CVS syllables, associated with tones 5 and 2, is 126.85 ms (n = 20, s.d. = 5.28) if C is an unaspirated [p] or [ts]; 111.00 ms (n = 10, s.d. = 12.55) if C is an aspirated [p^h] or [ts^h]; and 115.40 ms (n = 10, s.d. = 7.07) if C is a fricative. Similar results are obtained for CV:N syllables.

The reduction of the total diphthong durations in CD syllables which are associated with 55, 33, 22, and 35 tones and when initial C is an aspirated stop or affricate is contributed mainly by the reduction of the duration of the first target vowel [a], not the transition or the second target vowel in the diphthong. The mean duration of the first target vowel [a] in the CD syllables, associated with tones 55, 33, 22, and 35 is 197.34 ms (n = 35, s.d. = 14.98) if C is an unaspirated [p] or [ts], and 158.60 ms (n = 25, s.d. = 13.10) if C is an aspirated [p^h] or [ts^h]. The mean duration of the transition in the CD syllables, associated with tones 55, 33, 22, and 35 is 130.77 ms (n = 35, s.d. = 12.45) if C is [p] or [ts], and 135.36 (n = 25, s.d. = 11.88) if C is [p^h] or [ts^h]. And, the mean duration of the second target vowel [i] in the CD syllables, associated with tones 55, 33, 22, and 35 is 35.54 ms (n = 35, s.d. = 12.20) if C is [p] or [ts], and 34.08 ms (n = 25, s.d. = 11.57) if C is [p^h] or [ts^h].

The further reduction of the total diphthong duration in CD syllables, as a result of the effect the 21 tone, where C is an aspirated stop or affricate is contributed by the reductions of durations of the first target vowel and the transition, or the second target vowel in the diphthong.

As for the temporal structure in the diphthongs in those CD syllables, which are associated with tones 55, 33, and 35 and where C is a fricative, [s], the relatively smaller reduction of the total diphthong duration is contributed by both the reductions of durations of the first and second target vowels.

There is a tendency for the syllable-initial aspirated stops and affricates in CV: and CD syllables to have a slight increase in duration if the tone on the syllable is 21, as compared with the durations associated with a non-21 tone. The mean duration of [p^h] in [p^ha] associated with a 21 tone is 93.00 ms (n = 5, s.d. = 13.06), whereas the mean durations of [p^h] in [p^ha] associated with tones 55, 33, and 35 are 69.80 ms (n = 5, s.d. = 5.07), 73.21 ms (n = 5, s.d. = 9.05), and 85.20 ms (n = 5, s.d. = 5.02), respectively. The mean duration of the [ts^h] in [ts^ha] associated with a 21 tone is 140.40 ms (n = 5, s.d. = 20.07), whereas the mean durations of [ts^h] in [ts^ha] associated with tones 55, 33, and 24 are 113.40 ms (n = 5, s.d. = 12.22), 133.00 ms (n = 5, s.d. = 22.66), and 125.00 ms (n = 5, s.d. = 10.40), respectively. Similar results are obtained for CD, CV:S, CVS and CV:N syllables.

The syllable-initial consonant [s] in CV:, CD, CV:S, and CVS syllables has the longest duration, followed by [ts^h], [p^h], [ts], and [p] in descending duration, for example, in CV: syllables which are associated with a 55 tone, the mean durations (n = 5) of [s], [ts^h], [p^h], [ts], and [p] are 144.80 ms (s.d. = 12.44), 113.40 ms (s.d. = 12.22), 69.80 ms (s.d. = 5.07), 54.00 ms (s.d. = 12.69), and 8.40 ms (s.d. = 2.51), respectively

CONCLUSION

Temporal compensation does not seem to take place between syllable-initial consonant and the post-consonantal vowel or diphthong in the Cantonese CV:, CD, CV:S, and CVS syllables. The observation is evidenced by the fact that invariant syllable duration in Cantonese is not maintained, for example, the mean durations (n = 5) of the syllable-initial consonants [p], [ts], [ts^h], and [s] in CV: syllables associated with a 33 tone are 8.40 ms (s.d. = 2.30), 52.80 ms (s.d. = 10.62), 133.00 ms (s.d. = 22.66), and 150.60 ms (s.d. = 14.05), respectively, whereas the mean durations (n = 5) of the following vowel [a] in the syllables are 338.20 ms (s.d. = 14.91), 350.40 ms (s.d. = 18.48), 348.40 ms (9.61), and 353.80 ms (s.d. = 23.78)

The reduction of the vowel or diphthong duration in the syllables where the initial consonant is aspirated is not viewed as an effort of the speaker to maintain invariant syllable duration, rather a result of reduced subglottal pressure caused by the production of aspiration of the syllable-initial consonant.

The further reduction of the duration of the vowel or diphthong in CV: and CD syllables which are associated with a 21 tone and where C is an aspirated stop or affricate is assumed to be contributed by both the reduced subglottal pressure and the 21 tone. The duration of the vowel or diphthong in CV: syllables which are associated with a 21 tone is shorter than the duration associated with a non-21 tone even if the syllable-initial is zero or an unaspirated stop or affricate. This shows that tone does play a part in determining the temporal structure of the syllables. That the vowel or diphthong duration in the CV:, CD, CV:S, CVS, and CV:N syllables where the initial consonant is a fricative, [s], is not reduced or only slightly reduced is probably due to the fact that the production of [s] does not reduce the subglottal pressure for the production of the post-consonantal vowel or diphthong as much as an aspirated consonant does. It seems that temporal compensation takes place among the first target vowel, the transition, and the second target vowel in some syllables, as the total mean durations of the syllables [pai˥] and [pai˦] as well as [pai˧] and [pai˨] are almost the same, although the mean durations of the first target vowel, transition, and the second target vowel differ. The durational data by the female Cantonese speaker have not been presented, although the pattern of the temporal organisation is similar.

REFERENCES

[1] Allen, G. (1973), Segmental timing control in speech production. *Journal of Phonetics*, 1: 219-237.

[2] Bell-Berti, F. & K.S. Harris (1981), A temporal model of speech production. *Phonetica*, 38:9-20.

[3] Cao, J. F. (1992), Temporal distribution of bisyllabic words in Standard Chinese: an evidence for relational invariance and variability for natural speech. *Communication of COLIPS*, 2:58-66.

[4] Crystal, T.H. and A.S. House (1988a), The duration of American-English vowels: an overview. *Journal of Phonetics*, 16: 263-284.

[5] Crystal, T.H. and A.S. House (1988b), The duration of American-English stop consonants: an overview. *Journal of Phonetics*, 16: 285-294.

[6] Fowler, C.A. (1977), Timing Control in Speech Production. Bloomington, Indiana University Linguistics Club.

[7] Fowler, C.A. (1981), A relationship between co-articulation and compensatory shortening. *Phonetica*, 38:35-50.

[8] Kozhevnikov, V.A. & Chistovich, L.A. (1965), Speech: Articulation and Perception. Washington, DC., U.S. Department of Commerce, Joint Publications Research Service, No. 30: 543.

[9] Lehiste, I. (1971), Temporal organisation of spoken language. In L.L. Hammerich, et al, eds., *Form & Substance: Phonetic and Linguistic Papers Presented to Eli Fischer-Jørgensen*, pp. 159-169.

[10] Lehiste, I. (1972), The timing of utterances and linguistic boundaries. *Journal of Acoustical Society of America*, 51:2018-2024.

[11] Lindblom, B.E.F. (1968), Temporal organization of syllable production. *STL-QPSR* 2-3, 1968, 15.

[12] Lindblom, B.E.F. & K. Rapp (1973), Some temporal regularities of spoken Swedish. *Papers in Linguistics from the University of Stockholm*, 21:1-59.

[13] Nooteboom, S.G, (1972), *Production and Perception of Vowel Duration*. Utrecht, Utrecht University Dissertation.

[14] Peterson, G. & E. Lehiste (1960). Duration of syllable nuclei in English. *Journal of Acoustical Society of America*, 32:693-703

[15] [*Phonetica*, 1981]

[16] Port, R.F., S. Al-Ani, & S. Maeda (1980), Temporal compensation and universal phonetics. *Phonetica*, 37: 235-252. [End]

ON THE PERCEPTUAL CLASSIFICATION OF SPONTANEOUS AND READ SPEECH

Eleonora Blaauw
Research Institute for Language and Speech
Utrecht, The Netherlands

ABSTRACT

What prosodic cues do listeners use to classify speech as read or spontaneous? Two different types of spontaneous speech and matching read samples were examined. Results suggest that read speech is characterized by a typical prosodic make-up. Spontaneous speech seems to be characterized by the absence of that typical prosodic make-up; digressions from the 'read' prosody, in any direction, appear to induce listeners to classify the speech as spontaneous.

INTRODUCTION

When we hear someone speak, it is intuitively very easy to tell whether the speaker is talking spontaneously, or whether he is reading a text out loud. Research has confirmed this informal observation, and has shown that prosodic cues are important for the perceptual distinction [1,2].

The spontaneous-read distinction is not as simple and straightforward as it might seem at first glance. One can distinguish many different spontaneous and read styles, for example along a formal-informal dimension, or a careful-casual dimension [3]. These interfering dimensions do not seem to confuse listeners in making spontaneous-read judgments, however. This suggests that the spontaneous-read distinction constitutes a basic and meaningful difference to listeners. In addition, the cues that they use must be retrievable from different types of spontaneous and read speech. In this paper we aim to identify some of these cues. As it has already been shown that prosody plays an important role for the spontaneous-read distinction, we will limit our search to prosodic cues.

In looking for reliable prosodic cues to the spontaneous-read distinction, we should concentrate on characteristics that reflect fundamental underlying differences between spontaneously produced speech and speech read from text. After all, such characteristics should surface in any type of spontaneous and read speech, which renders them reliable.

Spontaneous speech has several fundamental characteristics that distinguish it from read speech, no matter how formal or informal the situation in which it is produced, or how careful or casual the produced speech. Spontaneous speech is produced impromptu, on the spot, which entails that much planning activity is required on the part of the speaker. In addition, it entails that spontaneous speech is highly flexible, and can be optimally adapted to the communicative situation. Read speech is largely prepared beforehand, at the stage where the text is written. The planning required from the speaker is therefore limited. In addition, the possibilities to adapt the speech to the communicative situation is limited; at the actual time of production, read speech is much less flexible than spontaneous speech.

In this paper we will focus on the fundamental difference in flexibility between spontaneous and read speech. It seems plausible that this difference affects the prosodic characteristics of spontaneous and read speech in a distinctive way, so that those characteristics form reliable cues for the perceptual classification of spontaneous and read speech.

The larger flexibility in spontaneous speech production is likely to be reflected in a highly flexible and variable prosody, strongly dependent on the communicative situation in which the speech is produced. The prosodic characteristics of a spontaneous intimate conversation between two close friends are bound to differ greatly from the prosodic characteristics of a spontaneously produced formal speech.

The lesser amount of flexibility in read speech may result in less variable prosody across different samples of read speech. A speaker reading a personal letter out loud to a friend or reading a speech out loud during a formal meeting will probably produce two speech samples with fairly similar prosodic characteristics. In addition, these prosodic characteristics are likely to be 'neutral', positioned somewhere in the middle of all possible spontaneous characteristics. Read speech is typically clear and careful, whereas spontaneous speech can easily digress in any direction. It can be casual and sloppy for example, but it can also be emphatic and highly expressive. It can be slow and hesitant, but it can also be produced at very high rates.

We think that listeners have little trouble in identifying all those different types of spontaneous speech correctly as spontaneous, despite the fact that the prosodic make-up of the speech varies enormously. Or perhaps we should say, thanks to the fact that the prosodic make-up of the speech varies enormously. If read speech indeed shows fairly stable, idiosyncratic prosodic characteristics, telling spontaneous and read speech apart becomes an easy task. Whenever the speech listeners are presented with exhibits those typically read prosodic characteristics, the speech can be classified as read. If the speech shows digressing prosodic characteristics, in any direction, the speech can be classified as spontaneous.

In summary, we hypothesize that spontaneous speech shows vastly different prosodic characteristics, dependent on the situation in which it is produced, whereas read speech shows relatively stable, average prosodic characteristics. Furthermore we hypothesize that listeners can classify different types of spontaneous and read speech correctly. They may do this by classifying speech with typically read characteristics as 'read aloud', and speech with characteristics that digress from the read values as 'spontaneous'.

To test these hypotheses, two speech corpora were selected. The first corpus consisted of casual spontaneous speech produced in an informal interview situation, and matching read speech (i.e. a read version of the interview based on a transcript of the spontaneous speech). The second corpus consisted of careful spontaneous speech, viz. so-called instruction monologues [4], and matching read speech. A selection of fluent spontaneous utterances and their read counterparts was made from both corpora. Classification judgments were elicited for those utterances. Subsequently, the values of four prosodic parameters were established for each of the selected utterances. These prosodic characteristics were then correlated with the percentage of 'spontaneous' judgments obtained from the listeners.

THE INTERVIEW CORPUS

The interview corpus consisted of one and a half hours of spontaneous speech produced by one male speaker, and a read version of large parts of the original interview produced by the same speaker, read from a written transcript. From this corpus 48 fluent spontaneous utterances and 48 matching read utterances were selected. In a perception experiment, each individual utterances was presented to 10 listeners, who were asked to classify each utterance as spontaneous or read. The average classification score was 79% correct (81% for the spontaneous utterances and 77% for the read utterances).

For each utterance we determined mean F0, standard deviation of F0, F0 range (both measures of the amount of F0 variation), and articulation rate. F0 range was defined as the distance between the lowest and the highest F0 value in each utterance. Standard deviation of F0 and F0 range are expressed on a logarithmic scale, in semitones. Articulation rate was defined as the number of syllables per second, excluding pause time.

Table 1: Acoustic characteristics of spontaneous and read utterances selected from the interview corpus; correlation (Pearson's r) between prosodic characteristics and percentage of 'spontaneous' judgments.

	spont.	read	r
mean F0 (Hz)	132	157	-.63
s.d. F0 (ST)	2.0	2.6	-.51
F0 range (ST)	8.2	10.4	-.48
art. rate (syll/s)	7.1	6.3	.52

In addition, Pearson's correlation was determined between the prosodic characteristics and the percentage of

'spontaneous' judgments for each utterance. The results are shown in Table 1.The differences between the two speech styles, as determined with a paired t-test, is significant at the 1% level for all four prosodic parameters. The t-values (df = 47) are 11.5, 5.3, 5.0, and -6.7 respectively. The correlation coefficients are significant at the 1% level as well. Thus, a lower mean F0, a smaller standard deviation of F0, a smaller F0 range and a higher articulation rate are significantly associated with more 'spontaneous' judgments. For more details on the collection and characteristics of this corpus the reader is referred to [5].

THE INSTRUCTION MONOLOGUE CORPUS

Spontaneous instruction monologues were collected from five male speakers. They were asked to give instructions to a listener on how to assemble the front view of a house from a set of cardboard pieces. Both speaker and listener had the same set of building blocks in front of them. They could not see each other, and the speaker did not receive any feedback from the listener. The monologues each lasted about five minutes. The spontaneous monologues were transcribed orthographically, and subsequently each monologue was read aloud by the original speaker. From this corpus 109 fluent spontaneous utterances and the 109 matching read counterparts were selected, divided over the five speakers. For more details on the collection of the corpus, the reader is referred to [6].

In a perception experiment, the selected utterances were presented individually to 21 listeners, who were asked to classify each utterance as spontaneous or read. The average classification score was 77% correct (79% for the spontaneous utterances and 75% for the read utterances).

For the utterances from the interview corpus we also determined mean F0, standard deviation of F0, F0 range, and articulation rate for each utterance. These measures were defined and determined in the same way as for the interview corpus (see above). The results are presented in Table 2. In addition, Pearson's correlation was determined between the prosodic characteristics and the percentage of 'spontaneous' judgments for each utterance. To compensate for absolute differences between the five speakers, z-scores were used for the correlation study. These z-scores were calculated separately for each speaker, across both speech modes. The correlation coefficients are shown in the last column of Table 2.

Table 2: Acoustic characteristics of spontaneous and read utterances selected from the instruction monologues; correlation (Pearson's r) between prosodic characteristics (z-scores) and percentage of 'spontaneous' judgments.

	spont.	read	r
mean F0 (Hz)	128	122	.27
s.d. F0 (ST)	3.2	2.9	.19
F0 range (ST)	13.9	11.7	.31
art. rate (syll/s)	5.2	5.8	-.58

The results show a reversal of the spontaneous-read differences in comparison to the interview corpus for all four prosodic variables The difference between the speech styles, as determined with a paired t-test, is significant at the 1% level for all four prosodic parameters. The t-values (df = 108) are 4.5, 4.2, 5.8 and -7.9 respectively.

A comparison between Tables 1 and 2 leads to the following observations. The characteristics of the read speech samples are, as predicted, fairly similar in both corpora, and intermediate between the values for the spontaneous interview and the spontaneous instruction monologues. Thus, the values for the spontaneous speech samples can be said to digress from the stable 'norm' values for the read speech. Mean F0 forms an exception; its value in the read samples does not lie between the values in the spontaneous samples. This is due to the fact that mean F0 is highly speaker-dependent; we did not use the same speakers in both corpora. In order to make the mean F0 values from both corpora comparable, they should be standardized, for example by expressing them in terms of the distance to the bottom of the speaker's range. We did not have the necessary data to carry out this standardization. For now, we will just assume that, had we used the same speakers, mean F0 would have shown the same pattern as the other prosodic parameters.

Thus, the production results seem to confirm the hypothesis that read speech shows stable prosodic characteristics, whereas the prosodic characteristics of spontaneous speech digress from the read characteristics in any direction.

The correlation coefficients for mean F0, standard deviation of F0 and F0 range are much smaller than they were in the interview corpus. Nevertheless, all four correlation coefficients are significant at the 1% level. Moreover, they are all reversed in comparison to the interview corpus. Thus, in this corpus, a higher mean F0, a larger standard deviation of F0, a larger F0 range and a lower articulation rate are associated with more 'spontaneous' judgments.

CONCLUSION AND DISCUSSION

The experiments described in this paper showed that listeners are able to identify different types of speech correctly as spontaneous or read. The prosodic characteristics of the two spontaneous samples showed large differences, whereas the two read samples both showed more or less the same average prosodic characteristics. The stronger the prosodic characteristics of an utterance digressed from these 'read' values, the larger was the percentage of 'spontaneous' classification judgments from the listeners.

It would be premature to conclude that all read speech shows typically read values on a whole range of prosodic parameters, by which a listener can recognize the speech as read. First of all, we only looked at a few prosodic parameters. Second, the method by which the two read samples used in the present study were collected biases the results towards this conclusion. Although the texts were based on different types of spontaneous speech, the settings in which the read samples were recorded were similar. Both read samples were examples of straightforward laboratory readings of a coherent running text. This is inevitable when one wants to collect matching spontaneous and read speech. However, future research should include read speech collected outside the laboratory, in different communicative settings. Possibly, such read samples will show larger prosodic differences than the present read speech samples. We maintain, however, that the lack of flexibility in read speech production will seriously limit the possible variation in prosodic characteristics. In some special cases this limitation may be overcome, for example when 'reading' a thrilling story to a child. In such a speech sample the prosodic characteristics will digress strongly from the average read values. However, we imagine that in a listening test such speech material would not be classified as read, but as enacted, or in some cases even as spontaneous speech.

REFERENCES

[1] Levin, H., Schaffer, C. and Snow, C. (1982), "The prosodic and paralinguistic features of reading and telling stories", *Language and Speech*, Vol. 25, pp. 43 - 54.

[2] Laan, G. & Van Bergem, D. (1993), "The contribution of pitch contour, phoneme durations and spectral features to the character of spontaneous and read aloud speech", *Proceedings Eurospeech '93, Berlin*, Vol.1, pp. 569-572.

[3] Eskénazi, M. (1993), "Trends in speaking style research", *Proceedings Eurospeech '93, Berlin*, pp. 501-505.

[4] Terken, J. (1984), "The distribution of pitch accents in instructions as a function of discourse structure", *Language and Speech*, Vol. 27, pp. 269-290

[5] Blaauw, E. (1992), "On the perceptual difference between read and spontaneous speech", In: M. Everaert, B. Schouten and W. Zonneveld (eds.), *OTS Yearbook 1992*, Utrecht: LEd, pp. 1-16.

[6] Blaauw, E. (1994), "The contribution of prosodic boundary markers to the difference between read and spontaneous speech", *Speech Communication*, Vol. 14, pp. 359-375.

EVALUATION OF DISCOURSE STRUCTURE ON THE BASIS OF WRITTEN VS. SPOKEN MATERIAL

Monique E. van Donzel & Florien J. Koopmans-van Beinum
Institute of Phonetic Sciences/IFOTT, University of Amsterdam, The Netherlands

ABSTRACT

In this paper we present the results of an experiment in which firstly we asked text analysts to evaluate the verbatim transcriptions of a retold story in terms of 'informational structure', using a method [2] based on linguistic knowledge and intuition. We then had listeners underline emphasized words and scale them for degree of emphasis in the spoken versions of the same story, but on the basis of the speech sound only. The prediction was that in the latter case linguistic knowledge may be overruled by the actual speech sound. Results show that this indeed seems to be the case.

INTRODUCTION

The structure of information in *written* texts usually becomes clear by the use of typographic means. In *spoken* texts it is generally assumed that the speaker may use various acoustic means to assign structure, for instance by accenting important words. In written texts words can also be perceived as being more or less important, in this case there is evidently no relation with accents.

In the often used elicitation method of *question/answer pairs* the informational structure (focus distribution) can be described using the labels 'new' vs. 'old' information, where 'new' usually refers to 'accented' and 'old' to 'not accented'. Focus is thus defined through intonation. However, this kind of definition may lead to circularity in that the possible acoustic features are already included in the definition itself.

How the focal structure of a whole *discourse* should be traced is less clear. Therefore, we developed a method [2] using various theories about discourse structure, in which the focal structure of a text is based on the informational structure rather than on the acoustic features, thus avoiding the circularity mentioned above.

The goal of our experiment was to see if there is a relation between the informational structure, based on linguistic knowledge, and prominence judgements of listeners based on the speech sound. Possible differences between different speaking styles and between sexes are discussed as well.

METHODS

Speakers, text analysts, and listeners

Four male and four female speakers, all native speakers of Dutch, were selected as speakers for the experiment. They were all students or staff members of the Institute of Phonetic Sciences. Five text analysts, all familiar and experienced with text analyses, participated in the evaluation of the written text. The speakers and text analysts participated on a voluntary basis. Seven male and nine female students and staff members of the University of Amsterdam, all native speakers of Dutch, participated as listeners in the experiment. The student listeners were paid for their participation.

Stimuli and recordings

The speakers were asked to read aloud a short story in Dutch (*Een triomf* by Simon Carmiggelt). After a short break they were asked to tell the same story in their own words, as detailed as possible (the 'retold' version). During the retelling of the story, a listener was present in the recording room, to create a natural telling situation. From this retold version a verbatim transcription was made by the first author, and the speaker was asked to read aloud this transcription the next day (the 're-read' version). All recordings were made in a sound treated room on DAT-tape.

Method of text analysis

In this section we will briefly present the method used to analyse the informational structure of the recorded discourses. This method is a combination of several theories about the structure of discourses [1, 3, 4]. Because of space limitations, we will discuss here only the labels at the word level.

Nominal constituents can be classified as follows, using so-called 'textual labels'. A *brand new* (bn) element refers to information that is completely new in the listener's context. This usually regards indefinite nouns or generic expressions. An *unused* (u) element is also new, but the listener can place the information it expresses directly in his/her discourse model. This are usually definite nouns or proper names. An element is labeled as *inferrable* (i) if the speaker assumes that the listener can infer it from the preceding context or from his/her knowledge of the world. *Evoked* elements have already been mentioned in the discourse. They can be I) *textually evoked* (et): the noun is evoked by a real pronoun, II) *displaced textually evoked* (etd): the noun cannot be evoked by a pronoun because the referent is too far back in the discourse, the full noun is used to avoid ambiguity, III) *situationally evoked* (es): the referent of a noun or pronoun can only be found in an extra textual context. *Modifiers* (mod) express some kind of degree or quality. *Orientations* (or) express temporal or locational orientations at the beginning of clauses.

Verbs are classified using the labels *unused*, *inferrable* and *evoked* in the same way as for nominal constituents. The verb phrase as a whole is labeled, the auxiliary and the main verb are considered as a unitary concept. Prepositions which are part of a verb are related to them by giving an index to both of them.

Written evaluation

The informational structure ('focal structure') of the transcribed retold versions of the four male and four female speakers was evaluated using the method described in section 2.3. The analyses were made by the first author. These analyses were presented to a panel of five text analysts, all familiar with discourse theories. The proposed text analyses were discussed and this resulted in an ultimate convention for labeling. Where necessary the proposed analyses were adapted. An example of parts of one of the texts and its analysis is presented in Table 1.

Table 1. Example of a text analysis.

het [es] eeh gaat [u] over twee mensen [bn] die wonen [u] in de stad [u]
en op een morgen [or] worden$_1$ ze [et] wakker$_1$ [u]
en dan [or] zien [u] ze [et] dat het heel hard [mod] gesneeuwd [u] heeft [i]
het [es] is dus een verhaal [bn] in de winter [i] [ai]
en ze [et] besluiten [u] om die dag [i] eens in het bos [u] te gaan kijken [u]
hoe het [et] er dan daar [et] uit ziet [i]
de stad [etd] uit het bos [etd] in
in het bos [etd] is het eeh heel heel dik [mod] besneeuwd [e]
de takken van de jonge bomen [i] die buigen$_1$ [u] over$_1$
en daar [et] moeten ze [et] soms [mod] onderdoor$_1$ kruipen$_1$ [u] ...

Perceptual evaluation

The 16 listeners were instructed to evaluate the retold versions and the re-read versions in terms of prominence, using only the speech signal which was presented over headphones. Each listener was presented with an individual tape containing four different spoken versions of the story (the first text was used as an exercise), either a retold version or a re-read transcription, from four different speakers. They were asked to underline the parts of the discourse they perceived as being emphasized by the speaker, on the basis of the speech sound only, so explicitly *not* on the basis of the written text, and then to judge the relative prominence of these parts on a scale from 1 (very emphasized) to 3 (less emphasized). These marks do however not necessarily represent the linguistic terms of primary, secondary and ternary stress. The verbatim transcription of the spoken text was used as an answer sheet. There was a two hour time limit to the task.

RESULTS

Textual structure and perceptual prominence

Each text was evaluated by three different listeners. For each of the eight verbatim transcriptions the analysis based on the text alone was taken as reference point. The perceptual judgements were compared to these analyses. For every text, style and listener a confusion matrix was made, in which the labels from the text analysis were matched against the

prominence judgements 1, 2 and 3. 'Zero labels' (0) were added to cover the cases in which a word was underlined but no judgement was given (zero perception) and the ones in which a word was underlined that did not have a proper label in the text analysis (zero text analysis). This resulted in 48 matrices (8 speakers x 2 styles x 3 listeners per text).

Overall matrix

To get a first impression of how the textual analysis might be related to the perceptual analysis, we normalized to percentages and summed all 48 matrices (Table 2). The three perceptually most relevant labels are *unused* (22%), *brand new* (17%) and *modifier* (16%). This is as can be expected since these labels represent words containing 'new' information. Thus, 55% of all underlined parts were 'new' items in the discourse ($p \leq 0.001$, df=1, χ^2=48.4).

When looking at labels referring to 'given' information, we find the following: *evoked textually* (8%), *evoked textually displaced* (14%) and *evoked situationally* (1%). Again, these relatively low percentages, apart from *etd,* can be expected, since evoked items will generally not be pronounced with much emphasis. However, *the evoked textually displaced* items seem to be perceived as more emphasized than other evoked items. This is not surprising either, since it is exactly these items that cannot be pronominalized, they have to be 'refreshed', and thus are 'new' in a certain sense. For example, 'the forest' is referred to at a later point in the discourse not by means of the pronoun 'it' but by repeating the full noun 'the forest' to avoid ambiguity.

Table 2. Overall matrix, normalized.

	0	1	2	3	total
or	0.00	0.51	0.96	0.34	1.82
mod	0.06	4.34	6.96	4.25	15.61
bn	0.00	5.14	7.69	4.55	17.37
u	0.08	6.43	9.95	6.03	22.49
i	0.04	3.97	6.27	3.73	14.01
et	0.04	1.59	3.95	2.59	8.16
etd	0.05	4.07	6.39	3.81	14.32
es	0.00	0.24	0.20	0.26	0.70
0	0.04	1.50	2.16	1.81	5.52
total	0.31	27.78	44.54	27.36	100%

The *inferrable* items represent information that is neither completely new nor completely evoked. From the parts perceived as emphasized, 14% is inferrable. This might suggest that this category is indeed a valid one in the analysis. The 'rest' group (7%) consisted of the items *orientation* (or) and zero judgements (0).

When looking at the relative prominence judgements (1, 2, or 3), we find that 28% of all items are judged with a *1*, 45% with a *2*, 27% with a *3* and 0,3% did not have a perceptual judgement. This indicates that listeners did use the whole scale of possibilities.

This first look at the data suggests that there does seem to exist a relation between the textual analysis and the overall prominence judgements of listeners. Elements that add new information to the discourse are perceived as emphasized more often than elements representing information that is already evoked earlier in the discourse. Information that can be inferred from other elements in the discourse is also perceived as emphasized in a number of cases. However, listeners do not seem to give a particular judgement (1, 2, or 3) to a particular textual label (or, mod, bn, etc.); so there does not seem to be a clear correlation between a certain judgement and a certain textual label. In almost half of the cases listeners judged a 2 ($p \leq 0.001$, df=1, χ^2=35.6), which may indicate that only in extreme cases a 1 or a 3 was judged. Therefore, in the rest of this paper we will take into account only the total percentage of judgements.

Differences between speaking styles and between sexes

In this section we will look at possible differences between the two speaking styles, and between the ways in which male and female speakers are perceived.

The first two columns of Table 3 present the overall percentage of judgements, for the retold and re-read speaking styles. There do not seem to be very large differences between the two styles; they differ at most 2%, these effects do not appear to be significant. We expected larger differences between the two speaking styles, since they are perceptually quite distinct. However, whenever the retold speaking style dominates, this is exactly for the major categories from Table 2 (*brand new, unused, inferrable* and *evoked textually displaced*). This might follow from the fact that the method of text analysis is developed from discourse theories based on spontaneous speech.

The last two columns present the overall percentage of judgements, for the male and female speakers separately. In some cases, the male and female speakers behaved differently. As for the major categories, the male speakers scored higher than the female speakers. The female speakers, however, emphasized much more *modifiers* than did the male speakers. This might suggest that the female speakers had a more elaborate way of telling, while the male speakers were more 'compact'.

Table 3. Overall percentage judgements, broken down for retold/re-read speaking style and for male/female speaker.

	retold	re-read	male	female
or	1.44	2.19	1.77	1.84
mod	14.34	16.89	12.94	19.53
bn	18.36	16.38	18.82	14.65
u	23.06	21.91	23.25	21.29
i	14.89	13.14	14.64	12.80
et	7.39	8.93	7.79	9.20
etd	15.32	13.33	15.05	13.63
es	0.68	0.73	0.71	0.83
0	4.52	6.51	5.04	6.25
total	100%	100%	100%	100%

Finally, something has to be said about the so-called 'zero judgements'. Overall, they cover about 5% of all labels, meaning that 5% of the words underlined by the listeners did not have a textual label or no judgement was given, and thus could not be classified. At a closer look, these words appeared to be mainly discourse markers (*well, thus, so,* etc.) or discourse connectives (*and, or,* etc.). However, cases in which an auxiliary was perceived as emphasized without the main verb being perceived as such, fall in this category as well.

DISCUSSION

In this section we will try to test our hypothesis that linguistic knowledge may be overruled by the actual speech sound in assigning structure to spoken texts.

The data show clear evidence for the three major categories *new, inferrable* and *evoked.* New words are expected to be perceived as being emphasized. Inferrable and evoked words, however, are not expected to be perceived as being emphasized so often, since these words represent information that is known at some level.

When looking at our results, we find that in exactly these cases there is a difference between the expected data and the observed data, especially when regarding the *inferrable* and the *evoked textually displaced* items: these are perceived as emphasized quite often. This indicates that in these cases, the actual speech sound does overrule linguistic knowledge, since emphasis is not expected.

Furthermore, the method of text analysis will need to be extended to *discourse markers,* to account for a part of the zero judgements, and to so-called *'contrastive accents'* to account for the occurrence of, among other things, emphasized auxiliaries.

ACKNOWLEDGEMENTS

The authors would like to thank Rob van Son for his help in processing the matrices, and Louis Pols for comments on the paper in general.

REFERENCES

[1] Chafe, W. (1987) 'Cognitive constraints on information flow', in: R.S. Tomlin (Ed.) *Coherence and grounding in discourse*, John Benjamins, Amsterdam/ Philadelphia, pp. 21-51.

[2] Donzel, M.E. van (1994) 'How to specify focus without using acoustic features', *Proceedings* **18**, Institute of Phonetic Sciences, University of Amsterdam, pp. 1-17.

[3] Mann, W.C. & S.A. Thompson (1988) 'Rhetorical Structure Theory: Toward a functional theory of text organisation', *Text* **8** (3), pp. 243-281.

[4] Prince, E. (1981) 'Toward a taxonomy of Given-New information', in: P. Cole (Ed.) *Radical Pragmatics*, Academic Press, New York, pp. 223-255.

SWEDISH CONSONANT CLUSTERS IN SPONTANEOUS SPEECH: PRELIMINARY ACOUSTIC-PHONETIC OBSERVATIONS

Robert Bannert
Department of Phonetics, Umeå University, Sweden

ABSTRACT

Compared to other parts of the Swedish phonological system, consonant clusters have not been studied sufficiently. This is especially true of the phonetic aspects. Therefore, the research programme "Variations within consonant clusters in spoken Swedish (VaKoS)" was launched, aiming at highlighting their occurrence and distribution in spontaneous speech, their temporal variations and the phonological processes operating on them. Preliminary data from the spontaneous speech of one female Standard Swedish speaker is presented.

INTRODUCTION

From a typological and universal point of view, Swedish is characterized by a relatively complex syllable structure. Within a morpheme, the syllable nucleus may be preceded and followed by three consonants, e.g. *straff* (penalty), *växt* (plant). Within the syllable as the phonotactic domain, five consonants may follow, e.g. *skälm+sk+t* (inflected form of roguish). Consonant clusters of eight consonants, at least in the canonical forms, can arise across word boundaries, e.g. *skälmskt skratt* (roguish laughter). However, in fluent speech, some consonants are normally deleted. The prosodic feature of quantity intersects in an intriguing way with consonant clusters in morphemes and syllables. Shedding light on this important area of Swedish phonology that has not yet received due attention, is the goal of the project "Variations within consonant clusters in spoken Swedish (VaKoS)".

The project aims at describing the occurrence, distribution and temporal variation of consonant clusters in phonetically controlled and spontaneous Standard Swedish speech. Phonological processes operating on these clusters will also be studied. Speech samples of one hour's length of five male and five female Standard Swedish speakers of similar background will be collected in the data base DUKoS (Databasen i Umeå för svenska konsonantgrupper). The investigation will also include perceptual experiments.

BACKGROUND

Phonotactic aspects of consonant clusters are treated in [1]. The domain of analysis is the word and a generative programme for possible consonant clusters in Swedish is developed. Compared to the morpheme as the domain of analysis, this approach yields larger clusters.

In a classic work [2], an attempt to write rules for phonological processes operating on consonant clusters was made. In contrast to [1], spontaneous speech was the goal of that investigation. It also treated clusters that arise across word boundaries.

Variations within consonant clusters from a phonological point of view are treated in [3]. Typical processes are deletion, assimilations and retroflexation. However, experimental studies concerning spontaneous speech and quantitative aspects of consonant clusters including phonological processes have not been conducted.

Experimental data concerning the temporal pattern of variation of consonant clusters consisting of /s, t, k/ in initial, medial and final morpheme position in phonetically controlled utterances were presented by [4]. A linear increase of consonant and cluster durations of about 40 ms due to focus accent applies to all positions.

GOAL

The goal of the present study is to collect preliminary acoustic-phonetic data on Swedish consonant clusters in spontaneous speech. Three main aspects are dealt with: first, the frequency of occurrence of consonant clusters with respect to word and morpheme boundaries. Second, the distribution of phonological processes effecting consonant clusters and their phonetic manifestations. Third, the segmental composition of the clusters.

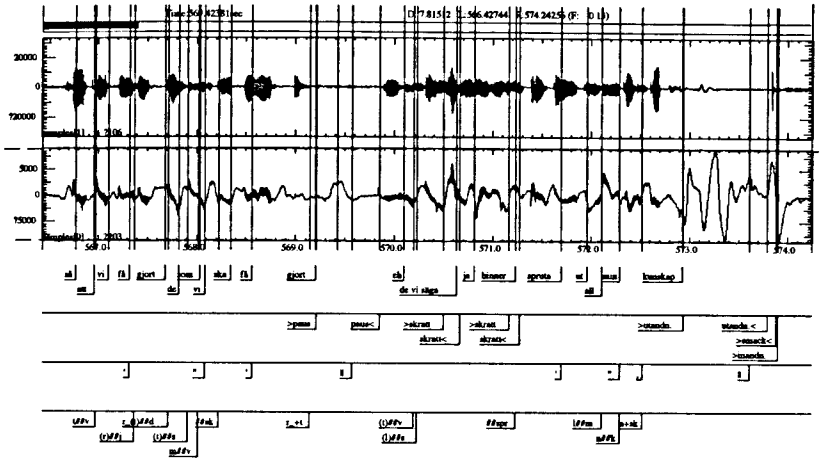


Figure 1. Illustration of processing the speech material. From top: speech wave, laryngograph signal, four levels of labelling.

MATERIAL

Two adult speakers, two males and two females, aged between 30 and 40, from the Stockholm area were recorded. The spontaneous speech of each speaker, approximately 60 minutes each, was organized into four different topics with the duration of about 15 minutes each. The topics included summer holiday, cooking, literature and body language. Recordings of the speech wave and the electro-laryngograph signal were made simultaneously. This signal makes it possible to clearly observe the actions of the glottis during consonant production. The preliminary data presented here comprises one topic about body language by the female speaker. A passage at the beginning of the text is shown:

"... å för mej blir de allså väldit väldit jobbit att försöka dels hålla tråden *så att vi få gjort de som vi ska få gjort eh de vill säga ja hinner spruta ut all min kunskap* å dels att ja hinner avläsa inte bara på dom som sitter längst fram ..."

The text is represented in a simplified orthography expressing the way of pronunciation in this spontaneous speaking style. The following prosodic markers are inserted:

Phrase boundaries:
‖ strong phrase boundary
| weak phrase boundary
Prominence levels:
"CV focus accent
'CV primary stress (word accent)
,CV secondary stress

ANALYSES

In a preparatory phase, the material was labelled on a SUN Sparc2 station using ESPS/waves+. Four levels of labelling were selected: (1) an orthographic representation as close as possible to actual pronunciation, (2) non-verbal behaviour like laughter, clicks, in- and exhalations, (3) prosodic features of phrase boundaries and four levels of prominence (see above) and (4) consonant clusters. For the first level, we found methodological inspiration in the work of the Gothenburg linguists [5]. Figure 1 gives an illustration of a short passage contained in the text sample given in italics above.

When specifying the consonant clusters in the speech sample, we started out from the underlying or canonical word form. Retroflex consonants in morphemes are analysed as single segments (phonemes). Thus 'bord' (table) contains an initial and only one final consonant. Retroflex consonants, however, arise as the result of the phonological process of retroflexation across morpheme and word

boundaries, e.g. 'smor+d' (greased), 'spar+sam' (economical), 'vår#dag' (spring day).

Together with the segmental composition of the clusters, morpheme and word boundaries are indicated. We found it necessary to differentiate between a simple word boundary within a compound word (mat#sal, 'dining-room') and a double word boundary delimiting words (##fönster##, 'window'). Sometimes it was difficult to decide the morphological status of word elements. Numerals like 'sexti' (sixty), words like 'månda' (Monday) and 'därför' (because) were treated as one morpheme. The reason for this decision is phonological and morphological: these words have the acute accent and for the language user of today these words cannot be divided into meaningful parts.

RESULTS

The preliminary results of this introductory study are presented in three groups: (1) occurrence of the consonant clusters expressed as their frequency, (2) the phonological processes, (3) their segmental composition.

Occurrence and frequencies

The speech sample studied had the length of about 12 minutes. It contained 1.000 consonant clusters. The ten most frequent consonant clusters are given under three conditions: (1) total in Table 1, (2) in morphemes (initially, medially, finally) in Table 2, (3) across word boundaries in Table 3. The first column in each table givs the frequency of occurrence, the second the consonant cluster in IPA notation and the third an example context. Segments in parentheses are deleted in the speech signal.

Table 1. Overall occurrences of consonant cluster. Total: 1.100.

39	(t)##d	att de
37	nt	inte
24	(r)##m	här med
23	nd	andan
17	st	fäste
17	n##s	den som
16	t##m	att medge
15	(t)##v	det var
14	(t)##s	det som
13	(r)##j	menar jag

Table 2. Consonant clusters in morphemes. Initial: total 57

10	st	stora
9	sk	skola
8	spr	språk
5	gr	grand
3	kr	kroppen
3	bl	blev
2	skr	skrift
2	pr	producera
2	fr	från
1	tr	tror

Medial: total 214

37	nt	inte
25	nd	andan
17	st	fastnat
8	rj	börja
8	ntr	intresse
8	ŋk	tanken
7	ld	ålder
6	ntl	egentligen
6	ndr	andra
5	kt	kt

Final: total 19

5	sk	grammatiska
4	st	bästa
2	ʂt	första
2	nt	instrumenten
2	nd	band
1	rk	starka
1	rb	verb
1	nt	instrumenten
1	ns	minns
1	kt	dialekt

The number of clusters within morphemes varies greatly between positions. The largest number is to be found in medial position, the smallest in final.

Table 3. Consonant clusters across word boundaries. Total: 620

39	(t)##d	att de
24	(r)##m	där med
17	n##s	kan se
16	t##m	att man
15	(t)##v	det var
14	(t)##s	det som
13	(r)##j	försöker ge
12	(r)##d	är det
12	n##m	kan många
11	m##d	dom där

The frequency distribution of the consonant clusters in the three contextual categories above is rather similar. While 252 clusters appear only once in the speech sample, 64 clusters only twice, 35 clusters only three times, etc., one cluster is to be found 37 times and the most frequent one 39 times (cf. Table 1).

Phonological processes

By far the most frequent phonological process observed is consonant deletion, in total 360 instances. Final /t, r/ are deleted most frequently. Table 4 gives all the deletions found in the speech sample.

Table 4. Consonant deletions and their frequency. n = 360.

t	126	h	6	k	2
r	104	l	6	p	1
g	44	j	5	ɖ	1
d	36	gt	5	lt	1
n	9	m	3	sk	1
v	7	s	3		

Segmental composition

Consonant clusters within morphemes will be used in order to illustrate their segmental composition. Table 5 gives an overview showing 2- and 3-consonant clusters, grouped according to the features obstruent, nasal and liquid.

Table 5. Segmental composition of 290 morpheme internal consonant clusters. Two and three segments.

Initial		Medial		Final	
2	3	2	3	2	3
	spr				
st		st		st	
sk	skr			sk	
pr					
kr				kt	
gr					
		nt	ntr	ʂt	
		nd	ndr	nt	
		ŋk		nd	
		nʃ		ns	
bl		ld	ntl	rk	
		rj		rb	

Three consonant clusters are only found in initial and medial position. In initial position, combinations of obstruents and /r/ dominate, while medially most clusters contain a nasal. In final position, pure obstruent clusters and combinations of nasal or /r/ followed by an obstruent appear as well.

CONCLUSIONS

We have developed a method and procedure that ensure a quantitative and qualitative processing of spoken Standard Swedish with respect to the analysis and phonetic description of consonant clusters. The largest difficulty encountered is morphological segmentation. We decided to give priority to the synchronic aspect and to disregard the diachronic perspective. The preliminary results of our project work are promising. The data on the occurrence and frequency distribution of the consonant clusters, the distribution of the various phonological processes and the specification of their segmental composition will contribute substantially to our knowledge of the phonotactic structure of Standard Swedish. Our results will also deepen our insights into the typological dimension of consonant clusters. Furthermore, this is of great interest to experimental phonology. One side-effect of our work, due to excellent computer facilities, will be a corpus of spontaneous speech labelled for prominence, morpheme, phrase and word boundaries, consonant clusters and even non-verbal signals. At the end, it will be available on CD-ROM. We plan also to publish the first frequency word list of spontaneous Swedish, sorted forward and backward, based on five hours' recording of five male and five female adult speakers.

REFERENCES

[1] Sigurd B. (1965) *Phonotactic structures in Swedish.* Lund: Uniskol.

[2] Gårding E. (1974), Sandhiregler för svenska konsonanter. In: *Svenskans beskrivning* 8, 97-106. Platzack C. (ed.). Lund.

[3] Eliasson S. (1986) Sandhi phenomena in Peninsular Scandinavian. In: Andersen, H. (ed) *Sandhi Phenomena in the Languages of Europe*, pp. 271-300. Berlin: Mouton de Gruyter.

[4] P. (1993), Temporal Variation of Consonant Clusters in Swedish. In: *Proceedings of the 3rd European Conference on Speech Communication and Technology*. Vol. 1, pp. 469-471. Berlin.

[5] Allwood, J. (1994), Modifierad Standard Ortografi MSO2. Mimeo. Department of Linguistics, University of Gothenburg.

SOME PROSODIC CUES FOR IDENTIFICATION OF LECTURE EXPRESSIVENESS

E. A. Nushikyan, N. A. Kravchenko
Odessa State University, Odessa, Ukraine

ABSTRACT

The aim of this paper is to manifest the results of the experimental research of prosodic organisation of text-lectures in English. The lectures under study were analysed from the point of view of their semantic and syntactic structure. The main task of this investigation is to prove the existence of tight connection between semantic value and expressiveness of a lecture and its prosodic organisation.

INTRODUCTION

A phonostylistic speech study widely spread in this country and abroad has an objective to reveal the peculiarities of prosodic and paralinguistic language means functioning in speech styles. Alongside with other speech styles the scientific prose research is of primary importance, its peculiar characteristics, scientific terminology leading to its language indices stability has always attracted the attention of linguists. Besides, the present state of constantly increasing volume of information results in leading role of scientific prose style among other styles. Lecture is one of the ways of transmission information. It is a very interesting object for investigation since it comprises some features of scientific and oratorical styles.

Many linguists refer lecture to the "influence texts" (Bahtin M., 1979, Leontyev A., 1974) but there are very few investigations of lecture from the view point of pragmalinguistics.

In this paper the prosodic analysis of lecture style in English is undertaken. It is a study of extralinguistic and linguistic factors taking part in structuring text-lecture. The extralinguistic factors are: the aim and the subject of utterance, the relations between a speaker and listeners, speaker's attitude towards the subject of utterance, social conditions of communication (Gaiduchic S., 1973). Among linguistic factors tempo, fundamental frequency and dynamic structure of the lectures were analysed.

RESEARCH MATERIAL

Three English lectures delivered by three American professors to the students of Odessa State University and their studio equivalents (written variants) read by the same speakers were taken as the subject of investigation. The material under study allowed to identify eight main parts in any lecture: 1) introduction to a lecture, 2) introduction to a new subject of the lecture, 3) introduction of new notions, 4) the main part of the lecture (the body), 5) deviations from the subject of the lecture, 6) appeal to the audience, 7) drawing conclusions, 8) the end of the lecture.

These parts of the lectures were instrumentally analysed with the help of IBM speech program, which enabled graphical presentation of tempo, fundamental frequency and intensity.

DATA ANALYSIS AND RESULTS

The comparative analysis of the lectures shows that there exists certain regularities in the prosodic organisation of above-mentioned parts of the lecture.

As far as temporal structure of the lectures is concerned the tempo of the introductory part of the lectures is rather slow. The variations in tempo from slow to quick of the main part of the lecture serve as cues of lecturer emotionality and expressiveness. Thus, the most important parts of the lecture are made prominent by changes in tempo - from quick in neutral deviations to slow in emotionally coloured utterances. While approaching to the end of the lecture all lecturers change their tempo to slow or even very slow in order to make it prominent (see Table 1).

Table 1. The average data of mean syllable duration of suprasentential units, included in different parts of the lecture.

part of the lecture	spontaneous lecture	studio lecture
introduction	239	179
introd. of new notions	245	177
introd. to a new subject	238	171
the body	249	185
deviations	212	165
appeal	214	173
conclusions	233	186
the end	252	202

The data of the table show that the spontaneous lecture is characterised by considerably larger value of mean syllable duration, that is the tempo of this kind of a lecture is much slower. Besides, in semantically unimportant parts of both types of the lectures the mean syllabic duration is less - the tempo is quicker than in introductory parts of the lecture and especially in the end of the lecture.

Pauses play an essential part in revealing lecture expressiveness. The analysis of textual pausation shows the considerable difference between spontaneous and studio lectures. Expressive lectures delivered to the audience are characterised by the greater amount of pauses (especially pauses of hesitation) and their longer duration (mainly in introduction, introduction to a new subject of the lecture and in the end of the lecture) whereas in neutral studio lectures pauses are much shorter. Short pauses prevail in introduction of new notions and appeal to the audience, middle pauses are found mainly in introduction and in the end of the lecture.

The coefficient of pausation is also helpful for differentiation of two types of the lectures (see Table 2)

Table 2. The coefficient of pausation in different parts of the lecture.

part of the lecture	spontaneous lecture	studio lecture
introduction	0,38	0,18
introd. of new notions	0,21	0,14
introd. to a new subject	0,36	0,20
the body	0,25	0,18
deviations	0,09	0,08
appeal	0,18	0,16
conclusions	0,19	0,15
the end	0,29	0,20

The data of this table manifest that in the emotional spontaneous lectures the quantity of the coefficient is considerably larger than in neutral studio lectures and depends on the place of the lecture these pauses occur.

The investigation of melodic component of lecture intonation includes the analysis of fundamental frequency and terminal tones of the utterances of different emotional tensity.

It is established that phrases in the suprasentential units of spontaneous lectures, characterised by a greater degree of expressiveness, are uttered with various terminal tones and more complicated ones than the same phrases in neutral studio lectures (see Figure 1).

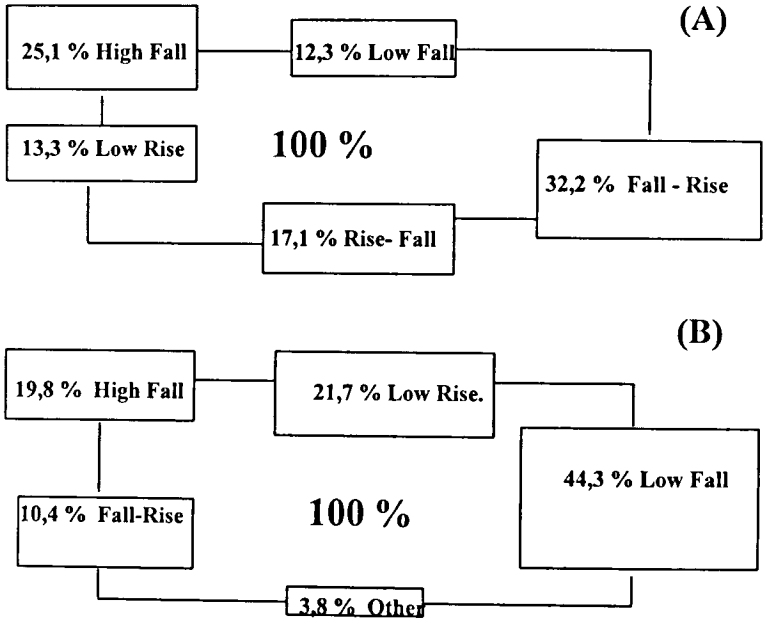


Figure 1. The frequency of different terminal tones in spontaneous lecture (A) *and studio lecture (B).*

The picture shows that in studio lectures simple terminal tones - Low Rise and Low Fall prevail, while in spontaneous lectures phrases in suprasentential units are pronounced with different complex tunes (mainly Fall-Rise, High Fall and Rise-Fall).

The analysis of fundamental frequency shows that higher fundamental frequencies are observed in spontaneous lectures, especially in introduction, appeal to the audience, and conclusions. Investigation allows to make a conclusion about close interaction of fundamental frequency interval of a separate utterance in spontaneous lectures with semantically important parts of the text.

Figure 2 illustrates fundamental frequency range in the main part of the lecture.

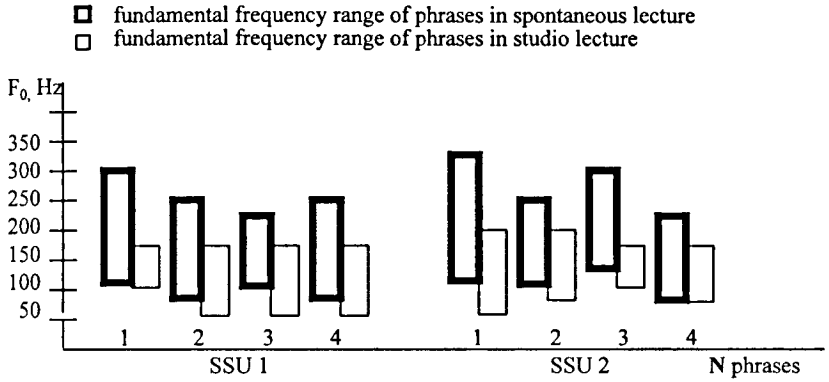


Figure 2. Fundamental frequency range of suprasentential units included in the main part of the lecture

This histogram shows the considerable difference between the value of this parameter in spontaneous and studio lectures. Besides, in the main part of the spontaneous lectures as well as in some other parts of them, the initial phrases are characterised by the highest fundamental frequency.

The analysis of dynamic structure of the lectures included investigation of maximum of intensity of phrases which constituted the suprasentential units of lectures and mean syllabic intensity of these phrases. The research shows that there exists correlation of dynamic structure of the lecture with its expressiveness and emotionality. Table 3 presents the data received by means of comparison of mean syllable intensity of phrases of spontaneous and studio lectures.

The table shows that the quantity of this parameter depends on the part of the lecture. Thus, the most semantically and emotionally important parts of the lecture, such as introduction of new notions, the end of the lecture, drawing conclusions, are characterised by more considerable difference between the dynamic structure of spontaneous and studio lectures. This fact proves the influence of expressiveness upon the parameter of mean syllable intensity.

Table 3. The ratio of mean syllable intensity of phrases included in suprasentential units of spontaneous and studio lectures.

part of the lecture	the ratio of mean syllable intensity
introduction	1,34
introduction of new notions	2,01
introduction to a new subject	1,28
the body	1,21
deviations	1,14
appeal	1,11
conclusions	1,63
the end	1,84

CONCLUSIONS

The obtained results are helpful for identification of lecture expressiveness. They reveal some regularities in prosodic organisation of text-lectures of different types. The analysis of the experimental material allows to make a conclusion about close interaction of prosodic means with different degree of expressiveness of the lecture parts.

REFERENCES

[1] Bahtin, M. (1979) Estetica slovesnogo tvorchestva, Moscow
[2] Leontyev, A.(1974) Lectsiya kak obshcheniye, Moscow
[3] Gaiduchic, S. (1973) Fonostylystychesky aspect ustnoy rechi, Minsk

UNSUPERVISED DECOMPOSITION OF PHONEME STRINGS INTO VARIABLE-LENGTH SEQUENCES, BY MULTIGRAMS

Frédéric BIMBOT, Sabine DELIGNE, François YVON.

ENST / Télécom Paris, CNRS - URA 820,
46 rue Barrault, 75634 PARIS cedex 13, FRANCE, European Union.

ABSTRACT

The multigram model allows the automatic extraction of variable-length regularities in strings of symbolic units. In this paper, we assess the multigram model as a phonotactic model. In our experiments on the MALECOT corpus, the multigram model outperforms the classical n-gram model for the description and the prediction of phoneme strings, measured in terms of test set perplexity. We also show that the model can be used to automatically derive segmental speech synthesis units.

1. INTRODUCTION

A string of graphemes or phonemes can be viewed as the result of a complex encoding process which maps a message into a stream of symbols. This string of symbolic units is far from being random, as the encoding process is subject to various phonotactical, lexical and syntactical constraints. In particular, combinations of letters form lexical items, which themselves are arranged according to grammar rules.

These constraints are responsible for a significant degree of redundancy in natural language symbolic representations such as phoneme strings or word strings. For instance, in the phonemic transcription of a conversation, all phonemes are not equally likely, nor are their two-by-two combinations (bigrams), their three-by-three combinations (trigrams), and so on...

This redundancy is partly exploited by probabilistic language models, among which the ***n-gram*** model [1] is very popular in language engineering. However, the underlying hypothesis of this model is that the probability of a given linguistic symbol (phoneme or word) depends on its n predecessors, n being fixed a priori and supposed constant over the whole text.

In opposition, the ***n-multigram*** model, recently developed [2] and extended [3], is based on the hypothesis that the dependencies between symbols are of variable-length (from 0, i.e independency, up to length n).

The multigram approach was previously tested with success as a ***language model***, i.e a model of word dependencies within a sentence [2][3]. In this paper, we report its performance as a *phonotactic model*, and we assess its application for the automatic extraction of formal speech synthesis units.

2. THEORETICAL ASPECTS

2.1. Formulation

In this section, we denote as $A = \alpha_1 \cdots \alpha_t \cdots \alpha_N$, a string of N linguistic symbols.

The conventional n-gram model assumes that the statistical dependencies between symbols are of fixed-length n along the whole sentence. The *likelihood* of A is then computed as :

$$\mathcal{L}_{gr}(A) = \prod_{t=1}^{t=N} p(\alpha_t \mid \alpha_{t-n} \ldots \alpha_{t-1}) \qquad (1)$$

where $p(\alpha_t \mid \alpha_{t-n} \ldots \alpha_{t-1})$ is the conditional probability of observing symbol α_t given that the history of $n-1$ symbols $\alpha_{t-n} \ldots \alpha_{t-1}$ has occured[1].

The n-multigram model makes a different assumption : under this approach, a stream of linguistic symbols is considered as the concatenation of independent variable-length sequences, and the likelihood of the whole string is computed as the sum (or the maximum) of the individual likelihoods associated to each possible segmentation.

Let Δ denote a possible segmentation of A into q sequences $s_1 \cdots s_k \cdots s_q$. For instance :

$$\begin{aligned} s_1 &= [\alpha_1\, \alpha_2], \\ s_2 &= [\alpha_3\, \alpha_4\, \alpha_5], \\ s_3 &= [\alpha_6], \\ &\cdots \\ s_q &= [\alpha_{N-2}\, \alpha_{N-1}\, \alpha_N], \end{aligned}$$

The n-multigram model computes the likelihood $\mathcal{L}_\Delta(A)$ of string A for segmentation Δ as the product of the probabilities of the successive sequences composing Δ :

$$\mathcal{L}_\Delta(A) = \prod_{k=1}^{k=q} p(s_k) \qquad (2)$$

[1] **Further in this paper, we recall the link that exists between the likelihood of a model and the explanatory capabilities of the model in terms of prediction.**

Input character strings	
1	blessedisthemanthatwalkethnotinthecounseloftheungodly...
2	buthisdelightisinthelawofthelordandinhislawdothhemeditatedayandnight
3	andheshallbelikeatreeplantedbytheriversofwaterthatbringethforthhisfruit...
4	theungodlyarenotsobutarelikethechaffwhichthewinddrivethaway

Output 5-multigram decompositions	
1	bless ed isthe man that walk eth not inthe couns el ofthe un godly ...
2	but his d e light is inthe law ofthe lord and inhis law do th he medit at e day and night
3	and he shall be likea t re e plant ed bythe river sof water that bring eth forth his fruit ...
4	the un godly are not so but are like the ch a f f which thew in d drive th away

Figure 1: Old Testament - King James Version (Psalms - first 5 verses). Character string decomposition using a 5-multigram model. Variable-length regularities are extracted without any supervision.

Denoting as $\{\Delta\}$ the set of all possible segmentations of A into sequences of maximum length n, the total likelihood of A is :

$$\mathcal{L}_{\mu gr}(A) = \sum_{\Delta \in \{\Delta\}} \mathcal{L}_\Delta(A) \qquad (3)$$

A decision-oriented version of the model can provide a maximum-likelihood decomposition of A as the segmentation Δ^* with highest individual likelihood :

$$\Delta^* = \text{Argmax}_{\Delta \in \{\Delta\}}\ \mathcal{L}_\Delta(A) \qquad (4)$$

and

$$\mathcal{L}^*_{\mu gr}(A) = \mathcal{L}_{\Delta^*}(A) = \max_{\Delta \in \{\Delta\}} \mathcal{L}_\Delta(A) \qquad (5)$$

For instance, with $A = abcd$ ($N = 4$) and a conventional tri-gram model :

$$\mathcal{L}_{3-gr}(abcd) = p(a|\phi\phi)\, p(b|\phi a)\, p(c|ab)\, p(d|bc)$$

with ϕ denoting the null symbol, whereas for a 3-multigram model :

$$\mathcal{L}^*_{3-\mu gr}(abcd) = \max \left\{ \begin{array}{l} p([a])\, p([bcd]) \\ p([abc])\, p([d]) \\ p([ab])\, p([cd]) \\ p([ab])\, p([c])\, p([d]) \\ p([a])\, p([bc])\, p([d]) \\ p([a])\, p([b])\, p([cd]) \\ p([a])\, p([b])\, p([c])\, p([d]) \end{array} \right\}$$

The maximum term indicates the maximum likelihood segmentation Δ^*, for instance : [ab][c][d].

2.2. Algorithm

The algorithm for estimating the multigram probabilities from a training corpus proceeds iteratively. After the sequence probabilities have been initialised by counting all co-occurences of symbols up to length n, a forward-backward procedure is implemented to refine these estimates. Once convergence is reached, a Viterbi procedure provides the maximum likelihood segmentation, either on the training set, or on a test set, as in Equation (4). A full formulation of the algorithm and additional details[2] can be found in [2] [3].

2.3. Illustration

Figure 1 shows the result of the multigram decomposition of an english text[3], from which all spaces between words were removed. The set of linguistic units, in this case, is composed of the 26 lower case letters of the alphabet, and the corpus on which the probabilities are estimated contains approximately 200 000 characters. After 10 training iterations of a 5-multigram model, convergence is obtained, and the dictionary of typical sequences contains approximately 1100 entries.

In Figure 1, we indicate sequence borders by a space. Some typical english words or morphemes are automatically extracted. Some frequent combinations of small words are often merged (*inthe, ofthe, inhis,...*), while rare words tend to be broken into smaller units (*t re e, ch a f f,...*). Occasionally, an unappropriate segmentation occurs (*river sof, thew in d, ...*). Nevertheless, it is quite clear that the multigram model, though using no prior knowledge, extracts variable-length regularities which are strongly correlated with the morpheme structure of the input text.

3. EXPERIMENTAL PROTOCOL

3.1. Motivation

The experiments reported in the rest of this paper are carried out on phoneme strings. Our experimental protocol is designed to assess objectively the multigram model as a description of syntagmatic aspects in phoneme strings, and to investigate its potential application as a tool for deriving variable-length segmental units for speech synthesis. In a first series of experiments, the multigram model is used to predict phoneme strings

[2] In particular, in what concerns the *pruning factor* [2].

[3] An excerpt from the Bible.

and evaluated in terms of perplexity. In a second experiment, it is used to build variable-length speech synthesis units, by merging diphones which frequently co-occur together. In this last case, the evaluation criterion is the reduction in the number of concatenations per sentence.

3.2. Database

Our corpus is the MALECOT corpus. It consists of approximately 200 000 phonemes (13 000 sentences) which were obtained by a manual phonemic transcription of informal conversations in the French language [4] [5]. We split our corpus into a *training set* (first 150 000 phonemes) and a *test set* (last 50 000 phonemes). The phonemic alphabet is composed of 35 symbols, namely : a, i, e, ɛ, u, o, ɔ, y, ø, œ, ə, ɑ̃, ɛ̃, ɔ̃, œ̃, p, t, k, b, d, g, f, s, ʃ, v, z, ʒ, m, n, ɲ, l, ʀ, j, w, ɥ. Spaces are removed from the corpus, so that the word borders are unknown.

3.3. Perplexity

As an objective measure of the multigram model ability in representing sequences of phonemes, we use the *perplexity* measure [1]. The perplexity of a model $\mathcal{M}$ on a string A is defined as :

$$X = 2^H \quad \text{where} \quad H = -\frac{1}{N} log_2 \mathcal{L}_{\mathcal{M}}(A) \qquad (6)$$

where N is the length of string A and $\mathcal{L}_{\mathcal{M}}$ the likelihood provided by the model, as in Equations (1), (3) or (5), for instance.

Consider now a string B of length N generated by a memoryless source[4], from an alphabet of X equiprobable symbols. As the probability of each symbol is $\frac{1}{X}$, the perplexity of B is $X' = 2^{H'}$ where :

$$H' = -\frac{1}{N} log_2 \mathcal{L}(B) \qquad (7)$$

$$= -\frac{1}{N} log_2 \left[\frac{1}{X}\right]^N = log_2 (X) = H$$

Hence $X' = X$. Perplexity can thus be viewed as the randomness in the data that can not be predicted by the model.

If two models provide different perplexity values on a same *test* corpus, the one with lower perplexity can be considered as more efficient in explaining the underlying process which generated the data, whereas a lower perplexity on the *training* set only indicates a better ability in rendering the particularities of the training data.

Our first set of experiments consists in comparing perplexity values provided by n-gram and n-multigram models[5], on the phoneme strings in the MALECOT corpus.

[4] The symbols emitted by a *memoryless* source are statistically independent from one another.

[5] Using Equations (1) and (5) respectively.

3.4. Average number of concatenations

Segmental speech synthesis generally uses acoustic diphone units[6] which are concatenated to each other in order to reconstruct a speech utterance. In practice, some speech synthesis defects come from discontinuites at the level of the concatenations.

The application of the multigram model to strings of formal diphones can extract sequences of diphones which frequently co-occur together. Diphones within such sequences can then be advantageously merged together into a longer *multiphone* unit. For instance, if the diphone sequence <as><sj><jɔ̃> has a high probability, a quadriphone unit <asjɔ̃> can be created and added to the list of segmental synthesis units, which will avoid two concatenations during the synthesis process, each time this group of phonemes will be met. However, the set of multiphone units must result from a compromise between the economy in concatenations and the number and volume of acoustic units in the segmental dictionary.

In our second set of experiments, we evaluate the benefit of multiphone units as the average number of concatenations per sentence in the MALECOT corpus as well as in terms of number of segmental units. We also give a rough estimate of the acoustic storage requirements, measured as minutes of speech[7].

4. RESULTS

4.1. Phoneme sequence modeling

Table 1 summarises the results obtained in terms of training and test set perplexity for n-grams and n-multigrams (with $1 \leq n \leq 5$)[8].

With more than 8000 entries versus less than 3000, the trigram (and 4-gram) model provides a lower prediction capability than the 5-multigram model (perplexity of 10.1 (or 10.0) versus 9.4). The n-multigram model also shows good generalisation properties from the training set to the test set.

Figure 2 depicts an example of n-multigram decompositions of a french sentence in its phonemic form, from our test set in the MALECOT corpus. Here again, the phoneme multigrams show a striking correlation with morpho-lexical elements, especially for $n = 5$. They could prove efficient as word- or subword-like units for speech recognition.

[6] An acoustic diphone can be understood as a *domino* composed of the transition between the "center" of a phone and the "center" of the next phone.

[7] On the basis of 50 ms for a border phoneme and 100 ms for an inside phoneme.

[8] A pruning factor of 1.0 was used for the n-multigram training. A Good-Turing estimate was used for unseen n-grams, whereas a fixed penalty was used for unseen n-multigram of length 1. See details in [2][3].

model order	n-gram model $n=1$	$n=2$	$n=3$	$n=4$	$n=5$	n-multigram model $n=1$	$n=2$	$n=3$	$n=4$	$n=5$
training set perplexity	25.8	13.7	8.9	5.3	3.3	25.8	16.8	12.4	9.9	8.3
test set perplexity	25.8	13.9	10.1	10.0	15.7	25.8	16.8	12.7	10.7	9.4
number of entries	35	937	8455	28786	51197	35	352	1119	1891	2683

Table 1: MALECOT corpus : training set perplexity, test set perplexity, and number of entries for the n-gram and the n-multigram models ($1 \leq n \leq 5$), for phoneme string modeling and prediction.

i	l	e	t	e	v	i	d	ɑ̃	d	a	j	œ	ʀ	k	i	l	f	o	d	ʀ	a	i	v	ə	n	i	ʀ			
il		e		te		vi		dɑ̃		d		aj		œʀ		k		il		fo		d		ʀa		i		və		ni		ʀ
il			ete			vid		ɑ̃		da			jœʀ			ki			lfo		d		ʀa		i		vən			iʀ	
	ilet				evid			ɑ̃	d			ajœʀ			k			ilfo		d		ʀa		iv			ənʀ			
	ilet				evidɑ̃							dajœʀ						kilfo		d		ʀa		i			vəniʀ			

il	le	et	te	ev	vi	id	dɑ̃	ɑ̃d	da	aj	jœ	œʀ	ʀk	ki	il	lf	fo	od	dʀ	ʀa	ai	iv	və	ən	ni	iʀ					
ile		ete		evi		idɑ̃		ɑ̃d		daj		jœʀ		ʀk		kil		lfo		od		dʀ		ʀa		ai		iv		vən		niʀ
ile		ete		evid			dɑ̃d		da		ajœʀ			ʀki			ilfo		od		dʀ		ʀa		ai		iv		vəni		iʀ
ile		ete		evi		id		dɑ̃d			dajœʀ			ʀki			ilfod			dʀ		ʀa		ai		iv		vəniʀ		

Figure 2: French sentence : "il est évident d'ailleurs qu'il faudra y venir" (MALECOT corpus - test set). 1- 2- 3- 4- and 5-multigram phoneme segmentations and 2- 3- 4- and 5-multiphone decompositions.

4.2. Multiphone units

Figure 2 illustrates the result of multiphone decompositions on a test sentence. Here, the elementary symbol is a diphone, and a sequence of diphones is represented as a tri-, quadri- or quintiphone. Table 2 reports detailed results concerning the number, size and repartition of multiphone units obtained by the multigram model, for different orders[9].

multiph. order	2	3	4	5
nb diph.	759	567	582	583
nb triph.	0	1828	938	859
nb quadriph.	0	0	1365	612
nb quintiph.	0	0	0	879
total	759	2395	2885	2933
missing diph.	466	658	643	642
grand total	1225	3053	3528	3575
≈ vol. in mn	2	8	12	14
nb conc tr. set	13.6	7.5	6.2	5.8
nb conc test set	13.5	7.7	6.6	6.3

Table 2: See text.

Table 2 shows, for instance, that the set of 5-multiphones (last column) is composed of 583 diphones, 859 triphones, 612 quadriphones and 879 quintiphones, i.e a total of 2933 units. As 35 × 35 = 1225 diphones are necessary to guarantee a 100 % coverage of any text, 642 other diphones must be added to the dictionary, which leads to a grand total of 3575 units, i.e less than 3 times the number of diphones. The 5-multiphone set would require approximately 7 times more space than the diphone set, to be stored in its acoustic form. In counterpart, it can be expected that a sentence could be synthesized with twice less concatenations, as the average number of concatenations per sentence on the MALECOT test corpus falls from 13.5 to 6.3. This should have a significant impact on synthetic speech quality.

[9] A pruning factor of 2.0 was used for this experiment.

5. CONCLUSION

The multigram model provides a powerful framework for the unsupervised description, decomposition and prediction of phoneme sequences, and an interesting tool for the automatic design of segmental speech synthesis units. More generally, it appears as a relevant approach for the modeling of natural language syntagmatic aspects, which are usually based on variable-length schemes.

References

[1] F. JELINEK: *Self-organized language modeling for speech recognition.* Readings in Speech Recognition. Ed. A. Waibel, K.F. Lee, Morgan Kaufmann Publ. Inc., 1990.

[2] F. BIMBOT, R. PIERACCINI, E. LEVIN, B. ATAL: *Modèles de séquences à horizon variable: multigrams.* XXèmes JEP, 1994. To appear in IEEE-SP Letters, as: *Variable-length sequence modeling : multigrams.*

[3] S. DELIGNE, F. BIMBOT: *Language modeling by variable length sequences : theoretical formulation and evaluation of multigrams.* IEEE-ICASSP, 1995.

[4] A. MALECOT: *New procedures for descriptive phonetics.* Papers in Linguistics and Phonetics to the Memory of Pierre Delattre. Mouton, 1972.

[5] J.P. TUBACH, L.J. BOE: *Un corpus de transcriptions phonétiques: constitution et exploitation statistique.* Report ENST-85D001, 1985.

ARTIFICIAL VISUAL SPEECH (AVS) CONTROLLED BY FUZZY METHODS

H-H. Bothe
Technical University of Berlin, Electronics Institute, Berlin, Germany

ABSTRACT

This paper describes a new approach to modelling visual speech movements with the help of a complex fuzzy-neural network (FNN), putting a particular emphasis on the input coding. The FNN uses a set of characteristic key-pictures derived from video films with prototypic speakers. A later facial animation may be created by arranging a sequence of key-pictures with respect to the given phonetic input text and a subsequent calculation of interim frames. For this selection, The FNN consists of a radial basis function network, a multilayer perceptron and a self-organizing Kohonen map.

Goal of the described work is the development of a facial animation program for the teaching of lip-reading that may be applied in schools or rehabilitation centers for hearing-impaired people. The present version creates grey-scale films on the computer screen that correspond to a given input text. It is implemeted on PC (MS-DOS) and is prepared for additional connection to a synchronized speech synthesis computer.

INTRODUCTION

Since the experimental work of Menzerath and de Lacerda [1] it is known that the movements of the speech organs are structurally interrelated within a spoken context. The sound signals are created in the course of a fully overlapping *coarticulation*.

The resulting facial movements can be treated as visual speech. While the smallest speaker-independent perceptual units of the acoustical signal are the *phonemes*, the correlating visual speech units are called *visemes* [2].

In spite of a large data reduction visual speech contains usually sufficiant information to enable hearing-impaired people to lip-read. The largest part of information is derived from the movements of the mouth region.

Thus, this paper postulates the modelling of facial movements that are relevant for the process of lip-reading with the help of changes in the lip contours. Although the development of a general motion model on sophisticated animation computers is desirable, this work concentrates on the implementation of one realistic prototype model that can be used in schools or rehabilitation centers. The proposed motion model has to take the resulting limitations in account. Other approaches are, for instance, described in [3-5].

DATA ACQUISITION

A block diagram of the complete analysis-synthesis-system is shown in Figure 1. At first, the acoustic speech signal and video data of prototype speakers were recorded on videotape and analyzed for a text corpus of 84 sentences. Proposing a speech model leads to characteristic

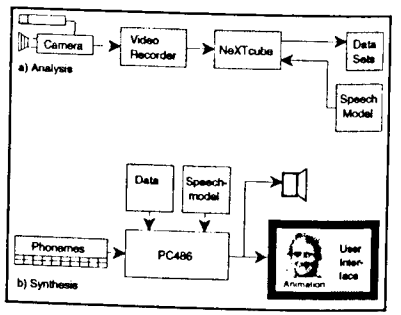


Figure 1. Visual speech analysis-synthesis system.

speaker-dependent movement data.The synthesis computer uses these data, together with the speech model, for the facial animation.

The data analysis consists of two steps, the fixing of the sound boundaries and the determination of phone-characteristic single pictures. Both steps have to be processed interactively.

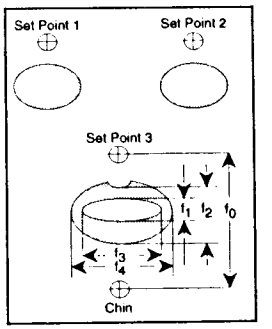
The phone boundaries were determined with the help of oscillogram, sonagram and acoustic feedback as described in [6]. Then, those video-frames fitting best with the subjective impressions of a well pronounced sound were indicated with the help of both the acoustic and visual material by different experts in lip-reading. In some cases, e.g. for the phonemes /h,g,k/, an exact determination was not possible; the production of these sounds is only weakly represented on the speaker's face and can usually not be perceived by people who lip-read. For this reason the existance of either one or none characteristic picture is proposed, concerning the modelling as well as the later facial animation [7].

The speech movements are characterized by continuous quasi-periodic opening and closing processes that are reflected in the courses of the visual features. Thus, the proposed motion model is based on a library of characteristic 'key-pictures', arranged in the extrema positions, that allow to track the courses of features. The interactively determined characteristic pictures of the phonemes are located in or at least close to these extrema.

In order to compose the library of key-pictures, the 'characteristic pictures' of the text corpus were classified by a FCM-algorithm (see [8]) with respect to lip shape and position by using specific visual features. The algorithm generates optimum location of the clusters automatically with respect to a given number of clusters. The cluster centers in the feature space compose the library of representative key- pictures.

In order to guarantee reproducable results some set points on nose and forehead and the contours of the lips were marked with a flourescent color. During the recording the persons were slightly radiated with UV light. An exemplary single frame shows Figure 2, together with the scheme of the feature extraction.

a



b

Figure 2. a. Examplary single frame, b. visual feature extraction of frontal view.

KEY-PICTURE SELECTION

A library of 'key-pictures' represents the properties of a speaker's speech movements. There is no a priori relation between the phonemes and the corresponding key-pictures. Thus, the unknown mapping function was trained in a complex artificial neural network with the help of the courses of visual features

for part of the spoken sentences. The other part is taken for the evaluation of the mapping quality.

The neural network considers forward and backward coarticulation effects by using an influence frame of 3+1+3 phonemes. The medium phoneme in its context is related to one (resp. none) key-picture. Pulling this 7×1-frame through a given sequence of phonemes creates a framework of pictures that represent important stages for the courses of visual features. The time distances between two key-pictures are calculated by averaging the time distances of the corresponding interactively determined pictures on the whole text corpus. The phoneme to be mapped is ignored if there was no characteristic picture to determine in the video film. A block diagram of the selection process is shown in Figure 3a.

Here, a key-picture out of the library is related to the phoneme /o:/ considering the dependencies of the preceeding and following phonemes /a:_n_h/ and /f_'_i/. The sign /'/ stands for an occlusional sound in a word boundary. The approximation of given facial movements for a sequence /Ph_1_Ph_2_..._Ph_8/ of phonemes by a sequence of key-pictures Kp_i is shown in Figure 3b. A morphing algorithm for grey-scale pictures creates interim frames at specific locations.

There are different approaches known for mapping phoneme sequences on speech pattern, as, for instance, the NETtalk algorithm [9]. This paper proposes the fuzzy neural network architecture seen in Figure 4 that considers similarities among the visual phonematic correlates. These serve for the input coding and, on the other hand, take influence on the network architecture [10].

For training, the FNN is cut at the viseme layer and trained in three steps by error-backpropagation:

- the phoneme-to-viseme-mapping with the help of the viseme structure of German (see [11]) and the recorded video material,
- the viseme-to-feature-vector mapping,
- the connected FNN with respect to the video material.

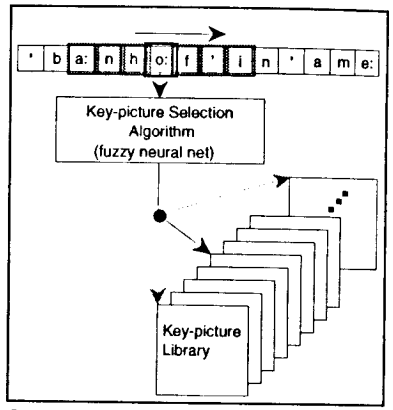


Figure 3. a. Selection, b. placement of key-pictures with respect to a given phoneme sequence Ph_1 ... Ph_8.

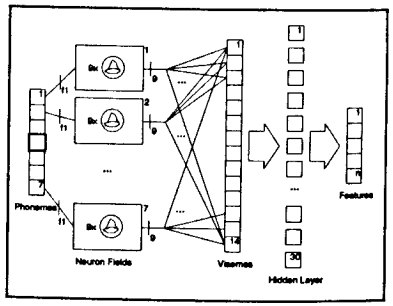


Figure 4. Block diagram of the fuzzy-neural network (FNN).

The exact placement of a key-picture in the frame of the phone boundaries is - depending on the context - also calculated with the help of an artificial neural network.

Evaluation of the Naturalness of the Facial Animation

For given phonetic input sequences, the resulting computer animation program produces similar courses of predicted visual features as measured in the video films. Despite of the fact that several artifacts still occur, the general tendency of opening and closing movements looks very much alike (see Figure 6).

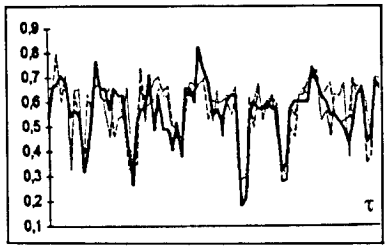


Figure 6. Measured (—) and predicted (- - -) courses of the visual feature $f_1(\tau)$.

Since even small local differences of the actual courses of visual features may result in a larger perceptual artefact and vise versa, the naturalness of the calculated videos has to be evaluated with the help of those who can lip-read, i.e. with hearing-impaired people. First results for a simple demonstration version of the animation program were investigated in a school for hard-of-hearing children and can be found in [12].

References

[1] Menzerath, P. and A. de Lacerda (1933), *Koartikulation, Steuerung und Lautabgrenzung*, Berlin.

[2] Fisher, C.G. (1968), "Confusions Among Visually Perceived Consonants", *J. Speech and Hearing Res.*, 11, 796-804.

[3] Storey, D. and Roberts, M. (1988), "Reading the Speech of Digital Lips: Motives and Methods for Audio-visual Speech Synthesis", *Visible Language* 22, 112-127.

[4] Cohen, M.M. and D.W. Massaro (1990), "Synthesis of Visible Speech", *Behaviour Research Methods, Instruments & Computers*, 260-263.

[5] H.H. Bothe, G. Lindner and F. Rieger (1993), "The Development of a Computer Animation Program for the Teaching of Lipreading", In: E. Ballabio, I. Placencia-Porrero and R. Puig de la Bellacasa (Eds.), *Technology and Informatics 9, Rehabilitation Technology: Strategies for the European Union*, Amsterdam, 45-49.

[6] C. Heise and H.H. Bothe (1993), "Phone and Syllable Segmentation by Concurrent Window Modules", *Proc. of the EUROSPEECH'93*, Berlin, 669-672.

[7] H.H. Bothe and F. Rieger (1994), "Zum Zusammenhang von akustischen und visuellen Korrelaten lautlicher Artikulationsprozesse", *Tagungsberichte der 20. Deutschen Jahrestagung für Akustik (DAGA)*, pp. 1337-1340, Dresden.

[8] Bezdek, J.C. (1981), *Pattern Recognition with Objective Function Algorithms*, London.

[9] Sejnowski, T.J. and C.R. Rosenberg (1986), "NETtalk: A Parallel Network that Learns to Read Aloud", *Electrical Engg. and Comp. Science Tech. Rep. JHU/EECS-86/01*, J. Hopkins University.

[10] H.H. Bothe (1995), "Fuzzy Input Coding for an Artificial Neural Network", *Proc. ACM Symposium on Applied Computing'95*, pp. 450-454, Nashville, USA, 1995.

[11] G. Alich (1961), *Zur Erkennung von Sprachgestalten beim Ablesen vom Munde*, (Doc. Thesis), Bonn.

[12] H.H. Bothe and F. Rieger (1993), "Computer Animation for Teaching Lip-reading", *Proc. 2nd European Conference on the Advancement of Rehabilitation Technology (ECART)*, pp. 4.4, Stockholm.

A NEW SPEECH ANALYSIS SYSTEM: ASSESS (AUTOMATIC STATISTICAL SUMMARY OF ELEMENTARY SPEECH STRUCTURES)

R. Cowie, M. Sawey and E. Douglas-Cowie
School of Psychology / School of English, Queen's University, Belfast, UK

ABSTRACT

ASSESS is a set of programs which automatically segment the speech signal, and then provide systematic statistical measures of the structures elicited. The attributes that it describes are relevant to distinguishing a number of speech varieties, normal and abnormal. We have applied it to speech pre and post cochlear implant, schizophrenic speech, and markers of emotion in speech.

INTRODUCTION

This paper is about extending a natural approach to measuring speech properties. The general aim is to develop systems which analyse speech automatically and create a battery of statistical descriptors. The distinctive extension that we consider is to incorporate preprocessing which extracts from the raw signal key features and relationships, whose properties can then be measured and summarised. This has the potential to produce much richer descriptions than approaches which treat the signal as a whole. We have developed a system which uses this strategy, and tested its value in several projects. It is called ASSESS, as an acronym for Automatic Statistical Summary of Elementary Speech Structures.

ASSESS grew out of concern to measure speech attributes which lead listeners to place people in emotionally and socially significant categories. It is not self-evident that statistical measures are relevant to the problem: they might be dominated by idiosyncratic features of speech at one extreme, or by phonetic and syntactic structure at the other. However, a good deal of evidence suggest otherwise. ASSESS builds on that evidence, and incorporates relevant measures in a systematic framework.

Its immediate origins lie in research on the way speech changes when people become deaf. We were able to show that listeners made distinctive (and usually adverse) judgements about deafened people's speech [1], but our phonetic analysis, using a traditional segmental analysis, showed so few abnormalities that our data were used to argue that there were no significant abnormalities [2].

We turned to statistical techniques because impressionistically, spectrograms of deafened and control speakers looked quite grossly different. Studies confirmed that such an approach captures some effects of acquired deafness on speech [3], [4], particularly the way intensity rises and falls. Patterns of change in F0 also appear to distinguish deafened and hearing speakers statistically: the distribution of types is more obviously anomalous than the nature of individual tokens [5]. A less intuitive feature of deafened speech involves spectral balance. Unusually high proportions of energy concentrate around the region associated with F2 [3].

There are indications that other speech varieties may be distinguished in related ways. Formal and informal speech styles appear to differ statistically [3]. In Belfast speech at least, social distinctions involve the distribution of pitch patterns which invite summary in statistical terms [6]. Certain forms of schizophrenia seem to be marked by the statistical distributions associated with F0 [7]. Statistical properties related to intonation bear at least broad relationships to the vocal expression of emotion [8],[9],[10]. Voice quality has often been identified as a factor in reactions to and evaluations of a speaker [11], and some types of voice involve distinctive balances between broad regions of the spectrum [12].

ASSESS develops the intuition that these observations reflect a domain which is both coherent and significant. It is a prototype which is competent enough to let us probe the nature of the domain and its relation to the way people respond to voices. Its structure offers a useful overview of the tasks that systems of this kind need to address. Its performance has allowed us to carry out reasonably large scale studies, which help to validate the approach, and to refine it in the light of experience.

SYSTEM OVERVIEW

ASSESS has four components.
(1) InASSESS saves samples from tape using a CED 1401 signal capture unit. Sampling is at 20kHz: frequencies above 10kHz are removed by low pass analogue prefiltering. Memory limitations restrict a single sample to 8.2 seconds. As typical inputs are much longer, they are divided by hand using a display of the time waveform to find natural cut-off points.
(2) QuantASSESS is a signal processing stage which extracts basic descriptions - an intensity profile, initial estimates of F0, and a 1/3 octave spectrum.
(3) QualASSESS takes the output of QuantASSESS, identifies key features, and creates initial statistical descriptions.
(4) SumASSESS integrates data from files which form a single passage and generates graphical and numerical summaries. Additional programs integrate data from multiple passages and replay stored graphical representations.
QuantASSESS and QualASSESS are currently written in TurboPascal and run on IBM PCs. Both can run a large batch of files from InASSESS, though each one is analysed independently. SumASSESS and supporting programs are written in QuickBasic and run on a Macintosh (allowing access to Macintosh statistical packages and graphics: convenience at that level is important).

Hardware seriously limits the utility of the current system. Capture is the greatest bottleneck, and work is in hand on a replacement using a SUN SparcII to capture long samples directly. The other components will then be transferred to the same platform and adjusted accordingly.

EXTRACTING KEY FEATURES

QuantASSESS has three functions.
(1) It recovers intensity from voltage. Energy is integrated over periods of 25.6 milliseconds, which we call 'slices'.
(2) It creates a spectrum. An FFT is calculated for each slice, and used to estimate the output of 18 1/3 octave filters, with centre frequencies running from 128Hz to 6.5kHz.
(3) It makes initial estimates of voice pitch using an algorithm which looks for upstrokes in the time waveform that may signal vocal cord openings. The intervals between these strokes give direct but noisy pitch estimates. QuantASSESS then finds sequences of strokes which could correspond to evenly spaced vocal cord openings, allowing that a few real strokes may have been missed, or spurious strokes introduced. Confidence values are associated with pitch estimates derived from these sequences: they depend on the number of missing or spurious strokes assumed.

QualASSESS recovers descriptions of key contours, and finds key transition points - beginnings and ends of silences and fricative bursts, maxima and minima on the intensity and F0 contours, and plateau boundaries. The idea of plateaux derives from work on schizophrenic speech [13]. They are 'flat' regions around an inflection where contour height has changed by less than 5% of the distance to the next inflection. Falls and rises run between plateau boundaries.

QualASSESS uses heuristics aimed at robustness rather than precision - this is essential because even a few outliers could seriously distort statistics.

Operations on the intensity contour begin with an averaging procedure which smooths out fluctuations within the duration of a typical syllable (about 150ms). Secondary smoothing removes 'zig-zags' which break a trend that is present on both sides of the 'zig-zag'. Inflections and plateaux are then found straightforwardly. A preliminary identification of pauses is also made, using the principle that levels associated with pauses are likely to be much more common than the levels just above them.

Fricative bursts are considered next. Slices are tentatively classed as fricative if the ratio of energy above 2.5kHz to that below 625Hz exceeds an empirically set threshold. Then a second pass checks for slices which are more likely to be pauses, using tests based on their duration, spectrum, and energy level. Evidence from all three is summed to decide whether slices are reassigned as pauses.

Pause finding continues by finding the average spectrum for slices which have been classified as pauses to date. All slices are compared to the resulting pause spectrum, and those with similar profiles are reclassified as pauses. So are short periods flanked by pauses on both sides.

The next task is to obtain robust estimates of voice pitch. To limit the impact of dubious pitch estimates, data

from QuantASSESS are completely ignored if they fall within a pause or a fricative burst, or exceed reasonable pitch limits. The variance of the pitch estimate is also calculated, and low confidence is assigned to data in periods with high variance. A 'snake' is then fitted to the pitch estimates. This is modelled on an elastic rope, stretched across the sample and pulled towards each data point by a spring whose strength reflects confidence in the estimate. The snake settles to a stable shape in an iterative process driven by the springs and the elasticity of the rope. This provides a robust estimate of the pitch contour. Peaks, troughs, and plateaux are then identified.

Transitions are now used to partition spectral information.

Three main average spectra are created - for fricatives, for pauses, and the 'main spectrum' for all other periods. Sub-spectra are also constructed for two significant types of episode, fricative bursts and peaks in the intensity contour. Peaks are considered as the best available approximation to vowel centres.

Two summaries are produced for each type. One sums the energy in each filter position: the other sums energy squared. Sums of squares are used later to calculate the variance of energy at each position. Describing variance for each filter gives a spectrum-like result which shows the variability of energy at each filter rather than its intensity. These variance spectra are designed to indicate whether episodes such as fricative bursts or vowels show normal or reduced variation from moment to moment.

The final part of QualASSESS deals with pitch again. QualASSESS uses a crude, but serviceable definition of tune boundaries: they are defined by pauses lasting longer than 150ms. It summarises the shape of each tune by fitting a curve with two components, a straight line and a quadratic (U-shaped) curve, with its centre on the midpoint of the tune.

STATISTICAL SUMMARY

SummASSESS links QualASSESS files derived from a single passage. It also has a normalising function. For each passage it reads a 'calibration' file to set a scale for intensity measurements. If the calibration file contains a tone of known intensity, this can be used to give true dB values. If not, median intensity can be set to a default value (usually 60dB).

SumASSESS generates statistical descriptions in two forms: a graphic summary showing the pitch and intensity contours, fricative bursts, and subspectra; and a statistical table consisting of three main blocks, dealing with properties of relatively high-order units; properties of spectra; and properties of the main contours, intensity and pitch.

Three sub-blocks describe high order units.

(1) Tunes. Tables give the number of tunes in a passage, and the average and standard deviation of several properties - duration, parameters of the fitted curves, number of inflections per tune, maximum and minimum pitch values within a tune, and the slope and the duration of opening and closing segments.

(2) 'Sound blocks'. These are stretches of intensity contour between two pauses. Mean and standard deviation are given for number of peaks per block, maximum height, and duration: and also for the duration, height, and slope of segments which open and close sound blocks.

(3) Frication. This uses measures based directly on the fricative spectrum and descriptors which deal with fricative bursts, including number of bursts and mean and standard deviation of burst properties such as duration, amount of energy, midpoint of energy, and spread of energy across the fricative region.

The spectrum section deals with five spectra, namely main; fricative burst average and variance; and intensity peak average and variance. For each, a matrix of measures describes four broad bands, covering the regions which tend to contain F0, F1, F2 and frication. For each band, SummASSESS specifies total activity, average activity per filter channel, variation between filter channels within the region, and centralisation or spread of activity across the region. The shape of each spectrum is summarised by the slope of a fitted line, and measures of its midpoint. Together these capture most properties of the spectrum that previous work suggests may be relevant [3], [11].

Contour measures form two parallel matrices for intensity and pitch. The basic unit is a row which specifies a feature's frequency, the mean and standard deviation of its value, and non-parametric measures - median, 10% point, 90% point, and inter-quartile range. Each matrix covers the following features: magnitude (dB for intensity, Hz for F0) for all points, peaks, troughs, rises, and falls: and duration measures for pauses, rises, falls, and plateaux.

PERFORMANCE

Results have been obtained from ASSESS in two major projects and a subsidiary one. They confirm the potential of the approach. The first major project [14], studied speech in 75 deafened speakers before and after cochlear implantation, plus 51 controls. It confirms and extends earlier observations on pre-implantation speech, and provides objective evidence that implantation produces changes - not all in the right direction. The second major project deals with the speech of 72 schizophrenic patients. Preliminary results show differences between them and controls, in a wider range of attributes than previous studies have considered. The subsidiary project attached to that studies vocal expression of emotion. Results show distinctions between passages expressing different emotions in spectral balance, range of pitch movement, timing of pitch movement, timing of intensity changes, and intensity distribution [15].

In none of these areas does ASSESS offer a complete analysis. Rather it is a broad brush tool, able to examine substantial bodies of speech and to establish that effects of certain general types exist. That seems a useful addition to the repertoire of phonetic science.

ACKNOWLEDGEMENT is due to Dr. D. Howard for pitch extraction algorithms.

REFERENCES

[1] Cowie, R. & Douglas-Cowie, E. (1983), "Speech in postlingual deafness", in M. Lutman and M. Haggard (eds.), *Hearing Science and Hearing Disorders*, London: Academic Press, pp. 183-230.

[2] Goehl, H. & Kaufman, D. (1986), "The real thing: a reply to Cowie, Douglas-Cowie & Stewart", *J. of Speech & Hearing Disorders* vol 51, pp.185-187

[3] Cowie, R. & Douglas-Cowie, E. (1992), *Postlingually acquired deafness: speech deterioration and the wider consequences*, Berlin: Mouton de Gruyter.

[4] Cowie, R., Douglas-Cowie, E. & Rahilly, J. (1991), "Instrumental measures of abnormalities in deafened speech", *Proc 12th ICPhS*, Aix-en-Provence, pp. 350-353.

[5] Rahilly, J. (1991), "Intonation patterns in postlingually deafened and normal hearing people in Belfast", Ph.D dissertation, Queen's University Belfast.

[6] Douglas-Cowie, E., Cowie, R. & Rahilly, J. (1994), "The social distribution of intonation patterns in Belfast", in J. Windsor-Lewis (ed.), *Studies in General and English Phonetics, In Honour of J.D. O'Connor*, London: Routledge, pp. 180-186.

[7] Leff, J. & Abberton, E. (1981), "Voice pitch measurements in schizophrenia and depression", *Psychol. Medicine*, vol. 11, pp. 849-852.

[8] Frick, R. (1985), "Communicating emotion: the role of prosodic features", *Psych. Bulletin*, vol. 97, pp. 412-419.

[9] Scherer, K. (1986), "Vocal affect expression: a review and a model for future research", *Psych. Bulletin*, vol. 99, pp. 143-165.

[10] Murray, I. & Arnott, J. (1993), "Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion", *J. Acoust. Soc. Am.*, 93 (2), pp. 1097-1108.

[11] Knapp, M. (1972), *Nonverbal communication in human interaction*, New York: Holt, Rinehart and Winston.

[12] Hammarberg, B., Fritzell, B., Gauffin, J. et al. (1980), "Perceptual and acoustic correlates of abnormal voice qualities",*Acta Otolaryng* 90 pp.441-451

[13] Andreasen, N., Alphert, M. & Merrill, J. (1981), "Acoustic analysis: an objective measure of affective flattening", *Arch. Gen. Psychiatry*, vol 38, 281-285.

[14] Cowie, R., Douglas-Cowie, E., Sawey, M. et al. (1995), "The effects of cochlear implants on speech production in postlingually acquired deafness", *Proc.13th ICPhS*, Stockholm.

[15] McGilloway, S., Cowie, R. & Douglas-Cowie, E. (1995), "Prosodic signs of emotion in speech: preliminary results from a new technique for automatic statistical analysis", *Proc.13th ICPhS*, Stockholm.

SEPARATION OF SPEECH FROM SIMULTANEOUS TALKERS

R.I. Damper, J.R. Thorpe and C.H. Shadle
Department of Electronics and Computer Science
University of Southampton
Southampton SO17 1BJ, UK

ABSTRACT

The separation of speech from two simultaneous talkers is a problem of some practical and theoretical importance. We describe a prototype separation system based on harmonic selection using comb filters. Hermes' subharmonic spectrum method is used to produce a number of (weighted) pitch estimates, with pitch tracks for the two talkers then found by constrained dynamic programming. The system has successfully separated composite male/female /hVd/ tokens but performance is currently rather variable.

INTRODUCTION

The separation of a target speech signal from contaminating, competing signals is a problem of some significance, having applications to improved speech recognition and signal-processing hearing aids. An especially interesting instance of the problem arises when the (single) contaminating source is the speech of another talker. Not only is this a common situation in practice, but separation is likely to be maximally difficult since the target and contaminating signals will share obvious similarities.

Early approaches to this problem [1] were monaural, estimating the fundamental frequency (or 'pitch') of each talker (f_0^1 and f_0^2 respectively), then selecting components of the frequency spectrum and assigning them to a talker according to their harmonic relation to the estimated pitch(es). This *harmonic selection* method assumes that the speech of at least one of the two talkers is voiced, and requires f_0^1 and f_0^2 to be well spaced so that it is obvious which talker is which.

Harmonic selection can be viewed as implementing one of the perceptual grouping principles advanced by Bregman [2], whereby human listeners are able to aggregate auditory features arising from distinct sound sources to effect separation. Other putative grouping principles are based on onset and/or offset synchrony of features, a common rate of amplitude modulation, and cues suggesting a common spatial origin.

Clearly, any implementation of harmonic selection is critically dependent upon a robust pitch detection algorithm (PDA) but most PDAs assume a single voice only [3,4]. More recent work on talker separation [5,6,7] has, therefore, focussed on improved PDAs. However, given that a common spatial origin is likely to be important to grouping, and thereby separation, attention has also been paid to binaural techniques [7,8]. Denbigh and Zhao [7] state that the major advantage of their binaural technique is the ability to recover from talker-allocation errors when f_0^1 and f_0^2 tracks cross.

We describe here the implementation of a prototype monaural separation system which has been successfully applied to the two-talker problem. In the next section, we detail the speech data employed in this study. We then describe the use of Hermes' subharmonic spectrum (SHS) pitch detection algorithm [9] to obtain several

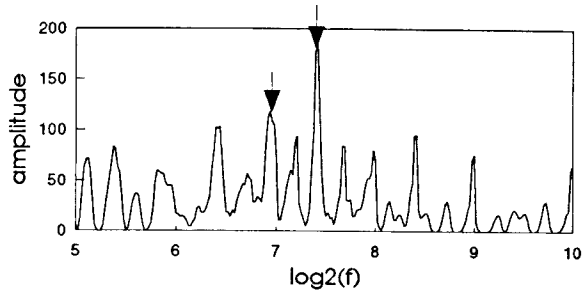


Figure 1. Subharmonic spectrum of typical frame of composite token shown here for $\log_2 f = 5$ *to 10, i.e. 32 to 1024 Hz. Actual fundamental frequencies for the two talkers are shown arrowed.*

weighted estimates of f_0, without consideration of talker identity at this stage. A dynamic-programming (DP) tracking algorithm is then described. This is used to correct pitch errors and to allocate optimal f_0 tracks to each of the two talkers. Results of separation using comb filters are then detailed.

SPEECH DATA

The speech data used in this study were a subset of those recorded by Deterding [10], consisting of /hVd/ tokens spoken by 3 male and 3 female adults and sampled at 10 kHz. A small number of composite tokens was then formed by adding (arbitrarily selected) pairs of male and female tokens. Male/female pairs were chosen to minimise problems of crossing pitch tracks – since the present, prototype implementation is monaural.

Processing was based on frames of 512 samples with 50% overlap. Each frame was multiplied by a Hanning window, padded with a further 512 zeros, and a 1024-point FFT taken. The resulting frequency resolution is, therefore, 9.77 Hz.

SHS ALGORITHM

Hermes' SHS algorithm [9] is an improved version of the harmonic compression PDAs of Schroeder [11] and Noll [12]. These rely on compressing the frequency scale of a spectral representation by integer (harmonic) factors and then taking either the product or the sum of the compressed representations, e.g. Noll's harmonic sum spectrum is defined as:

$$S(f) = \sum_{k=1}^{K} |F(kf)|^2$$

where $F(f)$ is the Fourier spectrum and there are $(K-1)$ compressions. The fundamental f_0 then appears as a peak in the product or sum spectrum, as there is consistent reinforcement of the fundamental by the compressed harmonics.

The problem with these algorithms is that there is a loss of data when used with sampled signals, since certain of the sample points in the compressed spectra fall between those in the original ($k = 1$) spectrum. This severely limits the value of K which can be employed (to about 5). The SHS algorithm avoids this problem by substituting harmonic compression on a linear frequency scale by harmonic shift on a logarithmic scale. Also, the amplitude spectrum (rather than the power spectrum) is used, with decreasing weight given to the more compressed spectra:

$$S(\log_2 f) = \sum_{k=1}^{K} w(k)|F(\log_2 f + \log_2 k)|$$

where here $w(k) = 0.84^{k-1}$ and $K = 9$.

Since the linear-to-log frequency conversion results in logarithmically-spaced

sample points, the spectrum is resampled by cubic spline interpolation at 48 points per octave after conversion. There is also a broadening of spectral peaks at lower frequencies; accordingly, peaks are thinned to a constant width of 3 samples in the log frequency domain. Figure 1 shows a typical subharmonic spectrum with the actual f_0^1 and f_0^2 marked by arrows.

Since even the best PDA will make frame errors, we do not attempt to identify f_0^1 and f_0^2 uniquely at this stage. Rather, the SHS algorithm produces six weighted estimates of possible fundamental for later DP pitch tracking as follows. The 3 largest peaks of the SHS are selected and weighted 1, 2 and 3 respectively. The largest peak (weighted 1) is then assumed to correspond to f_0 for the dominant talker. This estimate of f_0 and its harmonics are then used to subtract corresponding peaks from the thinned Fourier spectrum, and the SHS algorithm re-run to produce 3 new f_0 estimates, again weighted 1,2,3. As a consequence of the use of a log frequency scale, the resolution of the f_0 estimates is non-linear (being $48\log_2 f$).

No distinction is made between voiced and unvoiced speech, both being treated identically.

DP PITCH TRACKING

By maintaining multiple candidate f_0 values, improved pitch estimates can be obtained by dynamic-programming (DP) tracking. We use the method described by Ney [13] which performs a DP optimisation constrained by a (weighted) 'measurement' cost and a 'smoothness' cost.

The input to the DP algorithm is an $n \times m$ time-frequency matrix, where n is the number of possible f_0 values and m is the number of frames in the composite token. Because f_0 is assumed to lie between 32 and 512 Hz, values ≤ 32 are considered to be 0 while values ≥ 512 are considered to be 512 Hz. Hence, there are $n = 48(\log_2 512 - \log_2 32) = 192$ 'frequency indices', according to the logarithmic resolution of the PDA. The cells of the matrix contain the measurement cost, and are initialised to a high value, W_{ini}. The weights of the 6 f_0 estimates ($W \in \{1, 2, 3\}$) from the SHS algorithm are then entered in the appropriate cells.

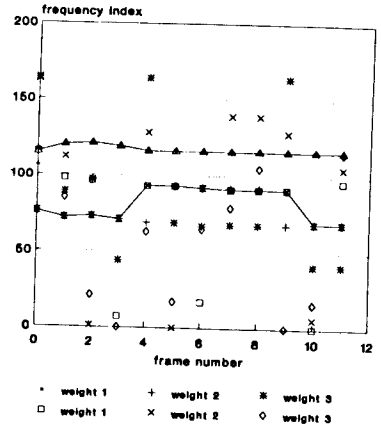


Figure 2. Time-frequency matrix for typical composite token and the pitch tracks (solid lines) for the two talkers found by dynamic programming.

The smoothness cost, D, was implemented as the absolute difference between frequency indices for consecutive frames, so penalising departure from a constant pitch value. The total cost is then the linear combination $W + \alpha D$.

With W_{ini} and α set empirically (at 100 and 0.2 respectively), the DP algorithm was applied to the matrix to find the optimal pitch track for one of the talkers. The values in cells on this track were then replaced by W_{ini} and the algorithm re-run to find the optimal pitch track for the second speaker. This is shown in Figure 2 for a typical composite token.

It is difficult to validate the pitch tracks found. However, use of a commercially-available speech analyser (Kay CSL) gave excellent agreement for one speaker and reasonable agreement for the other.

SEPARATION ALGORITHM

First, the Fourier spectrum is differentiated to find all its maxima, which are listed. The separation algorithm then takes the larger of f_0^1 and f_0^2, and uses this to calculate tentative values for the harmonic frequencies. These are then matched to the list of maxima; if there is no peak at the exact harmonic value, the points either side are checked to see if they are maxima.

Each harmonic peak thus found becomes the centre of one tooth of a comb filter. Each tooth is 5 FFT points wide, and has a Hanning window shape. Multiplication of the Fourier spectrum by the comb filter response then yields a frame of separated data corresponding to the higher f_0. Peaks allocated to this speaker are then deleted from the list of maxima, and the process repeated for the lower f_0.

When this has been done for all frames, separated tokens are produced by overlap-add re-synthesis.

CONCLUSIONS

As judged by informal listening, the prototype separation system works extremely well for some of the composite tokens but less well for others. Separation is better for female than for male talkers – the male separated tokens being more affected by cross-talk. Given the relatively small database used, this may simply reflect lower pitch variation among the female talkers which results in more accurate pitch tracking.

REFERENCES

[1] Parsons, T.W. (1976) "Separation of speech from interfering speech by means of harmonic selection", *Journal of the Acoustical Society of America*, vol. 60, pp. 911–918.

[2] Bregman, A.S. (1990) *Auditory scene analysis*, Cambridge, MA: MIT Press.

[3] Hess, W. (1983) *Pitch determination of speech signals: algorithms and devices*, Berlin: Springer-Verlag.

[4] Hermes, D.J. (1993) "Pitch analysis", in *Visual representations of speech signals*, M. Cooke, S. Beet and M. Crawford (eds.), Chichester, UK: Wiley, pp. 3–25.

[5] Weintraub, M. (1987) "Sound separation and auditory perceptual organization", in *The Psychophysics of Speech Perception*, M.E.H. Schouten (ed.), Dordrecht: Martinus Nijhoff, pp. 125–134.

[6] Stubbs, R.J. and Summerfield, Q. (1990) "Algorithms for separating the speech of interfering talkers: Evaluation with voiced sentences, and normal-hearing and hearing-impaired listeners", *Journal of the Acoustical Society of America*, vol. 84, pp. 1236–1249.

[7] Denbigh, P.N. and Zhao, J. (1992) "Pitch extraction and separation of overlapping speech", *Speech Communication*, vol. 11, pp. 119–125.

[8] Banks, D. (1993) "Localisation and separation of simultaneous voices with two microphones", *IEE Proceedings Part I*, vol. 140, pp. 229–234.

[9] Hermes, D.J. (1988) "Measurement of pitch by subharmonic summation", *Journal of the Acoustical Society of America*, vol. 84, pp. 257–264.

[10] Deterding, D.H. (1990) *Speaker normalisation for automatic speech recognition*, DPhil Thesis, University of Cambridge, UK.

[11] Schroeder, M.R. (1968) "Period histogram and product spectrum: new methods for fundamental-frequency measurement", *Journal of the Acoustical Society of America*, vol. 43, pp. 829–834.

[12] Noll, A.M. (1970) "Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum, and a maximum likelihood estimate", *Proceedings of Symposium on Computer Processing in Communications*, New York: University of Brooklyn Press, pp. 779–798.

[13] Ney, H. (1983) "Dynamic programming algorithm for optimal estimation of speech parameter contours", *IEEE Transactions on Systems, Man and Cybernetics*, vol. 13, pp. 208–214.

SPEAKER IDENTIFICATION USING A SPECTRAL MOMENTS METRIC WITH THE VOICELESS FRICATIVE /s/

R-M. de Figueiredo and S-L. Olivier
Department of Legal Medicine-FCM-UNICAMP, Campinas, Brazil

ABSTRACT

FFT spectra of 26 voiceless fricatives /s/ (10 speakers) were treated as a random probability distribution from which the first four moments were computed. In the first experiment a discriminant analysis based on the four moments resulted in correct classification of speaker identity ranging from 60% to 90%. In a second experiment, a cross-validation test showed that new samples may be correctly matched with the reference material in 8 cases out of 10.

INTRODUCTION

Fricative sounds have already been shown to carry relevant information concerning some of the speakers' characteristcs such as sex and identity [1, 2, 3, 4, 5]. However, the experiments conducted in the studies which proved that to be true concentrate on perceptual evaluation of fricatives produced in isolation and/or in the same phonetic context, neglecting the expected variation intra-speaker in the production of fluent speech. The present work aims at evaluating the efficacy of the fricative /s/ in identifying speakers, using speech samples extracted from fluent reading in various phonetic environments and at two diferent speech rates.

METHOD

Subjects

Ten male subjects aged between 24 and 42 were selected for study. the speakers were free from any speech defect and spoke general Brazilian Portuguese.

Materials and recording conditions

The speakers were asked to read a text extracted from a cientific journal in two different conditions: (a) at a normal and comfortable rate, and (b) as fast as possible, while mantaining intelligibility. Speech samples were all recorded analogically using a high-quality equipment (GRADIENTE Esotech DII tape recorder, and REALISTIC 33984-C microphone) in an acoustically isolated room with no specific reflection characteristics.

Fifty-two voiceless fricatives ocurring in various contexts were extracted from this basic material (twenty-six for each speech rate condition). Only fricatives in CV stressed syllables were used. V is one of the seven Brazilian Portuguese oral vowels (/a, ɛ, e, i, ɔ, o, u/). The words that were analysed in this study were not balanced for vowel context. It means that the number of cases in each vowel context is not, necessarily, the same.

Procedures

The CSL 4300B (KAY Elemetrics Corp.) was used for all acoustic analysis. The signal was digitized by 12-bit ADC at a sampling rate of 25 KHz. Following sampling, a digital high-pass filter with a 200 Hz cutoff frequency was applied to the speech waveform, in order to reduce low-frequency extraneous interference resulting from room vibration.

Only the median third of each fricative was selected for the extraction of each cross-sectional spectrum, in order to minimize any effects of anticipatory co-articulation with the neighbouring vowel. For each [s]-kernel a 512-point fast Fourier transform (FFT) was computed.

After normalized by peak, each FFT-spectrum (only range 0.5-10KHz) was treated as a random probability distribution from which the first four moments (mean, variance, skewness and kurtosis) were calculated. Figures 1 and 2 show pairs of spectra that differ in some of these values.

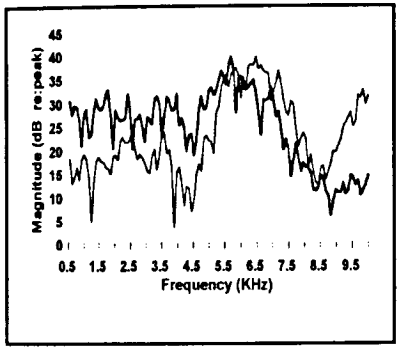


Figure 1: Two spectra, of different speakers, that differ in skewness and mean. The thin-lined spectrum has higher mean and slightly negative skewness, while the thick-lined spectrum has lower mean and markedly positive skewness.

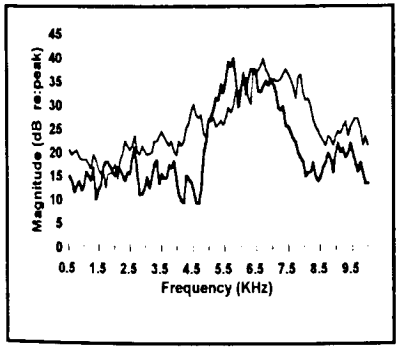


Figure 2: Two spectra, of different speakers, that are basically different for the value of the kurtosis (positive in the thick -lined spectrum and near zero in the thin-lined one)

The values of mean, standard deviation, skewness and kurtosis, derived from the cross-sectional spectra, served as input for a stepwise discriminant analysis accomplished with the program BMDP-7M [6]. The program finds the combination of variables that best predicts the group (speaker) to which a case (represented by the four moments) belongs. At each step, the variable that adds the most to the separation of the groups is entered into the discriminant function. In the end of the process, all variables that, in any way, contribute to the separation of speakers (according to a predetermined minimal F-value) enter the discriminant function.

At a first stage, the efficacy of this spectral moments metric was tested by using only the 26 fricatives produced at a normal speech rate. At a second stage, the results undergo a cross-validation test, in an attempt to classify the 26 fricatives produced under the fast speech rate condition, on the basis of the discriminant functions obtained in the first stage.

RESULTS

Table 1 shows a classification matrix obtained in the first stage of the experiment. The basis for the results found at this point was only the 26 [s]-kernels extracted from the speech samples at normal speech rate. Table 2 shows the results of the cross-validation test. At this point, the newset of 26 fricatives extracted from the fast speech was classified according to the discriminant function obtained in the first test. The observation of table 2 reveals that only two out of ten speakers were not correctly classified (S3-F and S6-F). It should also be noticed that the percentage of correctedness in general decreases considerably in relation to the first test (see table 1), in which only fricatives extracted from speech samples at normal speech rate were employed.

CONCLUSION

The results suggest that voiceless fricatives /s/ in CV stressed syllables are, potentially, good indicators of the

identity of the speaker, even if extracted from fluent speech and in different vowel contexts. Nevertheless, due to expressive alterations in the speed of production, the percentage of correct classification decreases considerably.

It is also important to observe that the efficacy of fricatives [s] in identifying speakers is doubtful in the forensic paradigm, in which the recording quality and bandwidth, both present in this experiment, should not be expected. On the other hand, in speaker automatic verification systems, in which it is possible to control a series of conditions (background noise, sound quality, etc) the use of fricatives seems to be potentially interesting.

Table 1. Classification matrix showing the percentage of cases classified in each group (speaker). The cells in boldface show the percentage of correct classifications.

Subject	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
S1	**92.3**	0.0	0.0	0.0	3.8	0.0	0.0	0.0	3.8	0.0
S2	0.0	**65.4**	7.7	15.4	0.0	0.0	0.0	7.7	3.8	0.0
S3	3.8	7.7	**76.9**	3.8	0.0	3.8	0.0	0.0	0.0	3.8
S4	3.8	7.7	11.5	**61.5**	3.8	0.0	0.0	7.7	0.0	3.8
S5	0.0	0.0	0.0	0.0	**80.8**	0.0	7.7	0.0	11.5	0.0
S6	0.0	0.0	3.8	3.8	0.0	**73.1**	0.0	0.0	0.0	19.2
S7	0.0	7.7	3.8	0.0	0.0	0.0	**88.5**	0.0	0.0	0.0
S8	0.0	3.8	0.0	7.7	0.0	7.7	0.0	**73.1**	7.7	0.0
S9	0.0	0.0	0.0	0.0	7.7	0.0	0.0	0.0	**76.9**	15.4
S10	0.0	0.0	15.4	3.8	0.0	11.5	0.0	0.0	7.7	**61.5**

Table 2. Results of the cross-validation test showing the percentage of classifications of the test group (only samples of fast speech: S1-F, S2-F, etc) in relation to the reference group, based on the discriminant functions obtained in the first phase. The cells in boldface highlight the higher percentage on each line.

Subject	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
S1-F	**61.5**	11.5	0.0	0.0	15.4	0.0	0.0	0.0	7.7	3.8
S2-F	0.0	**34.6**	0.0	26.9	0.0	23.1	0.0	3.8	11.5	0.0
S3-F	**42.3**	3.8	38.5	0.0	0.0	11.5	0.0	0.0	0.0	0.0
S4-F	15.4	23.1	0.0	**53.8**	0.0	0.0	0.0	7.7	0.0	0.0
S5-F	3.8	0.0	0.0	3.8	**57.7**	0.0	3.8	0.0	15.4	7.7
S6-F	0.0	**46.1**	0.0	0.0	0.0	30.8	0.0	19.2	0.0	3.8
S7-F	0.0	3.8	7.7	0.0	7.7	0.0	**80.8**	0.0	0.0	0.0
S8-F	19.2	26.9	0.0	3.8	0.0	7.7	0.0	**42.3**	0.0	0.0
S9-F	0.0	11.5	11.5	0.0	7.7	0.0	0.0	0.0	**46.1**	23.1
S10-F	0.0	0.0	7.7	11.5	0.0	19.2	0.0	0.0	15.4	**46.1**

REFERENCES

[1] Ingemann, F. (1968), "Identification of the speaker's sex from voiceless fricatives, *JASA* 44, 1142-44

[2] Schwartz, M. (1968), Identification of speaker sex from isolated, voiceless fricatives", *JASA* 43, 1178-1179

[3] La Riviére, C. (1974), "Speaker identification from turbulent portions of fricatives", *Phonetica* 29, 246-252

[4] Wu, K. and D.G. Childers (1991), "Gender recognition from speech. Part I: Coarse analysis", *JASA* 90, 1828-1840

[5] Hirson, A. and M. Duckworth (1993), "Glottal fry and voice disguise: a case study in forensic phonetics", *J. Biomed. Eng* 15, 193-200

[6] Jennrich, R. and P. Sampson (1990), "7M: Stepwise discriminant analysis", *BMDP Statistical Software Manual*, Univ. Calif. Press, 339-358

ACOUSTIC CHARACTERISATION OF SPEECH DATABASES: AN EXAMPLE FOR THE SPEAKER VERIFICATION

M. Falcone and U. Contino
Fondazione Ugo Bordoni, Roma, Italy

ABSTRACT

In this paper we propose a simple set of possible measures to describe speech databases in terms of acoustic features. Features may be 'global' or 'target dependent', i.e. they may or may not be functions of the objective of the corpus design. We focus our attention on the specific problem of speaker verification. In particular we analyse the SIVA database, collected over the telephone line in our Institute during the last summer.

INTRODUCTION

Lasting no more than ten years ago, there were no speech databases available. Although the exigency of speech database was a reality also at that time, only with the success of data driven algorithm in speech research (read HMM and NN), the availability of speech database become a need. The industry also promoted and pushed initiatives in this direction as they well know that no commercial applications are possible without large databases. Thanks to CD technology evolution [1], nowadays there are no difficulties in realising and distributing such databases. The pioneer in this field was the TIMIT. Its prototype was available in 1988, and from that time it is a reference in a widespread research and industry sites.

Today we have more than one hundred of CDs as public database; and many others (probably more than one thousand) have been collected for 'commercial' purpose, i.e. for setting up specific voice applications.

So, in conclusion, we now have a lot of databases. But if we are starting a new research, or we are developing a new application, and we understand that a speech database is needed, do we have enough information to make a choice? Probably no!

In fact the description and the characterisation of a database is usually much more expensive than its 'realisation'. For example you may imagine the effort, in term of man power i.e. in term of money too, needed to make (or just to check) manually a transcription.

Speech technology may overwhelm these problems when word-spotting, automatic text alignment, segmentation, etc. algorithms will reach sufficient performance, higher than the actual one. Today these are far from desired target.

On the other side, a speech database, may be characterised under a pure acoustical point of view. As it is a collection of speech *signals*, these may be characterised objectively, without human decision, simple by well defined measures and algorithms.

Generally speaking we may say that there are three different levels of possible description of a speech database:

- *descriptive*: the design of the collection, the population description, the instrumental set-up, etc.
- *annotation*: all the possible annotations and transcriptions, including word transcription, phonetical labelling, prosodic annotation, etc.
- *acoustical*: the measures related to the physical description of the signal.

Excluding the *annotation*, that is often the most important, expensive and difficult one, it will be commendatory that each database has a detailed *descriptive* and *acoustical* description, in order to make clear its possible use.

We shall explore the "sea of speaker verification", that is a small part of the "ocean of speech technology" in this direction, aiming to define a set of possible measures that should be attached to the speech database, in order to give a clear and useful description of the physical characteristic of the signals.

AVAILABLE DATABASES

As speaker recognition, that includes speaker verification and speaker identification, is just a marginal field, there are few public databases on this topic. Here it is a list of what it is available, i.e. the databases utilised in the most important experiments.

For a more detailed description of these see [2].

TIMIT & NTIMIT

Certainly this is the most famous database. Even if it was designed for speech recognition, it has been widely used also in speaker recognition.

Its telephonic version NTIMIT, has a detailed technical description. This is the unique case of acoustic description, that we know, and it is devoted to describe the transformation of the original database in a telephone quality speech database.

KING

It is the first database designed for speaker verification. It is also famous for the "great-divide", an effect related to some variations in the acquisition instruments. The effect is described in term of system performance, and not in relation to the characteristic of the speech signal (that is of course a more reasonable and interesting description).

YOHO

A database collected under a US federal contract in speaker verification. The public version of this database contains compressed speech file. It is not clear the "degradation" (if any) of the speech after the LPC based compression.

SPIDRE

A selection from the most famous "switchboard" database. Also in this case, there is no acoustical description available.

SIVA

The database we collect over the telephone PSTN line last years [3]. It contains 18 repetitions of 20 male speakers. Each session contains a list of isolated words, a dialogue and a read passage, for a total of about 180 seconds.

It is the one used in this work.

ACOUSTIC CHARACTERISATION: A PROPOSAL

Definition and standardisation of acoustical measures in speech are available only for telephonic speech [4]. Many of these may easily be moved to any other kind of speech signal, but the main problem is: which measures must be performed; using which instruments or algorithms; how the results should be grouped and reported; how to create a 'standard report' that will be easy to use and undertake a familiar look.

This is absolutely not a trivial task, and a definitive and comprehensive definition must be validate by the appropriate international commission and institute as CCITT, NIST, etc.

We do not intend here to give an exhaustive contribution. With this paper we only want to promote this initiative, and give a first contribution in this field. The amount of work and of graphical representation we have done cannot, obviously, be shown here; they will be part of the final release of the SIVA database. It is also our intention to run the same procedures on the previous speech databases, in order to identify different characteristics of the signals.

MEASURES: SOME EXAMPLES

The speech signal we use is a standard 8kHz sampled signal, coded with the American mu-law format. All the analysis are executed on a 256 points window, with a 128 points shift, i.e. with half frame of overlapping. Where spectral transformation is used the signals have been preenphatised with a factor of 0.95, and frames have been windowed using the Hamming mask.

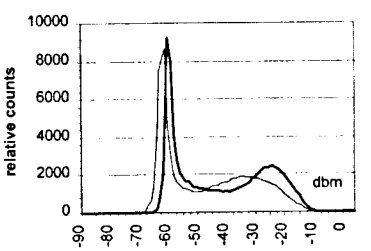


Figure 1. Two speakers' mean energy distribution, speech peaks are at 10dbm

Energy

It is a trivial measure. Nevertheless it is very important that the given values are 'objective', that is no offset is present and the scale reference is correct in relation to the international recommendations. For these reason is quite important that the algorithm respect the CCITT G.711 [5] recommendation, where the numerical values (both in mu-law and a-law) of a

1kHz tone that corresponds to a 0dBm energy are given. Energy normalised histograms clearly give an overview of the recording quality.

These measures should be reported for each session; for each speaker and eventually for 'speaker groups' (e.g. the speakers calling from the same city, or using the same handset, etc.).

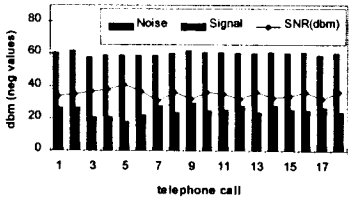


Figure 2. Signal to Noise Ratio (SNR) for one speaker collection

Signal to noise ratio - SNR

It is based on the energy histograms and it is not (unfortunately) an error free measure. More appropriately we can say that it is an estimation, i.e. it is given as the result of the estimation of the mean signal level and the mean noise level. The procedure to measure these mean values range from simple max. estimation to adaptive filtering, and their performance change depending on the speech quality. A human supervision may solve this problem when 'speech' signal to noise ratio is near to the zero value, or when extra signals are added to the speech. Usually, for standard telephone quality signals, automatic methods are adequately. Results may be reported exactly as in the previous case.

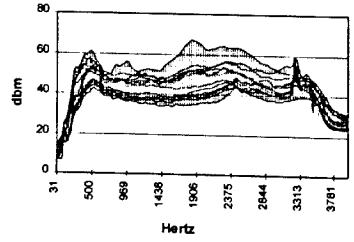


Figure 3. Power spectrum for one speaker, 18 calls existence zone

Long term spectrum

Power spectrum is another classical measure. As for the energy, also in this case it is very important to respect the 'reference signal' so that results may be objectively compared among different databases. This representation is very useful for diagnostic purpose. If the SNR value may insinuate a suspicion, that something is wrong in the signal, the analysis of the long term spectrum will solve in round numbers your doubts. It is difficult to define which kind of 'averages' make sense, as in this specific case a mean over several signals, may mask some important information. So, grouping must be done very carefully and to the averaged spectrum should be added its standard deviation. The first and second order statistical description (mean and standard deviation) of the long term spectra will be, under our experience, sufficient for a diagnostic analysis, if grouping is correct.

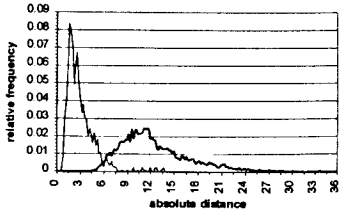


Figure 4. Intra-speaker and inter-speaker variability distribution

Inter-Intra speaker variability

With this measure we are moving towards the specific field of speaker recognition. Inter speaker variability is also important in speaker independent speech recognition, while intra variability is mandatory for speaker dependent speech recognition. In our specific case, they are both crucial. To measure 'variability' a metric, and a matching algorithm must be defined. A plot of an inter or intra speaker variability do not make any sense if the *object*, the *metric* and the *pattern matching* strategy are not defined. Of course comparison between different databases must be done only if these three quantities are the same, otherwise you are not comparing speakers (or signals) but the quality of the speech

model, of the mathematical choices you have done, i.e. your recognition or verification system.

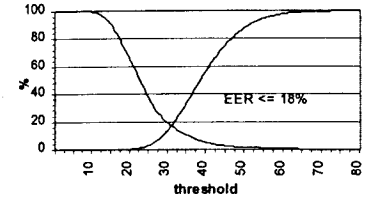


Figure 5. FA and FR using a short utterance (less than 1.5s)

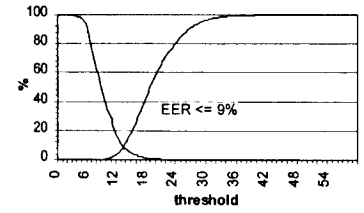


Figure 6. FA and FR using the complete telephone call (about 180s)

"Reference" system performance

It is very important to have an anchor point, a reference that gives the origin of the Cartesian axis where you want to plot the system performance. It seems reasonable that the speaker verification system should be based on the same metrics and parameters used in the inter-intra speaker variability.

Table 1. Performance of the reference system, using 13 LPCC with zero mean

		Training	EER% (FR for FA=0.01)	
		30s	60s	150s
Test	5s	16.5 (47)	12.5 (41)	15.5 (48)
	10s	13.5 (41)	9.0 (25)	13.6 (43)
	15s	8.5 (30)	7.5 (18)	10.7 (32)

As it is always possible to measure inter-intra variability and define a reference system using exactly the same base, it will be foolish to do it in a different way. More difficult is to define reference tests, as usually each database contains a set of files that are not comparable across different databases. The definition of test procedures is probably the most important and difficult point, and we do not address this problem here. In this work we have used a 12th order LPCC parametrisation and a AHSM [6] as distance between two speech samples. We have run several tests; in table one we summarise these reference system performances.

CONCLUSION

Far to define the ultimate recommendations for an acoustical characterisation of speech database, we have outlined the exigency of parting three manifold characterisations of speech databases: one of these is the acoustical description. We suggest a set of possible measure for the speaker verification case, and we report the analysis obtained for the SIVA database. According to the experience we done, these analysis are very useful for the researcher and for the application developer that, starting from the acoustical description, easily obtain a clear and objective view of the characteristic of the database, i.e. check the usefulness of the speech database in relation to his specific purpose.

ACKNOWLEDGEMENT

We would like to thanks all speakers that gave their voices for the SIVA database. They are all friends and colleagues that contribute to the collection without any reward. We are very grateful to all of them for their patience and availability.

REFERENCES

[1] IEEE ASSP Magazine (1985), "*Digital Audio*", Vol.2, N.4

[2] Godfrey J., Graff D., Martin A., Pallet D. (1994), "*Public Databases for Speaker Recognition and Verification*", ESCA Workshop, Martigny, pp.39-42

[3] Contino, U. (1994), "*Una base dati vocale per applicazioni di verifica del parlatore*", (in Italian), FUB Int. Report

[4] CCITT-ITU (1987), "*Handbook on Telephonometry*", Geneve 1987

[5] CCITT Blue Book (1988), "*Pulse Code Modulation of Voice Frequencies*", Rec.G.711, Fascicle III.4, pp.175

[6] Falcone, M., Paoloni, A. (1994), "*Text-Independent Speaker Verification Based on Multiple Reconstruction of Selected Speech Zones*", ESCA Workshop, Martigny, pp.173-176

IDEM: A SOFTWARE TOOL TO STUDY VOWEL FORMANTS IN SPEAKER IDENTIFICATION

M. Falcone, A. Paoloni, N. De Sario
Fondazione Ugo Bordoni, Roma, Italy

ABSTRACT

We introduce the new version of the IDEM system, that is a software package, running under Windows, for speaker identification. The recognition algorithm is, in summary, based on the comparison of a set of parameters, e.g. the pitch and the first three formants of the five vowels /a/, /e/, /i/, /o/, /u/ estimate in the stable portion of speech. In particular we describe the SPREAD module, that is the decision module that performs the identification task.

INTRODUCTION

The IDEM system [1] is a set of software tools to perform speech analysis and speaker identification tests on a personal computer, under the Windows graphical environment. It originally utilised a special high cost professional audio board, but now it works also with the standard audio boards compatible with the MPC definition like the Sound Blaster, it only required that the board support the 16bit audio. It was designed to help operators to carry on a speaker identification test in forensic, and special attention was paid to realise an efficient and simple interface as expert and non-expert people should use this package as well. Several revisions of the product have been released. The present one is the V1.8, but new features and powerful characteristics will be added in the future. An update 32bit version for the NT platform or Chicago software platform is also under evaluation.

OVERVIEW OF THE SYSTEM

The system consists of eight applications plus the acquisition one, that is usually bounded with the audio board or with the operating system of the computer, when MPC workstation is used. In the previous versions of the system the acquisition module was part of the package as it manage, under Windows, a specific audio board, but now it is not part if the IDEM package anymore, as the audio is under the direct control of the operating system. Figure 1 shows all the modules of the latest version of IDEM.

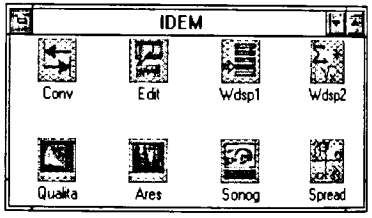


Figure 1. ARES and SPREAD are two applications of the IDEM package

"Conv" is the unique new module; it converts the speech and data file from the old to the new format and vice versa, this make possible to move from the old to the new version without any shocks. For a description of all the modules see [1], here we only describe ARES and SPREAD.

ARES

This module, written in MSVC 1.5, is dedicated to the spectral analysis of fixed length signal window. On the top of the main window a 2.5 second waveform of the audio signal is represented. The cursor is a tick line, wide as the selected zone (you may select any power of two cursor from 128 to 4096 points, default is 512). In the bottom left you have the zoom of the selected window. In the bottom right the power spectrum of the selected window, optionally in this window you can plot the LPC and the CEPSTRUM smoothed power spectrum. Fine tuning is possible by clicking the two buttons on the zoomed signal window. According to the defined number of formant you want to estimate (from one to four, default is three) in the power spectrum window you have some vertical bars, that you may move using the mouse. The position (in Hertz) of the line you are moving is monitored on the left side of the window. Just down the waveform you have plotted two scalar quantities (default are pitch and energy).

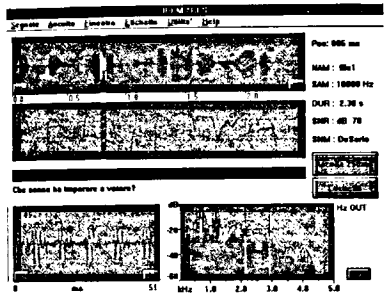


Figure 2. Main window of the ARES module.

Once you have find the signal portion from which you want to estimate the formant value, you had to move the vertical lines on the supposed formant frequencies. Now you may fix (i.e. save in a file) the information that includes the pitch value, the formants values, the vowel, the context of the word.

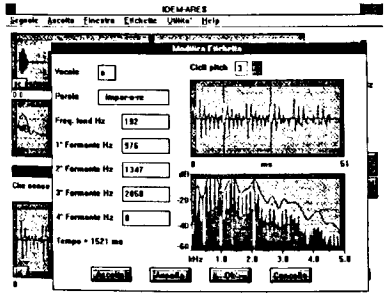


Figure 3. Expanded (modify parameters) window of the ARES module.

The symbol of the labelled vowel will now appear aligned to the audio wave on the screen. If you want to modify, cancel or control any data, just double click on this symbol and a new window where you may read/modify all these parameters appears on the screen.

SPREAD

This application has been written in Visual Basic 3.0. SPREAD (SPeaker Recognition by Automatic Decision) reads the file created with ARES, or with any other manual or automated method that produces compatible files, and let you set up an "*experiment*" where the parameters (in our case the formants of the vowels) associated to each single speaker may be analysed and compared. You can reach the final decision following a simple step procedure: first step, load the data files; second step, check data consistence and eventually data filtering; third step, run statistical decision test and create reports and documentation. SPREAD also contains several tools that can be utilised along the "*experiment*" execution, to obtain a deeper insight of data, i.e. of the formants distribution.

Double clicking on the icon runs the application, and the main window is opened. Only two menus are active at this stage: "Utility" and "Experiment".

Utility

Under the "Utility" you have the help, that follow the Window standard, and the program configuration. It is possible to select the word processor (e.g. Notepad, Word, Write, etc.) to be utilised for document creation and manipulation, as well the symbols that will be associate to the different speakers in the graphical reports. Once you have made your choice, the configuration is automatically saved, the Utility menu will not be utilised furthermore, unless for the 'exit' command that close the session.

Experiment

It must be clear that SPREAD is based on the "experiment" object.

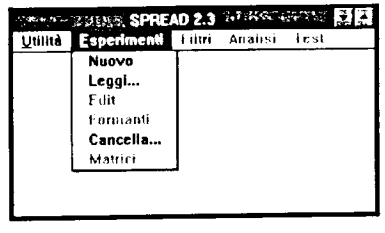


Figure 4. The main window of SPREAD, when run the application

You can work on, modify or delete previous experiments, or create new one. Inside the active experiment you must define: which formants (at least two) you want to use for each vowel; which data files you want to load (you can load and download files as you like it); the reference matrices that model the

population [2]. Once you made your choice and loaded the files, you are ready to the intermediate step 2.

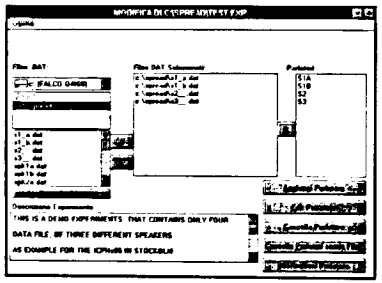


Figure 5. The edit window where the data files are added or deleted inside the experiment.

Filters and Analysis

If you believe that your data are not *clean*, you maybe want to run some filtering. There are five possible filters: one of these is mandatory and it is automatically executed before the test. This filter check data consistence, and looks for missing values or singular matrix, i.e. it checks all the possible causes that make the mathematical procedure senseless. The other filters check for "out of range" values, or performs decimation according to the standard deviation or to the population reference or internal matrix.

Once your data have been validated you may want to look at them. The "analysis" menu has three choices:

- create a report file containing all the information, for each speaker, or for each phoneme (including all data values, mean, standard deviation, occurrence, covariance matrix, etc.);
- plot the data on a Cartesian axis, you may plot the data of any numbers of speakers, for any combination of the used vowels. The variables to plot may be chosen by a selection menu, usually the standard F1 versus F2 plot is used as shown in figure 6. The graph is automatically updated when you click on the menu, so that it works as an interactive graphical environment. It is easy in that way to compare different speakers.

Many other options, as zoom, colour selection, title and legenda insertion, etc. are also available.

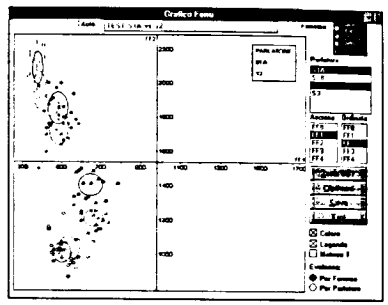


Figure 6. The F1 vs. F2 plot of the four vowels (a,e,i,o) for two speakers

Plots as the previous one only give an immediate overview of the distance among different speaker in a two dimensional space (e.g. F1, F2, or F2, F3, etc.) and their intrinsic limitation must be clear. In fact they may give a wrong indication due to the limitation that only two dimensions are shown.

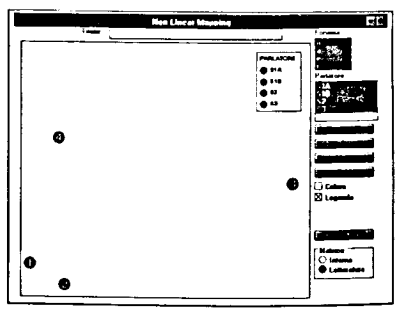


Figure 7. Result of non linear mapping procedure. File 1 and 2, belonging to the same speaker are quite near.

An alternative measure that take into account all the variables [3], may be utilised. In this case you have a mapping of an n-dimensional distance in a two dimensional space. The non linear mapping is an iterative procedure based on a randomised initial disposition of the points. When convergence is reached, speakers with the 'same' voice looks as neighbourhood in a plane space, as shown in figure 7. This representation is useful only for diagnostic purpose, it is not a real test as it is the result of a iterative approximation methodology, in other words it 'may' happens that two

speakers that have short distance in this space are not identified, while others with greater distance are identified.

Test and report creation

The last step is the execution of the identification test. Under this windows is shown a table containing on both axis the different speakers name. Each cell contains the value/result between the indexed speakers. You may have numbers in these cells (e.g. the distance between speaker for a given vowels), or the YES/NO identification result of a CHI-2 or a Hotelling test. In this case you have the default value of p=0.01 but it it possible to change this value by a simple click. In figure 8 the results for the four files we have used in this paper are shown. As the non linear mapping indicated, the file S1 and S2 belong to the same speaker, this according a CHI-2 test with p=0.01, i.e. with a probability of 99% of correct identification.

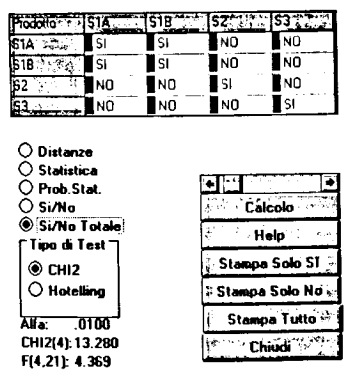


Figure 8. The result of the identification test, is easy to understand

It is also possible to estimate the false identification error, i.e. the probability that a unknown speaker will be identified with some of the speakers used in the experiment. This measure is possible with both analytical and simulated methods, as shown in figure 9.

A set of ASCII reports, with different degree of information, may be automatically created after the execution of the tests. It is also possible to create report for single speaker, or in relation of the test result itself. For example you may have a report that describes only the data that give a positive (or negative) identification score. Many other combinations are possible, but we have no space to describe them here.

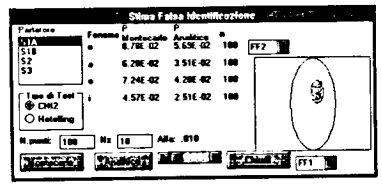


Figure 10. The False Identification Error is computed with analytic method, and with Montecarlo simulation

CONCLUSION

We introduced the last version of the IDEM system. In particular we describe two modules: ARES that let you create data files containing the value of the vowel formants, it has a easy to use interface with audio and visual feedback; and SPREAD that starting from these files let you analyse and filter the data and perform identification tests. This package is now available only for Italian language and it is used, mainly, for forensic purpose. We are currently evaluating a new and powerful (English) version for the Windows NT platform.

ACKNOWLEDGEMENT

We would like thank all that colleagues that contribute to the practical and theoretical development of IDEM, in particular Berardo Saverione and Antonio Federico, as well all the users that gave us useful hints and feedback.

REFERENCES

[1] Falcone, M., De Sario, N. (1994), "*A PC Speaker Identification System for Forensic Use:IDEM*", ESCA Workshop, Martigny, April 1994, pp.173-176

[2] Federico, A., Paoloni, A. (1993) "*Bayesian Decision in the Speaker Recognition by Acoustic Parametrisation of Voice Samples Over the Telephone Line*", Proceedings EUROSPEECH'93, Berlin, September 1993, pp.2307-2310

[3] Sammon, I.W. (1969) "*A Nonlinear Mapping for Data Structure Analysis*", IEEE Trans. on Computer, Vol.18, N.5

A STUDY OF INTRA- AND INTER-SPEAKER VARIABILITY IN VOICES OF TWINS FOR SPEAKER VERIFICATION

M. Mehdi Homayounpour(, **), Gérard Chollet(+)*
(*)CNRS/URA 1027, 19 rue des Bernardins, 75005, Paris, France
(**) Amirkabir University of technology, Hafez Street, Tehran, Iran

(+) IDIAP, C.P. 592, 1920, Martigny, Switzerland

ABSTRACT

This paper deals with the problem caused with similar voices like the voices of identical twins for text independent speaker verification. Three approaches to speaker verification were experimented: i) by human listeners, ii) by comparison of long-term spectra, and iii) by automatic methods [1]. A twin identification test was also conducted. Speaker verification experiments were achieved using LVQ3 and a Second Order Statistical Measure (SOSM). The results show that our automatic speaker verification systems discriminate the voices of identical twins worse than listeners familiar with them. It may be explained by the fact that twins relatives and their friends have received much more speech material for training than our automatic systems.

1. INTRODUCTION

Speaker verification algorithms perform well under controlled conditions, but their performance usually decreases when a user is recorded in other conditions or when he/she is in an emotional or pathological state, or when an impostor, an imitator or a person with a similar voice tries to be verified in his/her place. Twin brothers or sisters have similar voices in most cases. Rosenberg [4] and Cohen *et al.* [6] reported on speaker verification and identification experiments on voices of twins. Rosenberg did experiments with a single pair of twins. In his experiments, his automatic system performed better than human listeners. Cohen found that Cepstra and delta Cepstra yield adequate separation of voices of twins in a speaker identification task. Our experiments concern text independent speaker verification with 11 pairs of identical twins and siblings. These complementary experiments are described in sections 3 and 4. Section 2 specifies the content and recording conditions of the data base used for these experiments. Section 5 compares the results of speaker verification experiments done by human listeners and automatic systems.

2. THE TWIN DATA BASE

A telephone data base was recorded including recordings of 45 speakers consisting of 9 pairs of identical twins (8 males and 10 females) with similar voices, and 27 other speakers (13 males and 14 females) including 4 non-twin siblings. Each twin or sibling spoke for a total of 24 to 30 minutes in three sessions conducted with at least one week interval between sessions. In each session subjects were asked to read three different texts of one page. The speakers called from their office or from their home. Subjects were recorded over the telephone using an OROS AU32 PC-board at 16 bits linear form, 8 kHz sampling frequency.

3. LISTENING TESTS

For the aural method [5], listeners heard pairs of stimuli (55 pairs of 6s stimulus) extracted from the twin data base and decided whether they belonged to the same speaker or not. Two tests were conducted: In test I, there was no pair where both stimuli belonged to a twin pair, while test II included only pairs of stimuli belonging to twins or siblings. Test I was common for all the listeners but test II was different for the pairs of twin. Listeners were familiar or not with the twins or siblings.

Listening tests were conducted for the following purposes:

i) Is it an easy task for the human listeners to discriminate twins? What is the decrease of performance on twins.

ii) Is there a large difference in speaker verification performance when the listeners are familiar or not with the twins?

iii) Are the results of speaker verification by human listeners comparable to those of automatic systems on a twin data base?

In a further test (test III), family members of the twins were asked to listen at each time to a 6s stimulus of one of the twins and to identify him/her by using their a priori knowledge of the twins' voices. The result of this test and test II can serve to verify the hypothesis that when a listener is familiar with the twins (test III), he/she provide a smaller verification error rate (test II). Table 1, present the results of Test I.

Table 1- Results of speaker verification listening tests for test I and test II with Listeners Familiar With Twins (LFWT) and Listeners Not Familiar With Twins (LNFWT). FA and FR are False Acceptance and False Rejection error rates respectively. MER=(FA+FR)/2.

		FR(%)	FA(%)	MER(%)
LFWT	Test I	17.0	14.3	15.6
	Test II	20.2	16.4	18.3
LNFWT	Test I	16.8	14.6	15.7
	Test II	29.4	22.8	26.1

A 8.2% twin identification error rate (test III) was obtained for listeners familiar with the twins. It is much lower than the MER (18.3%) of test II for LFWT. Table 1 shows no bias in our population of listeners since identical results for LFWT and LNFWT are obtained on test I. On the contrary a highly significant difference is found between LFWT and LNFWT on test II. The error rates increase slightly from test I to test II when the listeners are familiar with the twins, while this increase is much more significant for LNFWT. A detailed observation of listening test results showed that when listeners are familiar with the twins, the twins identification error rate is directly proportional to twins speaker verification error rate. These results will be compared to those of our automatic approach in section 5.

4. AUTOMATIC APPROACH

Long-Term Spectra (LTS) and two automatic systems were developed and used for speaker verification. The automatic systems are based on a LVQ3 supervised neural net algorithm [4] and a SOSM measure[2]. 22 subjects comprising 18 twins and 4 siblings were considered as clients and 23 other subjects are impostors for our experiments with LVQ3 and SOSM.

4.1. Speech Analysis

Long silences were first removed. The recordings were pre-emphasised with a first order filter with transfer function of $1-0.95z^{-1}$. Each analysis frame of 30ms was multiplied by a Hamming window that was shifted by 15 ms. A vector of length 24 was retained which comprised 12 LPCC and 12 ΔLPCC [3]. Cepstral coefficients were normalised by subtracting from the cepstral coefficients their averages over the duration of the entire telephone call. This removes any fixed frequency-response distortion introduced by the transmission system.

The ΔLPCC coefficients represent the slope of the time-function of each coefficient in the cepstral vector; so it reflects the transitional information in speech signal. The regression slope is computed over 135 ms. Each coefficient in the feature vector was weighed by the reciprocal of its standard deviation obtained using 2s of training speech from each of the 22 clients.

4.2. Long-term Spectra

The identical twins have an identical, or at least very similar anatomy. So the speech differences between them is more related to their speech habits. This explains why most of our twins showed very close LTS when they were recorded over the same telephone line. LTS was very different when twins were recorded over different telephone line or handsets. Therefor LTS was rejected as a relevant feature to distinguish between twins.

4.3. LVQ3

Two speaker verification tests were conducted using a LVQ3 method adapted for speaker verification [1]. This technique allows to take other speakers into account during the training phase. A codebook for client i, contains three classes: one specifies client i, one for non-client i, and a class of noise and silence. The training data (reference vectors) for each client is obtained using 13.5s of speech from client i, 13.5s of speech from other clients having the same sex and 3.8s

of data representing silence, background and respiration noise. For a client i, the initial codebook contains 160 codes: 64 codes representing the class of client i, 64 codes representing the class of non-client i, and 32 codes representing noise. The initial codes were obtained by the classical LBG vector quantization algorithm using training data and were then tuned with the LVQ3 algorithm as explained in [1]. In the verification phase, the feature vector of a test utterance was compared to all vectors in the codebook and the code label of codebook-vector with the smallest distance to this feature vector was selected. This procedure was repeated for all feature vectors in the test utterance. A verification score was obtained which is equal to the number of testing vectors classified with the label I divided by the total number of vectors in test utterance minus the vectors classified as silence or noise. A speaker was accepted if his/her verification score was higher than a decision threshold, otherwise he/she was rejected.

Two experiments were conducted. They differ in the training phase. In the first experiment (1), training of a model for client i was done with data from any client of the same sex other than i. In the second experiment (2) training was done with 4 closest clients to i excluding twin or sibling. Verification tests were conducted with identical protocols for the two experiments:

x-a: tests on impostors

x-b: tests on twins and siblings.

where x reflects differences in training (x=1, 2). A test utterance duration of 6 seconds was used to conform with human listeners test conditions. The tests on impostors (test x-a) corresponds to protocol I of the listening tests while the tests on twins (tests x-b) is closer to protocol II of the listening tests. The FR obtained from the listening tests is applied to the FR/FA Receiver Operating characteristics Curve (ROC) of each client to find the corresponding FA. The mean of FR and this FA is considered as the error rate for this client (MER1). Similarly the FA of listening test is used to find the corresponding FR error rate and their average (MER2) is averaged with MER1 to find the Mean Error Rate (MER) for this client. The average of total error rates of all twins and siblings for the two sets of experiments is given in table 2. The speaker verification error rates are also presented by Equal Error Rate (EER).

Table 2. Results of speaker verification tests by LVQ3 method (experiments 1 and 2).

	MER	EER
1-a	18.0	13.1
1-b	30.0	30.3
2-a	19.6	21.9
2-b	31.1	34.6

4.4. SOSM

Another set of experiments were done with a SOSM technique [2]. The training speech data of the client i was used to compute a covariance matrix X for this speaker. A weighted symmetric sphericity measure $\mu(X, Y)$ is defined between a test covariance matrix Y and the reference covariance matrix X as the quantity:

$\mu_{SPH\ sym}(X, Y) = A+B$

where:

$A=\rho_{mn} \cdot \log(tr(YX-1)) + \rho_{nm} \cdot \log(tr(XY-1))$

$B= - \backslash F(1;m) \cdot (\rho_{mn} - \rho_{nm}) \cdot \log [\backslash F(det(Y); det(X))] - \log(m)$

with: $\rho_{mn} = \backslash F(m; m+n)$ and $\rho_{nm} = \backslash F(n; m+n)$

where m represents the number of training vectors and n the number of test vectors. For each client an individual covariance matrix was obtained using the same size of training speech material as used for training the LVQ3 models. Table 3 provides the results of the MER error rates obtained for experiments 3-a and 3-b

.

Table 3-Results of speaker verification test by SOSM method (experiment 3).

	MER	EER
3-a	13.5	8.7
3-b	30.0	28.5

4.4. LVQ3/SOSM

SOSM performs slightly better than LVQ3 on the protocol a. No significant difference is found on the protocol b. Both LVQ3 and SOSM show an increase in the verification error rate when a client's twin is considered as an impostor (tests x-b) compared to the case where speakers (non-clients) are impostors (tests x-a). The performance of our automatic systems degrades when a twin brother or sister tries to be verified in his/her place.

This decrease in performance is more important for SOSM method. A comparison of the results of experiment 1 and 2 for LVQ3 shows that when a larger number of speakers are taken into account for training a codebook for a client, a better speaker verification result can be obtained.

5. MACHINE vs. HUMAN

A comparison of Tables 1, 2 and 3, shows that neither human listeners nor our automatic systems are robust against voices of identical twins. Our automatic systems and listeners not familiar with the twins have about the same ability to discriminate between identical twins. The performance of human listeners didn't decrease significantly from test I to test II when the listeners are familiar with the twins. Our automatic systems behave in a way similar to listeners not familiar with the twins. The MER error rates which are obtained by taking into account the listening tests are higher than EER for both systems SOSM and LVQ3 methods. It shows that human listeners familiar with the twins and the two automatic systems studied here present different ROC (Receiver Operating Curve) characteristics.

6. CONCLUSION

Listening tests on twin voices showed generally an augmentation of false acceptance error rate for listeners not-knowing the twins and a smaller increase for listeners being member of the family or friends of twins. Human listeners familiar with twins may proceed with a first level of identification prior to discrimination. Long-term spectrum of speech was not found to be a relevant feature to discriminate between twins. Automatic speaker verification systems use only low level features which are related to the acoustic aspects of speech. The spectral representations of speech such as Cepstrum and delta Cepstrum parameters can not capture the behavioural differences between the twins. So a speaker verification system may take into consideration those features which represent the behavioural characteristics of a speaker to be more robust against the twins with similar voices. More efficient features and/or training procedures remain to be discovered to match the performance of listeners familiar with twins. But, of course, it should be noticed that twin relatives and friends have received much more speech material for training than our automatic systems.

ACKNOWLEDGEMENT

The authors would like to thank the twins and the other speakers and listeners who participated in data base recordings and listening tests.

REFERENCES

[1] M. M. Homayounpour, G. Chollet (1995) , "Neural Net Approaches to speaker Verification: Comparison with Second Order Statistic Measures",, *ICASSP'95.*

[2] F. Bimbot, L. Mathan (1994), "Second Order Statistical Measures for Text-Independent Speaker Identification", *ESCA Workshop on Speaker Recognition, Identification, and Verification*, pp. 51-54.

[3] M. M. Homayounpour, G. Chollet (1994), "A comparison of some relevant parametric representations for speaker verification", *ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, pp. 185-188.

[4] A. E. Rosenberg (1976), "Automatic Speaker Verification: A Review", *Proc. IEEE*, Vol. 64, pp. 475-487.

[5] Homayounpour, M. M., J. Ph. Goldman, G. Chollet, and J. Vaissière (, 1993), "Performance Comparison of Machine and Human Speaker Verification", *EUROSPEECH 93*, p. 2295-2298.

[6] A. Cohen, T. Vaich (1994), "On the Identification of Twins by Their Voices", *ESCA Workshop on speaker recognition, identification, and verification*, pp. 213-216.

[7] T. Kohonen (1990), "The Self Organizing Map", *Proceedings IEEE*, Vol. 78, pp. 1464-1480.

A STATISTICAL TIMING MODEL FOR FRENCH

E. Keller and B. Zellner

Laboratoire d'analyse informatique de la parole (LAIP)
Informatique — Lettres
Université de Lausanne
CH-1015 LAUSANNE, Switzerland

ABSTRACT

Numerous factors influence speech timing. Statistical analysis can identify an order of importance and mutual influences between such factors. A three-tiered (segment-syllable-phrase) model was created by a modified step-wise statistical procedure. It predicts the temporal structure of French of a single, highly fluent speaker at a fast speech rate. The model's predictions correlated with the 1204 syllables of the original corpus at $r = 0.846$.

INTRODUCTION

Research on French speech timing has documented influences at the segmental, syllabic and phrase levels. On the basis of numerous readings of a phonetically balanced short text, O'Shaughnessy [1, 2] proposed a model using 33 rules for the modification of segment duration according to segment type, segment position and phoneme context. For sound classes without prepausal lengthening, the model predicted durations with a standard deviation of 9 ms, yet was less accurate for the prediction of prepausal vowel durations.

The model supposes that timing phenomena can be captured by the segment. However, syllable-sized durations are generally less variable than subsyllabic durations, and may thus represent more reliable anchor points for the calculation of a general timing structure [3, 4]. Furthermore, stress variations and variations of speech rate tend to modify at least syllable-sized units, the syllable may be a psycholinguistic perception unit, and it may also be a minimal unit of rhythm. Syllabic duration can be influenced by the position in the prosodic group, the position in the word, degree of stress, the length of the prosodic group, the position according to the stressed syllable, semantic focus, proximity of syntactic boundaries, the lexical or grammatical status of the word, and emotional factors [5-24]. Some of these may be redundant, e.g., lexeme-final position may be redundant with phrase-final position.

Bartkova [5, 6] added supra-segmental coefficients to her formula for segment durations. Some depended on lexical/ grammatical status and on intra-word position, while others depended on the following consonant, the presence of a syntactic boundary, the presence of clusters, or the syllabic structure near a pause. A comparison of predicted and measured durations in 10 sentences gave a mean difference on segmental duration of ±15 ms. Such a difference can be a handicap for short segments. In our corpus the mean duration for /d/ was 50 ms and a 15-30 ms divergence would correspond to a 30-60% error.

The strategy of this study was to issue from segmental predictions, and to treat syllabic information as additional information. Beyond the syllabic level, word- and phrase-level information was also considered (syntactic, prosodic, rhythmic, intonational groups) [8, 15, 17, 19, 20, 25, 26], in order to account for syllable duration with the smallest number of factors. At each succeeding level, relevant parameters were chosen to explain the greatest proportion of the variance in the residue of the previous analysis. In this manner, a three-tier model based on segmental, syllabic and phrasal information was constructed.

METHOD

The Corpus and Segmentation

A fluent speaker of French was recorded with 100 phonetically balanced sentences. He spoke quite rapidly (6.5 syllables/sec. or more), with a normal, unexaggerated intonation. Acoustic recordings were made in studio conditions on DAT-tape. The digitized data was transferred to computer and was downsampled to 16 kHz.

The time occupied by phonetic segments was labelled with the Signalyze™ program according to a method defined in our laboratory. Specifically, segment transitions were analyzed according to three articulatory levels: labial, lingual and laryngeal. For example, the coarticulatory overlap at the /e/-/s/ transition was marked by symbols representing "onset of frication, associated with the lingual level", followed by "offset of fundamental frequency, associated with a cessation of vocal cord activity". Segmentation reliability was assessed by examining how and where points of transition between inferred articulatory states were marked. Interjudgmental agreement on *robustness* (the application of criteria to state transitions) was scored 1 (low) to 3 (excellent), and agreement on *precision* was scored on 1 (more than two Fo periods difference) to 3 (less than 1 Fo period difference in measurement). Over 50 types of state transitions, there were no cases of low robustness or low precision. The average robustness was 2.53 and the average precision was 2.68.

Analysis and Results

A modified step-wise statistical regression technique for segmental, syllabic and phrase level information was used to develop a model of the speaker's timing behaviour. An issue concerned the calculation of segment duration in a corpus where coarticulatory transition zones are marked explicitly. Is segment duration considered to be the steady-state portion of the signal, or does it include one or both zones of acoustically prominent coarticulatory overlap with adjoining segments? The issue was resolved with reference to durational variation. Since the coefficient of variation over the three zones was systematically smaller (average 0.375) than that of the steady-state zone (average 0.412), the combined duration of the three zones was considered to correspond to "segment duration". Syllable durations were constructed from segment durations by taking into account transitional overlaps (i.e., syllable 2 was overlapped with syllable 1).

The Segmental Model

Raw segment durations were non-normal in their distribution and a log transformation produced a close approximation to a normal distribution. Subsequent to log transformation, segments were grouped according to their mean durations and their articulatory definitions. Eight types of segments could thus be identified. Groups showed roughly comparable coefficients of variation, and an inspection of histograms and normal probability plots showed roughly normal distributions for all classes whose N was greater than 100.

Using the Data Desk® statistical package, a general linear model for discontinuous data (based on an ANOVA) was calculated with partial sums of squares. The following main and interaction factors (up to two-way) were postulated: Duration ($\log_{10}$(ms)) = constant + previous type + current type + next type + previous type * current type + current type * next type + previous type * next type.

Expressed in terms of a Pearson product-moment correlation, the model's predicted segmental durations correlated with empirical segment durations at $r = 0.696$. To test Model 1 in the syllabic context, syllable durations were calculated and were compared to measured syllable durations. The correlation coefficient was $r = .647$ ($N = 1203$, $p<.0001$). The residue from the model (= observed - predicted) was termed "Delta 1" and served as the basis for further factorial modelling at the syllabic level.

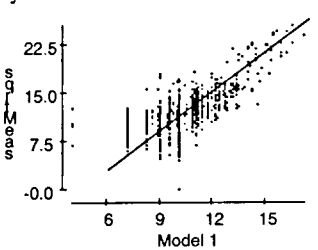


Figure 1. Prediction of the Segmental Model (Model 1): Syllable durations predicted by segmental durations ($r = .647$). Values of Figures 1-3 are in sqrt(ms).

The Syllabic Model

After much experimentation with syllable-level factors described in the literature, a three-factor model, including two-way interactions, was retained for the syllabic analysis: delta 1 = constant + function + position + schwa + function * position + function * schwa + position * schwa, where *"function"* distinguishes lexical vs. function word status, *"position"* identifies three positions in the word, "monosyllabic and polysyllabic-initial", "polysyllabic pre-schwa" and "other", and "schwa" indicates whether or not a schwa is present in the syllable. All main and interaction factors were significant at $p<.05$ by ANOVA.

Syllable durations obtained from the segmental model were additively combined with those for Delta 1 to produce the Syllabic Model (Model 2). Syllable durations showed roughly a square root distribution and were square-root transformed before analysis. Predictions for syllable durations were correlated with transformed observed durations at $r = .723$ (N=1203) (Figure 2). The residual data from this model was termed Delta 2.

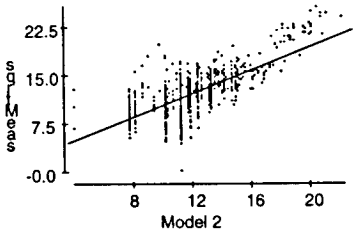


Figure 2. Prediction of the Syllabic Model (Model 2): Syllable durations predicted by segmental durations and syllable-level factors (r = .723).

The Phrase Model

Predictions of Models 1 and 2 showed a noticeable deviation from the regression line in the higher values. Specifically, most syllable durations in the > 280 ms range were underestimated. Furthermore, Delta 2 showed the most pronounced residual error for utterance-final syllables ending in a consonant. A phrase-final correction term was thus calculated for Model 3.

The predictions of Model 3 correlated with the observed square root-transformed syllable durations at $r = .846$ (Figure 3). The residual values from Model 3 varied quasi-randomly around 0. At the present time, it appears that only more sophisticated rules for the generation of the schwa vowel may still be able to improve this model's predictive capacity to some degree.

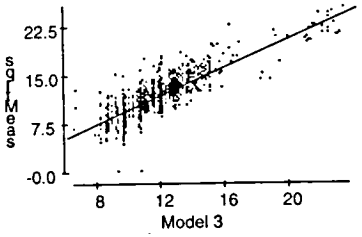


Figure 3. Prediction of the Phrase Model (Model 3): Syllable durations predicted by segmental durations, syllable-level factors and phrase-final lengthening (r = .846).

DISCUSSION

A general model for the prediction of the fast-speech performance of a highly fluent speaker of French was constructed. In view of current discussions surrounding segmental and syllabic contributions to timing models, it is interesting to note that segmental information accounts for a major portion of the variance explained by the model.

The correlation of 0.846 between predictions of Model 3 and the original data set is encouraging. Further improvements in the modelling may come about by the prediction of the presence vs. the absence of schwa, by explicit prediction of speech rate manipulation, and in longer texts, by a better modelling of pauses.

In the present fast-speech corpus, no phrase-level effects other than phrase-final lengthening were identified, in contrast to our findings on the production of French at a normal speech rate, where a systematic increase of lexeme-final syllable durations was observed over the extent of the prosodic phrase [25]. It seems likely that in conditions of considerably accelerated speech rate, our speaker sacrificed some of the "niceties" of phrase-internal timing modulation, and limited himself to a single, phrase-final durational marker.

ACKNOWLEDGEMENTS

Acknowlegements to N. Thévoz, A. Enkerli, H. Mesot, C. Bourquart, N. Blanchoud, and T. Styger. The research is supported by the Fonds National de Recherches Suisses (Projet Prioritaire en informatique, ESPRIT Speech Maps) and by the Office Fédéral pour l'Education et la Science (COST-233).

REFERENCES

[1] O'Shaughnessy, D. (1981). A study of French vowel and consonant durations. *Journal of Phonetics, 9*, 385-406.

[2] O'Shaughnessy, D. (1984). A multispeaker analysis of durations in read French paragraphs. *Journal of the Acoustical Society of America. 76*, 1664-1672.

[3] Barbosa, P., & Bailly, G. (1993). Generation and evaluation of rhythmic patterns for text-to-speech synthesis. *Proceedings of an ESCA Workshop on Prosody* (pp. 66-69). *Lund, Sweden.*

[4] Zellner, B. (1994). Pauses and the temporal structure of speech. In E. Keller (Ed.), *Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State-of-the-Art and Future Challenges* (pp. 41-62). Chichester, UK: John Wiley.

[5] Bartkova, K. (1985). Nouvelle approche dans le modèle de prédiction de la durée segmentale. *14ème JEP* (pp188-191). Paris.

[6] Bartkova, K. (1991). Speaking rate in French application to speech synthesis. *XIIème Congrès International des Sciences Phonétiques,* (pp 482-485). *Aix en Provence. Actes.*

[7] Campbell, W.N. (1992). Syllable-based segmental duration. *Talking Machines. Theories, Models, and Designs* (pp. 211-224). Elsevier Science Publishers.

[8] Delais, E. (1994). Prédiction de la variabilité dans la distribution des accents et les découpages prosodiques en français. *XXèmes Journées d'Etude sur la Parole* (pp. 379-384). Trégastel.

[9] Duez, D., Nishinuma, Y. (1985). Le rythme en français. *Travaux de l'Institut de Phonétique d'Aix, 10*, 151-169

[10] Duez, D. & Nishinuma, Y. (1987). Vitesse d'élocution et durée des syllabes et de leurs constituants en français parlé. *Travaux de l'Institut de Phonétique d'Aix, 11*, 157-180.

[11] Fant, G., Kruckenberg, A., Nord, L. (1991). Durational correlates of stress in Swedish, French and English. *Journal of Phonetics. 19*, 351-365.

[12] Fónagy, I. (1992). Fonctions de la durée vocalique. In P. Martin (Ed.), *Mélanges Léon.* (pp. 141-164). Editions Mélodie-Toronto.

[13] Grégoire, A. (1899). Variation de la durée de la syllabe en français. *La Parole, 1*, 161-176.

[14] Grosjean, F. (1983). How long is the sentence? Prediction and prosody in the on-line processing of language. *Linguistics, 21*. 501-529.

[15] Grosjean, F., & Deschamps, A. (1975). Analyse contrastive des variables temporelles de l'anglais et du français. *Phonetica, 31*, 144-184.

[16] Konopczynski, G. (1986). Vers un modèle développemental du rythme français: Problèmes d'isochronie reconsidérés à la lumière des données de l'acquisition du langage. *Bulletin de l'Institut de Phonétique de Grenoble, 15*, 157-190.

[17] Martin, Ph. (1987). Structure rythmique de la phrase française. Statut théorique et données expérimentales. *Proceedings des 16e JEP* (pp. 255-257). *Hammamet.*

[18] Mertens, Piet. (1987). *L'intonation du français. De la description linguistique à la reconnaissance automatique.* Thèse doctorale, Katholieke Universiteit Leuven.

[19] Monnin, P & Grosjean, F. (1993). Les structures de performance en français: caractérisation et prédiction. *L'Année Psychologique, 93*, 9-30.

[20] Pasdeloup, V. (1988). Analyse temporelle et perceptive de la structuration rythmique d'un énoncé oral. *Travaux de l'Institut de Phonétique d'Aix, 11*, 203-240.

[21] Pasdeloup, V. (1990). *Organisation de l'énoncé en phases temporelles: Analyse d'un corpus de phrases réitérées,* (pp. 254 - 258). 18émes Journées d'Etudes sur la Parole. Montréal, 28 - 31 Mai.

[22] Pasdeloup, V. (1992). Durée intersyllabique dans le groupe accentuel en Français. *Actes des 19émes Journées d'Etudes sur la Parole.* (pp. 531-536). Bruxelles.

[23] Wenk, B. J. & Wiolland, F. (1982). Is French really syllable-timed? *Journal of Phonetics, 10*, 177-193.

[24] Wunderli, P. (1987). *L'intonation des séquences extraposées en français.* Tübingen: Narr, 1987.

[25] Keller, E., Zellner, B., Werner, S., & Blanchoud, N. (1993). The Prediction of Prosodic Timing: Rules for Final Syllable Lengthening in French. *Proceedings,* ESCA Workshop on Prosody (pp. 212-215). Lund, Sweden.

[26] Saint-Bonnet, M. & Boë, J. (1977). Les pauses et les groupes rythmiques: leur durée et disribution en fonction de la vitesse d'élocution. *VIIèmes Journées d'Etude sur la Parole,* (pp. 337-343). Aix en Provence.

THE INFLUENCE OF NATIVE-LANGUAGE BACKGROUND ON SPEAKER RECOGNITION

O. Köster, N. O. Schiller and H. J. Künzel
Trier University, Trier, Germany, Max-Planck-Institute for Psycholinguistics, Nijmegen, The Netherlands, and Bundeskriminalamt, Wiesbaden, Germany

ABSTRACT

The influence of native-language background on the ability of speaker recognition was tested with different groups of subjects: group 1 had no knowledge of the target language (i.e. German), group 2 had some knowledge, and group 3 spoke the target language as its native language (control group). In a direct identification task, subjects had to recognize a speaker's voice with which they were familiarized before. The differences in performance between the groups were significant.

1. INTRODUCTION

In forensic speaker recognition, it sometimes occurs that the voice from a speaker of a foreign language has to be evaluated in a voice line-up or by an expert witness. The question arises in how far this process is influenced by the native-language background of the listener. Human listeners may make use of linguistic information when remembering voices (in addition to purely acoustic information) (cf. [1], [2]). Therefore, it may be the case that the performance in auditory speaker recognition is related to a listener's familiarity with the language under consideration.

Few studies have focussed on the effect of native-language background on speaker recognition. Goldstein et al. [3] found that native American English listeners showed no differences in recognizing speakers with and without a foreign accent and concluded that "[...] voice recognition is just as good (or as poor) for foreign voices as it is for native voices" (Goldstein et al. [3]: 220). Thompson [4] investigated monolingual English natives listening to speech samples from Spanish speakers, native English speakers and English speakers with Spanish accents, and found that the monolingual English listeners identified speakers of their own language best. Goggin et al. [5] tried to quantify the relationship between language familiarity and performance in speaker recognition. They concluded that "[...] voice identification is increased approximately twofold when the listener understands the language relative to when the message is in a foreign language" (Goggin et al. [5]: 456).

In the experiment reported here, we examined the performance of different groups of subjects in a speaker recognition task, with the groups differing in the degree of familiarity with the target language. Additionally, the influence of the voice transmission condition (hifi vs telephone) was tested. This is of primary interest in the forensic situation where most of the recorded speech material is transmitted via the telephone.

2. EXPERIMENT

To test the ability of listeners with a different native-language background in speaker recognition, a direct identification test was designed, in which four different groups of listeners had to recognize the voice of one German speaker in a set of six different German speakers.

2.1. Subjects

Subjects consisted of 53 female and 21 male listeners (n = 74). The age of the subjects was between 16 and 56 years (m = 26.28, SD = 11.85). Subjects were divided into three groups with respect to their knowledge of German. The first group consisted of native English speakers with no knowledge of German at all. The second group consisted of native speakers who had some knowledge of German.[1] The last group included native speakers of German (control group).

The first group of English speakers was further divided into two categories of age: group 1 (n = 15) included all subjects ≥ 30 years of age (m = 47.4, SD = 8.23), group 2 (n = 24) consisted of subjects under 30 years of age (m = 18.42, SD = 3.32). Subjects in both group 3 (n = 18; some knowledge of German) and group 4 (n = 17; German controls) were all under 30 years of age (group 3: m = 21.22, SD = 1.32; group 4: m = 26.28, SD = 3.38).

All subjects took part in the investigation voluntarily. None of them reported any hearing problems.

2.2. Speech material

The speech material used in the experiment was produced by six different male speakers. Speakers were of similar age (m = 29.67, SD = 5.45) and spoke Standard German with Hessian influences. The F_0 of the six speakers ranged from 86 Hz to 142 Hz (m = 109.5, SD = 18.7). All speakers had to read a small German text of approximately one minute in length onto a DAT recorder. Then three parts of the text between four and eight seconds in length were spliced out of the recordings of every speaker. To record exactly the same material under telephone transmission conditions, the speech samples were recorded again through a telephone line. Each of the six speech samples was re-recorded three times. In total, we obtained 108 speech samples[2]. All of the speech samples were randomized and re-recorded on DAT.

One speaker was designated as speaker X, the target voice. From speaker X, the hifi text was re-recorded on DAT five times to obtain a speech sample of approximately five minutes.

2.3. Method

All four groups of listeners were tested individually. Firstly, subjects were familiarized with the voice of speaker X by listening to speaker X's five minutes speech sample. Subjects were instructed to concentrate on the voice in order to try to memorize it. After this familiarization, response sheets were handed out to the subjects. After a short break of approximately five minutes, the subjects were given a forced-choice test. They were instructed to listen to the tape with the randomized speech samples carefully. After each sample the sujects marked "Yes" if they thought the voice was from speaker X and "No" if it was not. There were five seconds between each stimulus which the subjects considered to be enough time to make a decision. After every tenth speech sample, there was a sine tone of 300 Hz to help subjects to keep track of the task.

3. RESULTS

The design of the experiment allows to differentiate between two error categories: subjects could either reject the target voice speech sample when it actually came from speaker X (false rejection; FR) or identify a speech sample as the target voice when it was in fact produced by one of the dummy speakers (false identification; FI). Furthermore, FRs and FIs were split into the errors made under the hifi vs telephone transmission conditions to see whether there was a difference.

3.1. False rejections vs false identifications

If subjects were randomly identifying the speaker, we would expect an FRs to FIs error ratio of 1:5 (18 target voice samples compared to 90 dummy samples). The observed error ratios fall below the expected value in all four groups: group 1 made 67 FRs (m = 4.4, SD = 2.5) and 256 FIs (m = 17.07, SD = 14.11) (ratio = 1:3.82), in group 2 there were 141 FRs (m = 5.88, SD = 5.18) and 163 FIs (m = 6.79, SD = 8.09) (ratio = 1:1.16), in group 3 there were 26 FIs (m = 1.44, SD = 2.43) and 39 FIs (m = 2.17, SD = 4.07) (ratio = 1:1.5), and group 4 made 24 FRs (m = 1.41, SD = 1.97) and 37 FIs (m = 2.18, SD = 2.71) (ratio = 1:1.54).

[1] Subjects of group 3 were students of German; they took part in a university exchange program and had already been in Germany for several months when the experiment was run.

[2] 3 parts of the text x 2 transmission conditions (hifi vs telephone) x 3 repetitions x 6 speakers = 108 speech samples.

χ^2-tests revealed that the FR to FI error ratios fall significantly below the expected value of 1:5 in all four groups (group 1: $\chi^2 = 18.16$, $df = 1$, $p < .001$, group 2: $\chi^2 = 416.69$, $df = 1$, $p < .001$, group 3: $\chi^2 = 63.7$, $df = 1$, $p < .001$ and group 4: $\chi^2 = 57.41$, $df = 1$, $p < .001$). The respective error proportions are given in figure 1.

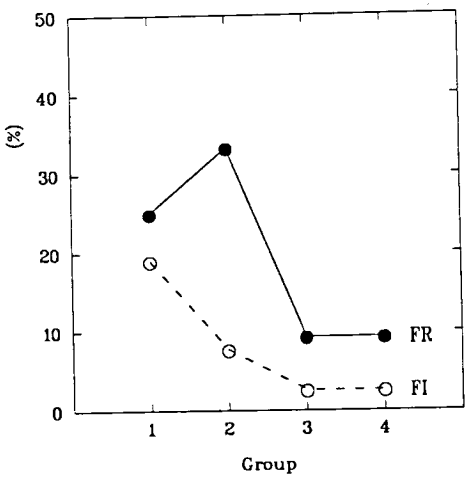


Figure 1. Error proportions for FRs and FIs, respectively.

To see whether there were differences in the amount of errors (FRs and FIs) between the four groups ANOVAs were carried out with the degree of knowledge of German as the dependent variable. The data therefore were arcsin transformed (cf. Winer [6]: 400). The results were highly significant for both the FRs, $F(3, 70) = 7.85$, $p < .001$, and the FIs, $F(3, 70) = 11.897$, $p < .001$.

Post-hoc analyses (pairwise comparisons; *Scheffé* tests) revealed that with respect to the FRs, group 2 made significantly more errors than either group 3 ($p = .003$) or group 4 ($p = .004$). Concerning the FIs, group 1 made significantly more errors than any of the other groups (group 2 $p = .007$, group 3 $p < .001$ and group 4 $p < .001$). None of the other differences between the four groups were significant.

3.2. Hifi vs telephone transmission condition

The ratio of speech samples recorded under hifi vs telephone transmission conditions was 1:1 (54:54). Within the 18 target voice samples (9:9) and the 90 dummy samples (45:45), the respective ratios were also 1:1. The expected error ratio both for FRs and FIs was therefore 1:1. Again, the observed ratios deviated from this *a priori* value in different ways (see tables 1 and 2).

group	FRs	hifi	teleph.	ratio h:t
1	**67**	**25**	**42**	**1:1.68**
2	**141**	**44**	**97**	**1:2.21**
3	**26**	**13**	**13**	**1:1**
4	**24**	**9**	**5**	**1:0.56**

Table 1. FRs in the two transmission conditions for the groups.

group	FIs	hifi	teleph.	ratio h:t
1	**256**	**99**	**157**	**1:1.59**
2	**163**	**71**	**92**	**1:1.3**
3	**39**	**16**	**23**	**1:1.44**
4	**37**	**5**	**32**	**1:6.4**

Table 2. FIs in the two transmission conditions for the groups.

With respect to FRs, group 2 made significantly more errors when the speech sample was recorded over the telephone ($\chi^2 = 9.96$, $df = 1$, $p < .005$). The difference between the number of errors for group 4 reached only marginally significance ($\chi^2 = 4.08$, $df = 1$, $p < .05$). But note that in this case there were fewer errors for the telephone transmission condition.

Concerning FIs, all four groups made fewer mistakes in the hifi condition. Significance was reached for group 1 ($\chi^2 = 6.57$, $df = 1$, $p < .025$) and for group 4 ($\chi^2 = 9.85$, $df = 1$, $p < .005$).

4. DISCUSSION

All four groups made significantly fewer FIs relative to FRs than could be theoretically expected. This means that, on the average, subjects were inclined not to identify a speech sample as coming from speaker X. This leads to the interpretation

that subjects were in general quite prudent with identifying a voice as the one from the target speaker. This result is in contrast to the result obtained by Künzel [7]. Künzel tested the speaker recognition abilities of linguistically naive listeners and found that in his groups on the whole subjects showed the tendency to identify two speech samples as coming from the same speaker even when this was not the case (cf . Künzel [7]: 35).

As the statistical analyses revealed, there were significant differences in performance in the speaker recognition experiment between the four groups. The results indicate that unfamiliarity with the target language affects the ability to recognize a speaker, as subjects with knowledge of German performed generally better than subjects without any knowledge of German. It seems that speaker recognition does not only involve purely phonetic features, but also incorporates linguistic information. The results further permit the interpretation that the degree of knowledge of the target language seems to be of less relevance because group 3 and 4 performed equally well.

The influence of the listeners' age on the performance in speaker recognition remains rather unclear. Whereas the younger subjects of group 2 made fewer FRs than the older ones of group 1, the situation is reversed with respect to the FIs; here, group 1 made significantly more errors than group 2. This last result is in accord with Künzel ([7]: 54) who found that the amount of FIs rose with increasing age.

The effect of the acoustic quality of the speech samples was investigated by recording the speech samples under hifi vs telephone transmission conditions. The speech signal is reduced to the bandwidth interval between 300 and 3400 Hz when transmitted over German telephone lines and contains additional noise. On the whole, performance was worse when the speech sample was recorded via the telephone. The only exceptions were the ratios of groups 3 and 4 for the FRs (see table 1). This leads to the interpretation that the acoustic quality of the speech sample is very important for speaker recognition purposes. In accord with what Künzel ([7]: 26) found, it seems that in the speech samples recorded via the telephone some speaker specific features that help in voice recognition are missing or obscured.

Acknowledgements

The authors would like to thank Bernadette Schmitt and Jan-Peter de Ruiter (both Max-Planck-Institute for Psycholinguistics, Nijmegen, The Netherlands) for helpful comments on the statistical analyses and Julie Christiansen (Max-Planck-Institute for Psycholinguistics, Nijmegen, The Netherlands) for proof-reading the paper. The authors take responsibility for all remaining mistakes.

5. REFERENCES

[1] Ladefoged, P., Ladefoged, J. (1980), "The ability of listeners to identify voices", *UCLA Working Papers in Phonetics*, vol. 49, pp. 43-51.

[2] Lorch, M. P., Meara, P. (1989), "How people listen to languages they don't know", *Language Science*, vol. 11, pp. 243-253.

[3] Goldstein, A. G., Knight, P., Bailis, K., Conover, J. (1981), "Recognition memory for accented and unaccented voices" *Bulletin of the Psychonomic Society*, vol. 17, pp. 217-220.

[4] Thompson, C. P. (1987), "A language effect in voice identification", *Applied Cognitive Psychology*, vol. 1, pp. 121-131.

[5] Goggin, J. P., Thompson, C. P., Strube, G., Simental, L. R. (1991), "The role of language familiarity in voice identification", *Memory & Cognition*, vol. 19, pp. 448-458.

[6] Winer, B. J. (1971), *Statistical principles in experimental design*, second edition. New York et al.: McGraw-Hill, 1971.

[7] Künzel, H. J. (1990), *Phonetische Untersuchungen zur Sprecher-Erkennung durch linguistisch naive Personen*, Stuttgart: Steiner (Zeitschrift für Dialektologie und Linguistik, Beihefte; 69).

A SEMI-AUTOMATED LX-BASED METHOD FOR THE MEASUREMENT OF VOICE ONSET TIME

Krzysztof Marasek
University of Stuttgart, Institute of Natural Language Processing
Experimental Phonetics, Stuttgart, Germany

ABSTRACT

The main goal of this study is to establish a reliable automated method of Voice Onset Time (VOT) estimation. It is shown, that VOT measurements can reliably and accurately be performed by combining the information about stop release from acoustic signal with the information about voicing initiation derived from the laryngographic signal. This task can be performed automatically.

PREFACE

VOT, defined as the time difference between stop release of a plosive and the start of vocalisation of the following vowel, is a common parameter in the investigation of speech and language disorders [2]. The proposed method uses two-channels recording of the speech signal. The Laryngograph was used to monitor the activity of the vocal folds (Lx signal) with the acoustic speech signal simultaneously recorded on the second channel (Sx signal). The starting time instant of vocal fold vibration is based on the Lx signal analysis, while the closure release impulse is found in the Sx signal.

THE Lx SIGNAL

The Laryngograph [3] enables direct measurement of vocal fold activity. Thanks its non-invasive measure method (laryngeal conductance measured by pair of electrodes situated on the neck, both-sides of the cricoid cartilage) and mostly high SNR ratio, the laryngograph is often used as a reference signal for pitch period determination. Nonetheless, the output from laryngograph (Lx) is not free from problems: it is influenced by vertical movements of the larynx (so called Gx signal) and it does not match some movements of the vocal cords. For some speakers the Lx regristration may even fail temporarily. The amplitude of the Lx depends on speech loudness. However, as it was confirmed in [1] the Lx signal matches exactly the vocal fundamental frequency. The Lx signal is also used in more detailed analysis of vocal folds activity. Of special interest is the use of the Lx signal to differentiate pathological modes of phonation. To achieve that, an undistorted form of the Lx should be used, i.e. the influence of the Gx should be eliminated, but the distortion of the shape of the Lx waveform shuld be avoided at the same time.

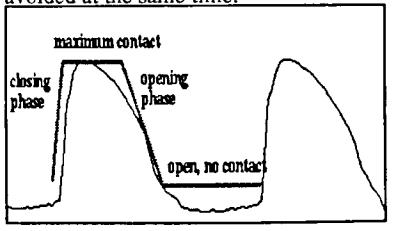


Figure 1. Phases of the Lx signal.

The individual pitch periods of speech signal are determinated in the Lx signal through position of the peak change of the laryngeal conductance, which is adequate to the instant of glottal closure. The changes of the conductance during one pitch period are presented on Fig.1. When the glottis is open, the conductance is low and flat. During closing phase, the conductance rises steeply and remains high during closure. Then, during the opening phase of the vocal folds the conductance is falling, but not so steeply as during the closing phase. The position of the maximum change of the conductance is determinated by zero-crossing and/or by thresholding of differentiated Lx signal. Previous to the use of this method it is necessary to pass the Lx signal through low-pass filter to avoid the influence of the Gx signal. The vertical movement of the larynx may be fast, so it is not easy to determine cut-off frequency of the filter. As it was pointed by Baken [1] such filtering may strongly influence the shape of laryngographic waveform, making it unsuitable for

further analysis. Hess and Indefrey [4] proposed more sophisticated algorithm (with very good temporal resolution), but it also needs filtering of the Lx signal and fails in case of rapid vertical movements of the larynx.

The proposed algorithm works on the raw, unfiltered Lx data. The algorithm may be divided into 3 steps:

1. The markers are set at the positions of local maxima. Markers are set only when the local maximum occurs after given time (thresholding in the time domain) and next samples differs significantly in amplitude to the maximum (thresholding in the amplitude). Then the positions of local minima are found, also with tresholding in time and amplitude domains. Further analysis is based on the pair of markers: maximum-minimum. The temporal difference between them may be used as pitch period estimate. The positions and amplitudes of maxima and minima are compared to their neighbours and, when they are significantly weaker and shorter, they are attached to stronger pairs (such situation occurs in creak-like or laryngalized phonation).
2. The parameters of whole record of the data are taken into account in the second pass of the analysis. The pairs of markers which occur in isolation or in very short train of markers (<4) are recognised as an error and removed if they occur between long unvoiced segments of speech (at least 200 ms). Differences in the pitch periods length are analysed and if the length quotient is greater than 5, an algorithm tries to recover the min-max pairs from the original signal. Thereafter the begining and the end of each voiced segment is marked. In every voiced segment each maximum-minimum pair is again analysed to find disturbances from mean length and mean relative amplitude within the pitch periods. If deviations are greater then given threshold, the weaker pairs are connected to the stronger ones (if a resulting pair is not too long regarding mean pitch length). The min-max pairs segment the Lx signal according to pitch period length.
3. The glottal shape encoded in the Lx waveform is established on the third pass of analysis. Based on minimum-maximum pairs the time instants of the opening, the open, the closing and the close phases of the glottis are found. To find the time instant of the approximated begining of the closing phase the 3-point smoothed Lx waveform is analysed and the point of maximum of the first difference is chosen. The starting point of opening phase is more difficult to find, especially when rapid movements of the larynx occur. It is assumed, that the opening starts at the same level of conductance as at the begining of the closing phase, so the next start of the opening (point 5 on Fig.2) is found as the crossing point of Lx waveform with straight line connected to the closing phase markers (dotted line between points 2 and 5 in Fig.2).

Every period of Lx signal is than described as presented on Fig.2.

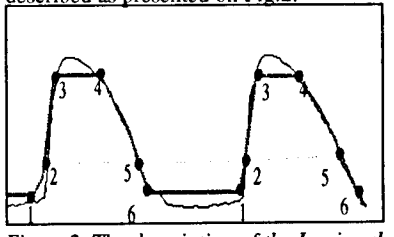


Figure 2. The description of the Lx signal using straight lines.

The shape is described using six idealized straight lines and deviations from those lines are used as indices for the signal classification.

THE STOP RELEASE

The impulse characteristic the stop release has to be found in speech channel of the record.

The direction of search for stop release depends on the form of the speech signal after the start of the vocalisation. In fact, the vocalised friction phases of plosives contains only low-frequency component. Thus, the zero-crossing (ZC) rate for every pitch segment is very low, distinguishing it from following vowel segments (after F2 release). The decision is made based on the ZC rate for the first 10 periods of the speech signal. If ZC rate is low (and its variation is also low) it is assumed, that the VOT will be negative. The closure impulse for the negative

VOT forms a short (and rather weak) noise-like burst on the top of speech signal (Fig.3). In order to find this impulse the differentiated speech signal is consulted, but only on its positive parts. The segment with the highest ZC-rate points the period with the closure (short noise burst). Within the founded segment the greatest jump in the amplitude of the speech signal points to the release of the impulse. Additional conditions on length and strength of the burst prevent accidental determination of the VOT.

The search for the stop release in positive direction is based on the difference in the signal energy between the silence (occuring before closure) and other phases of the plosive. In fact, between the begining of the stop release and the start of the following vowel (i.e. in the burst phase) some noise is present, thus short time energy shows a rapid step indicating the start of the plosive (given appropiate SNR and the initial position of the plosive). Thereafter, in a window where the energy changes most rapidly, the time index of the sample with the biggest difference is determinated as the starting point of the closure release. As a validation, the energy between segments before and after the stop release is compared (the following one should be bigger than the preceding one).

RESULTS

The method was tested on recordings done using the Laryngograph processor (Lx) and small electret microphone (Sx). The recordings included 3-4 logotomised words (like : /baba/, /papa/, /gaga/). The method was tested not only on normal speakers with modal voices, but also on patients (with neurological disorders) showing some voice disoders (breathy creaky voice). The recordings of the patients have substantially lower quality as the control ones, especially regarding the SNRs. The VOT was measured only in the initial position. The results are summarised in Table I.

DISCUSSION

As can be seen from Table I the results, althought quite good for so complicated signals, are not fully satisfactory regarding the percentage error. The most errors were caused not by troubles in the perfect localisation of the closure release impulse, but rather by the imprecise localisation of the start of vocal fold vibration. Within the vocalic segment, the first one or two periods of vibration are destroyed, their amplitude and duration is irregular and non-stationary (see Fig.4). To overcome this, a kind of soft-tresholding (the parameters are used with additional weights) in the amplitude and the time domain was used to find the begining of the Lx-vibration. This method was quite sucessfull for control speakers (the one major error within this group was caused by intentionaly unnatural, very long negative VOT) but failed for other groups of speakers, whose speech was very slow and quiet. It was observed, that for so quietly speaking persons, the Lx signal was distored or even lost for some moments. The Lx signal exceeded also the permited range of the A/D converter due to rapid movements of the larynx (swallowing). The minor errors (smaller than 1 ms, typically about 0,5 ms) are caused by distrurbances in location of the closure impulse. The overall description of the Lx signal, however, performed well and the single periods were precisely located.

CONLUSIONS

It was shown that VOT measurements can reliably and accurately be performed by combining the information about stop release from acoustic signal with the information about voicing initiation derived from the laryngographic signal [2] and that task can be performed automatically.

References.

[1]BakenR.J.(1992),Electroglottography, J. of Voice, Vol.6, No.2, 98-110.

[2] Blumstein S., Cooper W., Goodglas H., Stalender S., Gottlieb J. (1980), Production Deficits in Aphasia: A Voice-Onset Time Analysis, Brain &Language 9:153-170.

[3] Fourcin A.J. (1993), Normal and pathological speech: phonetic, acoustic and laryngographic aspects, in: Singh W., Soutar D., Functional Surgery of the Larynx and Pharynx, Butterworth Heinemann.

[4] Hess W.J.(1992), Pitch and Voicing Determination, in: Furui, Sondhi M.M. (eds.) Advances in Speech Signal Processing, Marcel Dekker Inc., 3-49.

Table I. The results of VOT measurements for different groups of speakers. The numbers given in square brackets [] describe segments, where VOT was found, but in reality there was no plosive at the begining of the voiced segment.

Group of voices	Number of VOTs	No. of small errors (< 1ms)	No. of big errors (> 1ms)	No. of not or false recognised VOTs
control - modal voice	15	3	1	[1]
aphasia - modal voice	12	4	3	1
dysarthia - creaky voice	12	5	3	2
parkinsons - breathy voice	12	4	3	[1]
Σ	51	16	10	6[2]
% errors		31	19	11(5)

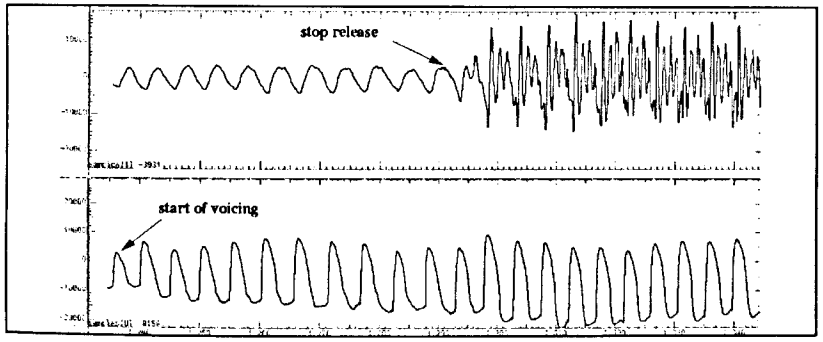


Figure 3. An example of the negative VOT.

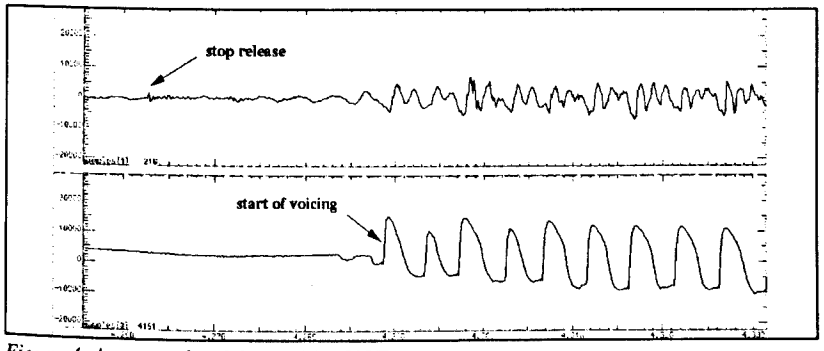


Figure 4. An example of the positive VOT.

CONTROLLED ELICITATION AND PROCESSING OF SPONTANEOUS SPEECH IN VERBMOBIL

M. Pätzold, M. Scheffers, A. Simpson and W. Thon
IPDS, Kiel, Germany

ABSTRACT

Within the framework of the German Verbmobil project, a large amount of spontaneous dialogue data had to be collected. This paper describes the recording environment and the means of elicitation and transcription which have been developed at Kiel to fulfill this task.

INTRODUCTION

The ultimate goal of the Verbmobil project [1] is the development of a portable translation system with voice input and output.

The aim of data collection in the first phase of Verbmobil was to provide a large amount of German spontaneous dialogues associated with appointment making. The data should consist of high quality speech signals together with their orthographic and phonemic transcriptions. Part of the signals should also be segmented and labelled. The dialogues should be elicited in a controlled situation, but still be as spontaneous as possible.

Part I of this paper describes the technical details of the recording environment and the signal processing developed to meet the requirements imposed on the speech recordings. Part II describes the elicitation of appropriate material and the subsequent steps involved in its transcription and segmentation.

PART I: SIGNAL RECORDING AND PROCESSING

The following requirements were imposed on the speech recordings:

- Synchronous capture of the speech signals of two dialogue partners.
- High quality recording (low background noise level, large dynamic range).
- The actually recorded "turns" should not overlap in time.

We furthermore needed to reckon with a recording session lasting up to an hour. The end product should be a series of speech files each containing one "turn", arrived at with as little manual labour as possible.

Recording Environment

To meet these requirements, a hardware/software recording environment has been implemented with the following features (see Figure 1):

- The dialogue partners are placed in separate sound-treated rooms. They communicate via headsets.
- The speech signals are recorded directly to hard disk into a multiplex stereo file (2x16bit/16kHz), on a PC AT486/66 platform with about 500MB disk space, sufficient for recording sessions in excess of one hour.
- Both microphone signals are recorded on DAT for backup purposes.
- A DSP Motorola 96002 controls the high quality analogue I/O channels as well as the digital I/O.
- The dialogue is controlled by speak-buttons and lights. Both speakers may request their input channel by pressing their button. Requests are granted on a "first come, first served" basis. Service is indicated to a speaker by his light being turned on. Thus, only one speaker's signal is recorded at a time.
- The experimenter may at any time during the recording session communicate with the dialogue partners via his microphone without interfering with the signal recording.

The DSP programme controls the A/D and D/A conversion and monitors the button actions.

If no buttons are pressed, a zero signal is output to both headphones, a distinct constant marker signal is recorded on both channels and both lights are turned off.

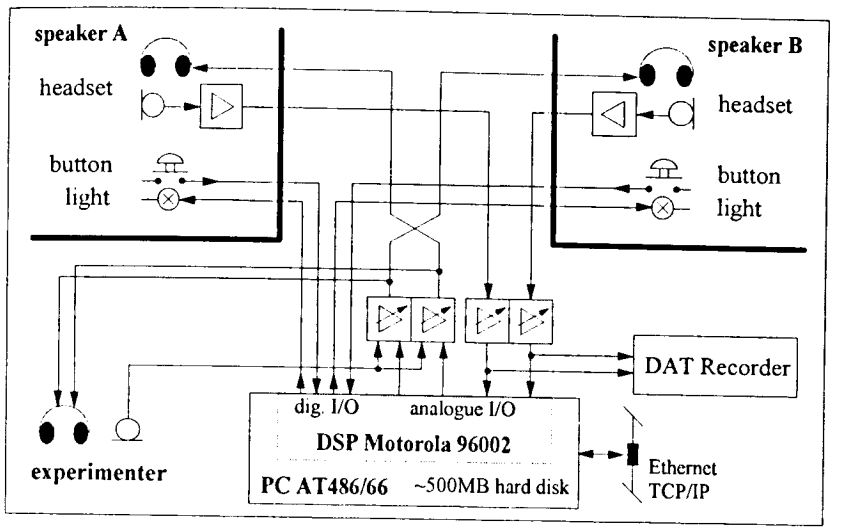


Figure 1. Recording environment: hardware set-up.

As soon as the first speaker presses his button, his speech signal is routed to his partner's headphones and recorded on his channel, his light is turned on until he releases his button.

If, during this time, the other speaker presses his button, his light remains turned off, a zero signal is still output to his partner's headphones, but a different constant marker signal is recorded on his channel.

Thus, the marker signals recorded on both channels reflect the exact timing of button actions. Furthermore, the DSP programme continuously checks the input sample levels and signals these to the PC programme.

Running in parallel, the PC programme does the actual data transfer to the hard disk and provides the user interface to the experimenter. During a test session a peak level display may be used to optimize the recording level for the extreme dynamic range of spontaneous utterances.

Signal Processing

The resulting data file is transferred via Ethernet to a UNIX workstation, where it is de-multiplexed and split into two files, one for each channel. At the same time the embedded markers are detected and converted to a list with the time intervals of the turns. After inspection and, if desired, correction of the interval markers (e.g. because a speaker has released his button for a short time within his turn), a second programme uses them to split the two files into the desired series of turn files. Starting from the original multiplexed file, the names of the respective output files are generated automatically and contain at the end stage a code for the dialogue scenario (e.g. appointment making), a code for the recording site, a recording identification number and information on the channel (speaker) and the position of the turn in the dialogue. Finally, a programme is available to convert the files from local format to the delivery format.

PART II: ELICITATION, TRANSCRIPTION AND SEGMENTATION

In this section we describe the elicitation of appropriate material and the subsequent steps involved in its transcription and segmentation.

Elicitation

The recordings had to contain the following material:

- names of months
- dates
- names of days
- names of holidays
- times
- deictic time expressions
- proper names
- names of towns
- spelling

To guarantee systematic coverage of the material the following scenario was developed [2].

Each speaker was given a set of calendar sheets each covering a two-month period, together with timetables covering the weekdays. The calendar sheets and timetables were placed face down in a pile in front of the speaker together with a pen for making notes.

Apart from the names of months, dates and the names of days, the calendar sheets also contained names of holidays, exemption blocks (shaded areas representing days on which the speaker could not make an appointment) and simple appointments. The timetables had the names of days, times and exemption blocks.

The calendar sheets and timetables served to elicit the names of months, the names of days, times and the names of holidays. Appointment entries in the calendar sheets were designed to elicit names of German towns, e.g. "Dienstreise nach Kiel" ("Business trip to Kiel").

In order to make speakers utter letter names, appointments had to be arranged at an exhibition ("IAA in Frankfurt") and at a conference ("ICPhS in Stockholm").

Finally, deictic time expressions were elicited using a portion of a timetable. The names of the days were left out and the speakers were told that the first day on the timetable was today, and that three meetings had to be arranged over the next two days, i.e. today, tomorrow and the day after tomorrow.

Each recording session was split up into eight tasks. Each task involved the speakers arranging three appointments in the period specified on a calendar sheet. The appointments were noted at the bottom of the calendar sheet and also briefly explained by the experimenter.

The first seven tasks allowed the twelve months of the year to be covered with six calendar sheets. The first task was used as a dummy to get the speakers accustomed to the set-up and enable recording levels to be set. The eighth task involved the elicitation of deictic time expressions using the cut-down timetable.

Before the first recording, the speakers were instructed on the tasks and on the use of the speak-button.

Orthographic Transcription

The transcription system provides an orthographic representation of the dialogues [3]. The system must fulfill two requirements. First, it must be simple to allow for a relatively fast transcription of a large amount of data. Second, it must attempt to meet the demands of both signal processing and linguistics.

As well as transcribing the lexical content, the system must also capture characteristic aspects of spontaneous speech.

For lexical items and semantic-syntactic structure the transcription is based on the Duden conventions [4] as far as possible. In addition, the following objects typical for spontaneous speech are included in the system:

- interjections
- agreement and negation particles
- particles indicating hesitation
- non-words (neologisms, slips of the tongue)
- laughing, coughing, lip-smacking etc.
- articulatory lengthening
- breathing and pauses
- breaks and repairs
- stretches of utterance, either poorly understood or not understood at all by the transcriber
- commentaries on idiosyncrasies in a speaker's production, stylistic and dialectal forms, etc.
- non-articulatory noises (finger-tapping, rustling of paper, etc.)
- interruptions in the recording, caused by incorrect use of the speak-button

- numbers
- abbreviations and spelling

The transcription symbols for these objects are chosen such that the objects can readily be identified and easily be distinguished from the lexical items.

Segmentation and Labelling

For segmenting and labelling the signals, the orthographic transcription is converted into a phonological transcription representing the canonical pronunciation of the utterance. The canonical transcription is used as the basis for a list of labels. These are to be time-aligned with the signals and where necessary modified to indicate differences from the canonical form (deletions, insertions and replacements).

Segmentation is discrete and exhaustive. The segmentation of each signal file begins at the onset of utterance and ends with the cessation of utterance. The placement of a label simultaneously symbolises the beginning of one segment and the end of the previous segment.

```
g097a003.s1h
ANS003: das +/is<Z>=/+ freut mich , da"s
        Ihnen das pa"st <A> <#Klicken> .
oend
 d a s+ Q l s z: =/+ f r 'OY t m l C+ , d a s+
 Q i: n @ n+ d a s+ p 'a s t h: . :k
kend
 c: d -h a s+ Q- -q l s z: =/+ f r 'OY t m l C+
 , d -h a s+ Q- -q " i: n @- %n+ d -h a s+
 p -h 'a s t -h h: . :k
hend
` 1809 #c:
  1809 ##d
  2994 $-h
  3379 $a
  3720 $s+
  4959 ##Q-
  4959 $-q
  4959 $l
  5662 $s
 10424 $z:
 10424 $=/+
    .   .
    .   .
```

Figure 2: Example of a label file. From top to bottom: orthographic transcription, canonical form, the modified labels after segmentation and part of the labels with their sample numbers.

The product of the segmentation and labelling is a text file containing the orthographic and canonical phonological transcriptions and a list of (modified) labels and their times (see Figure 2).

The system for segmenting and labelling was originally developed for read speech [5]. It has been extended to include labels and conventions for the objects introduced for spontaneous speech. As with phonetic-phonological labels, the new ones are time aligned with events in the signal [6].

In addition, a system for prosodic labelling is at present being developed [7].

ACKNOWLEDGEMENT

This work was partially funded by the German Ministry of Education, Science, Research and Technology (BMBF) in the framework of the Verbmobil Project under Grant 01IV101M7. The responsibility for the contents of this study lies with the authors.

REFERENCES

[1] Karger, R., Wahlster, W. (1994), *VERBMOBIL Handbuch,* Verbmobil Technisches Dokument 17, Saarbrücken: DFKI.

[2] Pätzold, M., Simpson, A. (1994), *Das Kieler Szenario zur Terminabsprache,* Verbmobil Memo 53, Kiel: IPDS.

[3] 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 11, Kiel: IPDS.

[4] *Der Duden,* (1991), 20th ed., Mannheim, Wien, Zürich: Dudenverlag.

[5] Kohler, K.J. (1994), *Lexica of the Kiel PHONDAT Corpus: Read Speech,* vol. I, AIPUK 27, Kiel: IPDS.

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

[7] Kohler, K.J. (1995), "PROLAB - The Kiel System of Prosodic Labelling", *Proc. XIIIth ICPhS.*

AUTOMATIC VOWEL QUALITY DESCRIPTION USING FOUR PRIMARY CARDINAL VOWELS

Shuping Ran, Phil Rose, Bruce Millar and Iain Macleod*
Computer Sciences Laboratory
Research School of Information Sciences and Engineering
Australian National University, Canberra, ACT 0200, Australia
** Linguistics Department, Faculties, ANU*

ABSTRACT

This paper investigates the possibility of describing vowels phonetically using an automated method. Models of the phonetic dimensions of the vowel space are built using two multi-layer perceptrons trained using four primary cardinal vowels. Test vowels processed by these perceptrons are placed onto a cardinal-like vowel chart. These automatically derived positions are compared with the positions of these vowels in a similar space as judged by a phonetician, and with the acoustic space derived from these vowels. The differences observed are discussed.

INTRODUCTION

Vowels are described in phonology and traditional phonetics with the three major parameters of height, backness and rounding, as well as additional parameters like nasality and tenseness. Although backness, height and rounding are often defined articulatorily, it is now widely assumed following Ladefoged [1] that the labels are primarily acoustic or perceptual, and relate to perceptually motivated transforms of F_1 (height) and effective F_2 (backness and rounding).

Vowels are traditionally described by phoneticians by listening to the vowels, and then placing a vowel symbol onto the cardinal vowel chart or assigning it appropriate diacritics according to learned auditory models. Figure 1 illustrates a three dimensional cardinal vowel system. This traditional method is very tedious, and is not feasible for non-phoneticians. This paper investigates the possibility of describing vowel quality without the skills of an experienced phonetician, using a novel method which automatically places a given vowel into a space which is defined by a set of reference vowels.

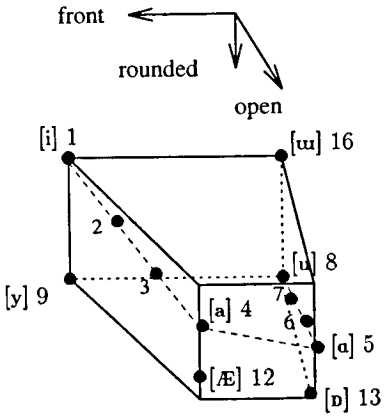


Figure 1. A three dimensional model of the vowel space (after Ladefoged [2])

A preliminary study [3] was carried out in which the vowels of four speakers of Australian English were analysed by this method. Models of each speakers' vowel space were trained using three reference vowels from an existing data corpus to encode the form of acoustic evidence for phonetic features which correlate with the dimensions of the vowel space (e.g. open-close, front-back). The reference vowels were chosen according to their relatively extreme positions on the cardinal vowel chart and their stability within Australian English. While the results of this study

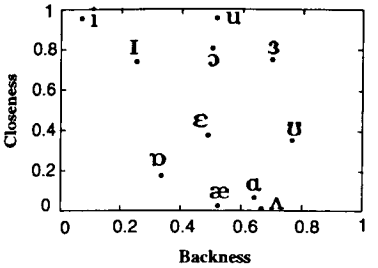


Figure 2. Test results averaged over six stop contexts of eight reference vowel model: 11 pseudo steady state vowels on a Closeness versus Backness plane.

were encouraging, it was clear that the choice of the reference vowels was crucial for more accurate positioning of the vowels on the vowel chart.

In a further study [4], eight cardinal vowels which represented the extremes of the dimensions: front-back, open-close, rounded-unrounded, produced by an experienced phonetician were used for the model training. English vowels in stop consonantal context produced by the same speaker were used for testing. The results showed that the method worked well with respect to the front vowels, but badly for the back vowels (see Figure 2). It was suspected that this result was due to the lip rounding of some reference vowels introducing some misleading information into the models.

In the present study, we aim to minimise this potentially misleading information by choosing a different set of reference vowels.

REFERENCE VOWELS

The reference vowels used in this study were derived from the vowel model expressed by Figure 1. The aim was to use primary cardinal vowels that were maximally extreme on the two dimensions of front-back and open-close. The four primary cardinals (vowel 1 [i], 4 [a]; 5 [ɑ] and 8 [u]) fit this specification.

Five repetitions of each primary cardinal were recorded in a sound booth by our speaker, who is an experienced phonetician trained in the British tradition. The reference utterances were hand segmented. The parts of the signal where F_0 remained stable were used for this study. An $F_1/(F_2-F_1)$ plot was made of these vowels from conventional wide band spectrograms, as shown in Figure 3.

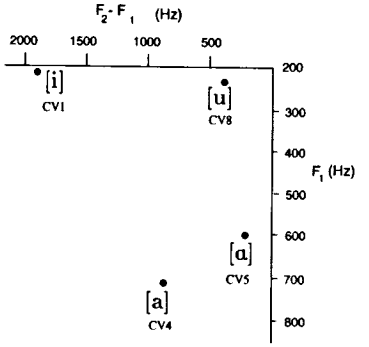


Figure 3. F_1 vs F_2-F_1 plot for phonetician's cardinal vowels CV 1 4 5 8.

ENGLISH VOWELS

Five repetitions of English vowels in the context of [stop][vwl]d utterances were produced by our speaker, where: [stop] represents one of the six phonemically voiced and voiceless labial, alveolar, and velar plosives of English (/b, p, d, t, g, k/); [vwl] represents one of the eleven nominally monophthongal phonemes (/i, ɪ, ɛ, æ, ɑ, ɒ, ɔ, ʊ, u, ʌ, ɜ/); and d is /d/. The [stop][vwl]d utterances were manually segmented and labelled according to the procedures described by Ran [5]. Only the pseudo steady-state vowel interval was of interest for this study.

These vowels were transcribed by the phonetician, and placed on a traditional chart showing height and backness, with rounding indicated separately -- see Figure 4. This figure shows an unremarkable auditory configuration typical for the British English accent of the speaker, with some apparent influence from Australian English. Thus the /u/ is considerably

fronted ([ʉ] >); the /ɔ/ is a close-mid [o]; the /ɛ/ is closer than open-mid, and the /ɜ/ is closer and more front. An $F_1/(F_2-F_1)$ plot of the English vowels from conventional wide band spectrograms also reflects this pattern (Figure 5).

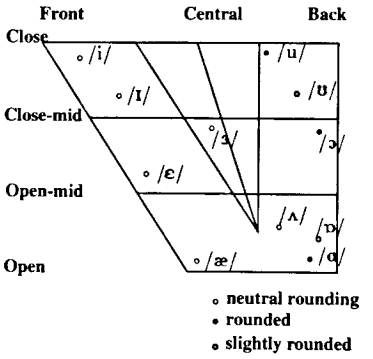


Figure 4. English vowel description by a phonetician.

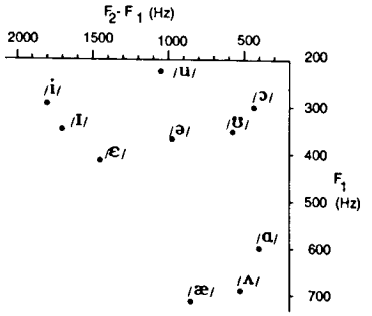


Figure. 5. F_1 vs F_2-F_1 plot for phonetician's English vowels in b-g *context.*

DATA PRE-PROCESSING

The data, including the pseudo steady-state English vowel intervals and the reference vowels, were processed in 'frames' of 12.8ms, with adjacent frames having a 6.4ms overlap, by passing them through a Hamming window, and then deriving 13 Linear Predictive Cepstral Coefficients (LPCCs) for each frame.

MODEL TRAINING

Two Multi-Layer Perceptrons (MLPs) were used to model the articulatory dimensions of front-back and open-close in order that they may be used as articulatory descriptors for backness and closeness. Each MLP was implemented with one hidden layer of two nodes and was trained by using the back-propagation algorithm. The inputs for this training comprised frames of four repetitions of the four reference vowels, and comparator outputs were their articulatory labels as shown in Table 1.

cardinal vowel	articulatory description	back	close
i1	front-close	0	1
u8	back-close	1	1
ɑ5	back-open	1	0
a4	front-open	0	0

Table 1. Articulatory labels for the reference vowels.

MLPs with one hidden layer were used because they are theoretically able to encode relationships of any complexity [6]. The number of hidden nodes was chosen by experiment starting with one hidden node, then incrementing the number one by one. The architecture which gave best performance on the training data was chosen. The number of hidden nodes for the backness and closeness descriptors was two.

VOWEL DESCRIPTION RESULTS

The cepstral data of the pseudo-static intervals of the English vowels were processed by the trained articulatory descriptors (i.e. the closeness descriptor and the backness descriptor), on a frame by frame basis. The outputs from the descriptors were the activation scores of the output nodes of the MLPs, which indicated with what probability a given input frame can be labelled with the articulatory label of the descriptor.

Figure 6 reports the results by combining the output from the two descriptors.

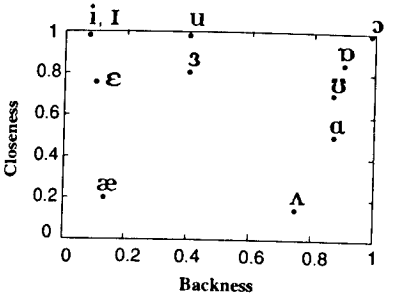


Figure 6. Test results averaged over six stop contexts of four reference vowel model: 11 pseudo steady state vowels on Closeness versus Backness plane.

The horizontal axis represents the backness, where the left represents maximal frontness and the right represents maximal backness. The vertical axis represents the closeness, where the top end represents maximal closeness and the bottom end represents maximal openness.

Analysis of Figure 6 reveals that, compared with the phonetician's auditory judgements (Figure 4) and the F_1/F_2-F_1 plot (Figure 5), the automatic method using the four reference vowels resolves the English vowels well. The positioning of the vowels approximates more closely to the positioning in the F_1/F_2-F_1 plot than to the positioning of the auditory judgements, especially for the back vowels.

Because of the restrictions of space, test results in individual contexts are not included here. The resolutions appear to be rather sensitive to differences in the consonantal frame. It can only be assumed that differential consonantal assimilation is occurring which is currently being studied.

Comparing the test results of eight reference vowel models [4] with that of four reference vowel models, the latter has improved substantially the description of the vowels, specially with respect to the back vowels. One noticeable problem is that some vowels (/i, ɪ, u, ɔ/) are positioned on the extremity of the maximum closeness which is unrealistic.

CONCLUSIONS

This study arose from our concern to improve the reference vowel set over that used in [4]. The results have clearly shown improved vowel positioning by choosing four primary cardinal vowels as reference vowels instead of all the eight cardinal vowels. The method provides a normalised system of automatic phonetic quality description. The challenges that remain include further understanding of the impact of consonantal context on the method and ways of accounting for it. It is also important to find ways of training naive speakers to produce reference vowels which may then be used to normalise automated phonetic description of their vowels.

REFERENCES

[1] Ladefoged, P. (1982) *A Course in Phonetics*, Second Edition, (Harcourt Brace Jovanovich:New York).

[2] Ladefoged, P. (1975), *Three Areas of Experimental Phonetics* (Fourth edition), (Oxford University Press, London).

[3] Ran,S., Millar,J.B., Macleod,I. (1994), "Vowel quality assessment based on analysis of distinctive features", *Proc. International Conference on Spoken Language Processing*, Yokohama, pp. 399-402.

[4] Ran,S., Rose, P., Millar, J. B. and Macleod, I. (1994), "Automatic vowel quality description using a cardinal vowel reference model", *Proc. of the Fifth Australian International Conference on Speech Science and Technology*, pp. 387-392.

[5] Ran, S. (1994), *Speech Knowledge Modelling for Speech Recognition: A Study Based on Distinctive Features*, PhD thesis, The Australian National University.

[6] Lippmann, R. P. (1987), "An introduction to computing with neural nets", *IEEE Trans. on Acoustics, Speech and Signal Processing*, 4(2), pp. 4-22.

VOWEL CLASSIFICATION BASED ON ACOUSTIC AND ARTICULATORY REPRESENTATIONS

Alain Soquet[1] and Marco Saerens[2]
[1]Institut des Langues Vivantes et de Phonétique, [2]IRIDIA,
Université Libre de Bruxelles

ABSTRACT

The objective of this paper is to compare different acoustic and articulatory representations on a vowel classification task. Classification results were obtained based on linear discriminant analysis and decision trees algorithm with cross-validation on the speakers. The cepstrum, the formants and the articulatory representations achieve similar performances with linear discriminant analysis. The decision tree algorithm provides accurate classification rules for the formants and the articulatory representations. The resulting articulatory rules are consistent with our knowledge on vowel production and could be efficiently used in knowledge-based systems.

INTRODUCTION

A problem of long-standing interest in speech analysis and recognition concerns the most appropriate representation for acoustic-phonetic decoding. In this work we will focus on vowels; results for plosives place of articulation identification can be found in [1]. Vowels are traditionally described in term of static spectral characteristics or articulatory configurations.

From the acoustic point of view, global spectrum descriptions in term of a small set of coefficients (for example the LPC coefficients) can be used. However, the standard representation for vowels consists in the first resonance frequencies of the vocal tract (the formants frequencies). It is well known that, even if the first formants frequencies are efficient cues to classify vowels, there exist an important speaker variability – for example, the differences between male and female speakers [2]. Moreover, in fluent speech, the target vowels are not always reached when produced in a consonant context [3]. Nevertheless, this phenomenon does not introduce any degradation in the human recognition capabilities [4]. A large amount of work continues however to focus on static description of vowels (see for example [5] and [6]).

From the articulatory point of view, one can describe a vowel by the configuration of an articulatory model that produces similar spectral characteristics. Unfortunately, the computation of the articulatory configuration from the acoustic parameters (the acoustic-to-articulatory inversion) is not a trivial problem. In previous work [7], we developed a tool that realizes this inversion in the framework of an articulatory model, based on the first three formant frequencies.

Our aim in this work is to compare several acoustic and articulatory representations on a vowel classification task.

ACOUSTIC REPRESENTATIONS

The speech signal was passed through a 5 kHz cutoff low-pass filter, and sampled at 10 kHz. The signal was then preemphasized ($1 - 0{,}95\ z^{-1}$) before further processing. Six different acoustic representations have been chosen. Three of them are directly computed from the speech signal, and are widely used in statistical speech recognition systems (e.g. HMM). The three remaining ones are related to formant frequencies, prevalent in knowledge-based recognition systems.

- **LPC** (LPCA): The LPC coefficients were computed with the autocorrelation method on a 25,6 ms frame multiplied by a Hamming window. The number of poles of the predictive filter was fixed to 12.
- **LPC cepstrum** (LCPS): The LPC cepstral coefficients were derived from the predictive coefficients obtained with an LPC analysis [8]. As before, we used the first 12 coefficients.
- **Cepstrum** (CPST): Cepstral coefficients were computed from a 16 ms frame multiplied by a Hamming window. The first 12 coefficients of the cepstrum were used in order to describe the spectral characteristics of the signal at the measurement point.
- **Formants** (FORM, BARK, MEL): The formant values were extracted semi-automatically on the basis of the different acoustic representations. We used 3 different scales for the frequency axis: Hertz (FORM), Bark [9] (BARK), and Mel [10] (MEL).

ARTICULATORY REPRESENTATIONS

Four articulatory representations were selected. The first one is computed from the LPC coefficients. The other three correspond to the control parameters of three different articulatory models. These control parameters are provided by a neural network performing the acoustic-to-articulatory inversion on the basis of the first three formant frequencies [7].

- **LPC area** (LAREA): The LPC area functions are computed from the LPC reflection coefficients as suggested by [11].
- **DRM** (DRM): The distinctive regions model [12] is an 8 regions acoustic tube with transversal control. The model is derived from acoustic properties of the uniform acoustic tube. The control parameters are the sections of the 8 regions. The length of the tube was kept constant (18 cm).
- **Maeda** (MAEDA): Maeda's model [13] is an articulatory model derived from X-ray sagittal cuts. A set of 7 parameters controls the shape of the sagittal cuts.
- **Lin Fant** (LF): Lin and Fant's model [14] is a geometrical model with longitudinal control. There are 3 main control parameters (two for the principal constriction, and one for the lips).

EXPERIMENTS

In order to study the effectiveness of these different representations for the classification of vowels, a set of vowel-consonant-vowel (V^1CV^2) was recorded (where C is one of the six plosives [p, t, k, b, d, g], and V^1 or V^2 one of the five vowels [a, œ, i, u, y]). The resulting 150 VCV were recorded by 11 male speakers, giving a total of 1650 tokens and 3300 vowels. The 10 representations are computed in the stable part of the vowel V^1 and V^2.

In a first experiment, vowel recognition results were obtained based on linear discriminant analysis [15] with cross-validation on the speakers: the tokens from each individual speaker are successively removed from the training set, and used as a test set. The results can therefore be considered as speaker-independent. Each training set consists in 3000 vowels, and each test set in 300 vowels. The results are presented in figure 1.

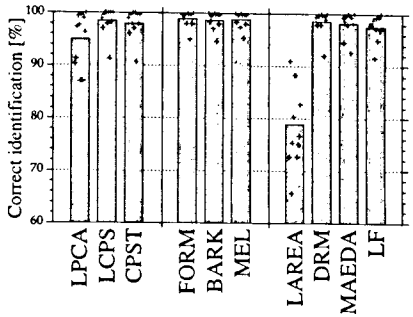


Figure 1: Results of discriminant analysis for the ten representations showing percent of correct classification: results for each test speaker (crosses) and averaged performance (in grey).

The formant based representations obtain the best performances on average. Their scores are, however, comparable with those of other acoustic representations – like LCPS – which have higher dispersions. We can observe that LPCA obtains lower performances.

The performances of LAREA – less than 80 % and a very high variability among the different speakers – are significantly lower than the other representations. On the contrary, the three other articulatory representations present performances comparable to the formant cues. Their dispersions are larger than for FORM but remain lower than for LCPS. On average, LF is less performant than DRM and MAEDA.

We observed that the vowel giving the largest amount of errors is [a], often confused with [œ]. This result is illustrated in figure 2 showing the scatter plot of the 3300 vowels on the two first discriminant axes for 4 representations.

In a second experiment, we used a decision trees algorithm named C4.5 [16]. This technique allows to build a tree that classify the data with a succession of tests involving just one attribute. The tree is

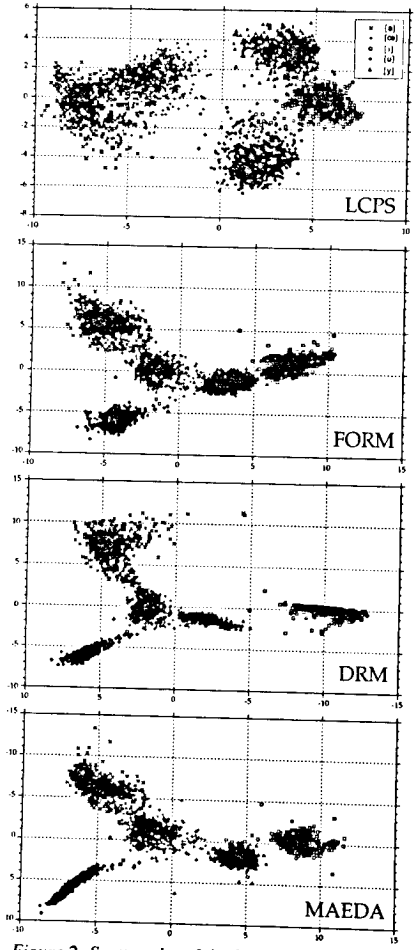


Figure 2: Scatter plot of the 3300 vowels on the two first discriminant axes for LCPS, FORM, DRM and MAEDA.

then pruned so that it becomes both simpler and more accurate on unseen cases. Finally, the algorithm generates a production rule classifier that is usually as accurate as the pruned tree, and more easily understood by people. This algorithm has been applied to four representations: LCPS, FORM, DRM and MAEDA. As for the discriminant analysis, the performances were obtained with cross-validation on the speakers. The results are presented on figure 3.

LCPS obtains bad performances in

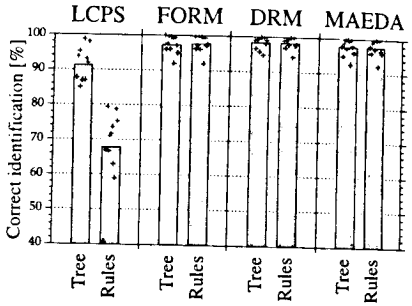


Figure 3: Results of decision trees and rules for four representations showing percent of correct classification: results for each test speaker (crosses) and averaged performance (in grey)

comparison with the other three representations. The performances of LCPS are even worse when using the rules. This result can be explained by the size of the trees generated by the algorithm (see table 1). The trees for LCPS are on average three times larger than for the other representations and their interpretation is quite intricate. Therefore, the algorithm does not succeed in generating a set of rules able to classify efficiently the five vowels. This indicates that the boundaries are complex and that the vowels cannot be separated with simple production rules.

Table 1: Comparison of results obtained with the decision trees (mean size of the trees – # – and correct classification scores) and with the rules after pruning the tree (mean number of rules – # – and correct classification scores).

	Decision trees		Rules	
Method	#	Classified	#	Classified
LCPS	121.5	91.28 %	8.8	67.83 %
FORM	39	97.43 %	9	97.66 %
DRM	25.7	98.25 %	6	98.31 %
MAEDA	34.3	97.67 %	7.3	97.67 %

On the contrary, the performances of the three other representations are similar to those obtained by discriminant analysis. Moreover, the rules generated by the algorithm are quite intuitive :

- The rules deduced from FORM make efficient use of the formant frequencies in order to discriminate the vowels.
- The rules deduced from the two articulatory representations DRM and MAEDA are very intuitive and consistent with our knowledge on the production of vowels. They use the main constriction and the lips opening to distinguish among the vowels.

Finally, it is interesting to note the small size of the tree for DRM, able to classify the five vowels with, on average, only six rules.

CONCLUSIONS

We compared 10 representations of the speech signal on a vowel identification task with two different classification procedures: the linear discriminant analysis and the decision trees algorithm. The cepstrum, the formants and the articulatory representations achieve similar performances with linear discriminant analysis. When using the decision tree algorithm, similar performances are only obtained for formant and articulatory representations. Indeed, for the cepstrum, the performances of the rule-based classifier are found to be significantly worse. This can be explained by an overfitting of the training set which results in very complex trees that are unable to abstract the data.

ACKNOWLEDGMENTS

This work was partially supported by the "Communauté Française de Belgique" and the "European Communities" in the framework of the ARC 93/98 — 168, ARC 92/97 — 160, and FALCON (6017) Basic research ESPRIT projects.

REFERENCES

[1] A. Soquet, and M. Saerens, "A comparison of different acoustic and articulatory representations for the determination of place of articulation of plosives," Proc. of ICSLP, pages 1643-1646, 1994.

[2] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," J. Acoust. Soc. Am., vol. 24, pages 175-184, 1952.

[3] B. Lindblom, "On vowel reduction," Speech Transmission Laboratory-Quarterly Progress and Status Report, Stokholm, vol. 29, 1963.

[4] W. Strange, "Dynamic specification of coarticulated vowels spoken in sentence context," J. Acoust. Soc. Am., vol. 85, pages 2135-2153, 1989.

[5] J. D. Miller, "Auditory-perceptual interpretation of the vowel", J. Acoust. Soc. Am., vol. 85, n°5, pages 2114-2134, 1989.

[6] J. Hillenbrand, and R. T. Gayvert, "Vowel classification based on fundamental frequency and formant frequencies," Journal of Speech and Hearing Research, vol. 36, pages 694-700, 1993.

[7] P. Jospa, A. Soquet, and M. Saerens, "Variational formulation of the acoustico-articulatory link and the inverse mapping by means of a neural network," In "Levels in Speech Communication Relations and Interactions," Amsterdam: Elsevier, pages 103-113 ,1994.

[8] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," J. Acoust. Soc. Am., vol. 55, n°6, pages 1304-1312, 1974.

[9] E. Zwicker, and E. Terhardt, "Analytical expresions for critical-band rate and critical bandwith as a function of frequency," J. Acoust. Soc. Am., vol. 68, n° 5, pages 1523-1525, 1980.

[10] G. Fant, "Speech sounds and features," MIT Press, Cambridge MA, 1973.

[11] J. Makhoul, "Linear prediction: a tutorial review," Proc. IEEE, vol. 63, pages 561-580, 1975.

[12] M. Mrayati, R. Carré, and B. Guérin, Distinctive regions and modes: a new theory of speech production," Speech Communication, vol. 7, pages 257-286, 1988.

[13] S. Maeda, "Une modèle articulatoire de la langue avec des composantes linéaires," Actes des 10èmes Journées d'études sur la parole, pages 154-162, 1979.

[14] Q. Lin, and G. Fant, "Vocal-tract area-function parameters from formant frequencies," Eurospeech, pages 673-676, 1989.

[15] "SPSS Reference guide," SPSS Inc., 1990.

[16] J. R. Quinlan, "C4.5: Programs for machine learning," Morgan Kaufmann Publishers, San Mateo, California, 1993.

A PHONETICALLY ORIENTED SPEECH DATABASE FOR MANDARIN CHINESE

Chiu-yu Tseng
Institute of History and Philology & Institute of Information Science, Academia Sinica, Taipei, Taiwan, R.O.C.

ABSTRACT

A phonetically oriented database for Mandarin Chinese speech is designed. Using two electronic corpora of 77,324, lexical entries and 5,353 sentences, 1,455 lexical items and 599 phrases/ sentences in discourse/short stories that cover all possible segmental, syllabic plus tonal combinations in Mandarin Chinese were generated. Tailored software is designed to perform phonetic and acoustic analyses for collected speech samples .

INTRODUCTION

The need to establish a large scale database of Mandarin Chinese speech has been existing ever since research in speech synthesis and speech recognition began in Taiwan over a decade ago. Synthetic speech and automatic speech recognition by computers offer the most optimal and efficient method of communication between humans and computers. [2, 3] While researchers in Taiwan have been actively conducting research in both speech synthesis and speech recognition in Mandarin Chinese without a large scale database, a consensus has been reached that a database that would provide orthographic, phonetic as well as acoustic information would be essential. The paper reports part of an ongoing project toward that goal. The project consists of a knowledge database, a corpus database, a parsing database, a speech database and finally an application end. Resources and specialties from various sectors in Academia Sinica, Taipei, Taiwan has been delegated. This report is the first attempt to describe the speech database only.

Although officially beginning in the fall of 1994, researchers at Academia Sinica have initiated and participated in several previous efforts to collect of speech database in Mandarin Chinese. We realize such a database would be crucial because the acoustic realizations of segments and tones and their interactions depend on complex interactions among many factors. At the present stage, we are prepared to deal with factors that are phonetic and acoustic, emphasizing the tonal aspect of Mandarin Chinese in particular and therefore using syllable as the basic unit . The long term goal is to establish a large scale database that would incorporate intra-speaker and inter-speaker factors. However, the present focus is a phonetically-oriented database that aims to include all possible intra- and inter-syllabic and tonal combinations in most frequently used words so that speech collected under such guidelines would enable us to investigate phonetic properties that would be of use for developing a speech synthesis and recognition system.

THE SPEECH DATABASE

The database consists two types: (1) a word database and (2) a continuos speech database. Both types are now being developed by collecting speech data from different speakers.

DESIGN OF THE DATABASE

Both the word database and the continuous sentence database are designed to be phonetically balanced.

For the word database, an electronic dictionary corpus called Modern Chinese Corpus [1] that include more than 80,000 lexical items were used. Software was designed to first select lexical entries of at most four syllables in structure. A total of 77,324 items were derived. Software was then designed to select items that cover all possible intra- and intra-syllabic plus tonal combinations from three sets of sub-corpus, i.e., the most frequently used 20,000, 40,000, and 77,324 lexical items from the text corpus. Table 1 summarizes the results.

Table 1. Statistical analysis of phonetically specified lexical items from 3 text corpora.

Total # of lexical items		20,000	40,000	77,324
mono-syllabic words	#	3,177	4,369	6,974
	# of possible tones	5	5	5
di-syllabic words	#	14,276	27,431	48,349
	# of possible tonal combinations	20	20	20
tri-syllabic words	#	1,674	4,380	11,562
	# of possible tonal combinations	78	92	97
quadri-syllabic words	#	873	3,820	10,349
	# of possible tonal combinations	235	300	354
# of possible inter-syllable combinations		1,351	1,536	1,649

Results demonstrate that choosing lexical items with the above-mentioned phonotactic and phonetic specifications from a corpus of only 20,000 most frequently used words would suffice. Therefore, the chosen word database consists of 1,455 frequently used words that the include 393 monosyllabic, 676 disyllabic, 145 trisyllabic and 241 quadrisyllabic words of altogether 3,144 syllables. Table 2 summarizes the results.

Table 2. Statistical analysis that shows the distribution of the chosen words that formed the described database.

Total # of lexical items		1,455
monosyllabic words	#	393
	# of possible tones	5
disyllabic words	#	676
	# of possible tonal combinations	20
trisyllabic words	#	145
	# of possible tonal combinations	78
quadrisyllabic words	#	241
	# of possible tonal combinations	235
# of possible inter-syllable combinations		1,351

The continuous speech database, on the other hand, consists of 599 sentences that are constructed from 5,353 sentences that included ten stylistic variations of narratives and/or speech. Duration of sentences/discourse varies from 2 to 180 syllables.

DATA COLLECTION

The initial goal of the database is intended to collect homogeneous speech to set up standard references phonetically and acoustically because large-scale speech data to be collected in later stages will include a variety of inter- and intra-speaker differences due to dialectical pronunciations. We recruited professional Mandarin language teachers whose production of Chinese is of the standard of professional narrators. Sound proof chambers equipped with PC486 and beyerdynamic M69N(C) microphone were used during recording sessions. The words and sentences were read at a normal speaking rate. Each complete set of speech data by each speaker came to 7 hours of recording time. Table 3 summarizes the speakers of our standard references.

Table 3. Summary of speakers whose speech serves as standard reference for the database.

age	gender	# of speakers
65 years and above	male	1
	female	1
35 - 65 years	male	1
	female	1
under 35 years	male	1
	female	1

SEGMENTATION AND LABELING

Segmentation and transcription of the database were done by hand in order to keep the quality of the reference speech data as high as possible so that it could serve as our basis for designing the software that would perform the initial segmentation and labeling when large-scale speech data is collected. Four windows displayed (1) waveform of the utterance; (2) spectrogram; (3) peak of auto correlation function, root mean square, probability of voicing, and fundamental frequency patterns; and (4) phonetic labeling respectively on one screen. are also displayed. A trained personnel inspects the display of the top three windows while segmenting the speech signal phoneme-by-phoneme. Note that at the current stage, only phonetic transcription is provided. Since it is a difficult task to define boundaries between phonemes, especially between two adjacent vowels, boundaries were defined as the center of the formant transitions between the two phonemes [4] while listening through headset at the same time. Figure 1 shows an example of segmentation and labeling.

When establishing the electronic files, tagging system was designed following specifications from the Linguistic Data Consortium (LDC) with additional tags for tonal information. Phonetic information is yielded to provide possible in-depth investigation of spoken Mandarin Chinese in general. Statistical analyses are were also performed to further yield results of phonetic phenomena that not available otherwise. Table 4a and 4b illustrates the kind of statistical results from analyzed speech data of one speaker.

Table 4a. Statistical analysis of Mandarin Chinese consonants from the described database produced by one speaker.

phone	mean (ms)	std (ms)	phone	mean (ms)	std (ms)
b	15	11	j	75	22
p	79	23	q	164	33
m	90	25	x	172	41
f	106	28	zh	73	98
d	15	13	ch	128	47
t	85	23	sh	202	107
n	71	24	r	90	65
l	70	25	z	129	90
g	24	9	c	188	113
k	101	20	s	204	97
h	111	33			

Table 4b. Statistical analysis of Mandarin Chinese vowels from the described database produced by one speaker.

phone	mean (ms)	std (ms)	phone	mean (ms)	std (ms)
i	236	97	iao	298	107
u	211	108	iou	336	104
yu	308	122	ian	304	116
a	249	87	in	291	92
o	247	80	iang	317	117
e	237	118	ing	283	88
ai	279	94	ua	311	115
ei	256	106	uo	275	127
ao	294	111	uai	296	84
ou	263	121	uei	265	103
an	268	103	uan	337	112
en	253	92	uen	289	98
ang	302	110	uang	333	117
eng	276	98	ong	283	100
er	278	99	yue	316	83
ia	304	104	yuan	341	97
ie	291	98	yun	296	92
iai	577	1	yung	313	111

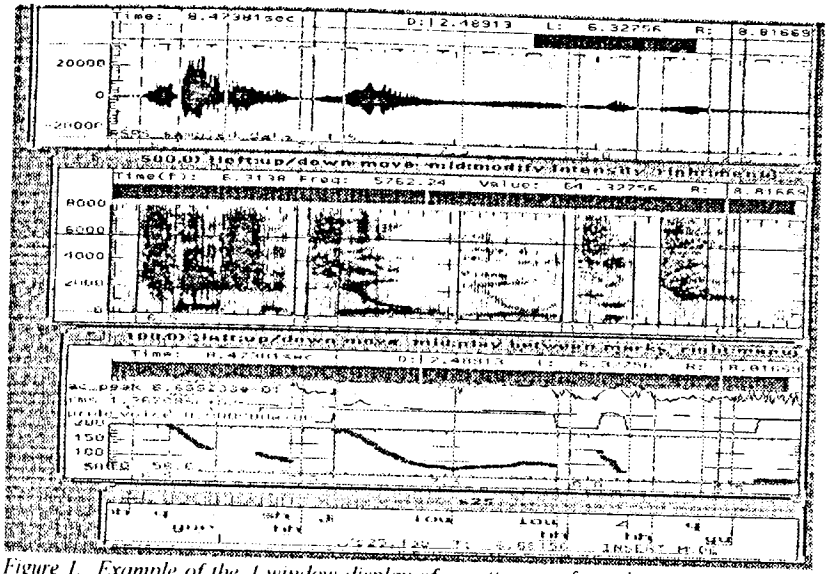


Figure 1. Example of the 4-window display of an utterance from the screen. From top to bottom, window 1 displays the waveform; window 2 the spectrogram; window 3 auto correlation functions, probability of voicing, and the fundamental frequency patterns; window 4 the phonetic labeling.

CONCLUDING REMARKS

An outline of a Mandarin Chinese speech database is described. At the current stage, it consists of two types of databases. Speech data are transcribed with fine acoustic-phonetic labels to meet a variety of needs for speech research. So far, data from six speakers, three males and three females have been completely digitized. The project is at its first year of a 5-year endeavor. Next year efforts will be devoted to collected speech data using statistical methods so that a large number of speakers, each providing a fraction of the above designed set, will participate.

REFERENCES

[1] Chen, K-J, and Huang, C-R, Modern Chinese Corpus (ongoing project at Institution of Information Science, Academia Sinica, Taipei, Taiwan, R.O.C.)

[2] Zue, V., Seneff, S, and Glass, J., "Speech database development: Time and beyond" Paper presented at the Workshop on Speech input/output Assessment and Speech Databases, Amsterdam, The Netherlands, 20-23 September, 1989.

[3] Lamel, L., Kassel, R.H. and Seneff, S. "Speech database development: Design and analysis of the acoustic-phonetic corpus" Proceedings of the Speech Recognition Workshop, Palo Alto, Ca., February 19-20, 1986.

[4] Kuwabara, H., Takeda, K, Sagisaka, Y, Katagiri, S., Morikawa, S, and Watanabe, T. "Construction of a large-scale Japanese speech database and its management system" S10b.12,1989 IEEE.

JITTER-MEASUREMENTS FROM TELEPHONE-TRANSMITTED SPEECH

Isolde Wagner
Bundeskriminalamt - FB Sprechererkennung, Wiesbaden, Germany

ABSTRACT

The present study investigates the validity of a new jitter-algorithm on telephone-transmitted speech samples which are particularly degraded by band-width limitation. The algorithm has been developed for specific use in forensic speaker identification and allows for the quantification of hoarseness not only from isolated sustained vowels but also from vowels in connected speech. The results of a pilot study are presented here. They show that the algorithm is valid to differentiate speakers with certain kinds of pathological hoarseness from speakers with normal voices.

INTRODUCTION

In order to quantify voice qualities which are perceived as hoarse, special attention has been paid to a phenomenon which refers to the temporal irregularities of the vibration process of the vocal folds. The phenomenon is called jitter. It is defined as the involuntary short-term variation of the voice fundamental frequency (fo) from one cycle to the next, in contrast to the voluntary and controlled long-term variation of fo which is the physical correlate of sentence intonation.

While several studies have proposed methods to allow for jitter measurements in high quality recordings of isolated sustained vowels [1,2,3,4,5,6 for example], there seems to be no reliable method of measuring jitter in connected or degraded speech. These problems, however, arise in forensic speaker identification (SI), where (a) non-cooperative speakers have to be examined who are not inclined to produce sustained vowels, and (b) the majority of the speech samples to be analysed are telephone-transmitted with a pass-band between about 300 to 3400 Hz. Therefore a new jitter-algorithm was developed by the Forensic Science Laboratory of the Bundeskriminalamt (Federal Criminal Police Office) and the University of Trier which was designed to yield reliable results even under forensic conditions.

JITTER-ALGORITHM

The new algorithm differs from previous ones by allowing jitter measurements either from sustained vowels or from vowels in connected speech, irrespective of the underlying sentence intonation. It has been implemented on a MEDAV SPEKTRO 3000 computer system and consists of two analytical procedures: (a) a new fo analysis method which is based on frequency demodulation procedures providing high resolution fo values, and (b) a new method for the computation of jitter taking up the basic idea of relative average perturbation (RAP) as suggested by Koike [4].

The two procedures have been explained in detail in an earlier study [7], however, the specificity of the new algorithm is described in more detail here. It consists in treating the high resolution fo contour as a multidimensional vector. A second, auxiliary vector is derived from the first using a method of approximation with a third order polynomial function - a contour with one point of inflexion, serving as a reference vector of the fo contour. In Figure 1, the second window from the bottom gives an example how the procedure works with jitter in a portion of 120 ms duration of the sustained vowel /a:/ produced by a hoarse male speaker at an average fo of 88 Hz. The steps represent the high resolution fo contour including short-term variations; the smooth curve represents the long-term variations derived by a third order polynomial function which describes the intonation contour. The deviation of the actual values from the polynomial function is calculated and the result is shown in the lower right corner in terms of RAP. The value is 2.8655 %.

EXPERIMENT

In order to test the validity of the algorithm based on speech samples which are degraded by band-pass filtering and thus do not contain the fundamental in the signal, the study uses a harmonic rather than the fo as a multidimensional vector in the procedure. The results of the jitter measurements obtained in this way are compared to the results of measurements on the basis of high quality recordings.

SUBJECTS AND MATERIAL

The material consists of recordings of seven male German speakers with normal voices and seven speakers with pathological hoarseness of various origins deviding the hoarse speakers into two subgroups: speakers who suffer (a) from *hypo-* and (b) from *hyper*functional dysphonia. The recordings were made under sound treated conditions using high quality equipment. Subjects were required to produce different types of sustained vowels, both in isolation and in a /mVm/-context, and also various sequences of connected speech.

ANALYSIS PROCEDURE

Recordings of both of the two sustained productions of the vowels /e/, /ɛ/, /a/, and /o/, and one sample of the same vowel from connected speech were digitized in the MEDAV SPEKTRO 3000 computer system in a two channel mode, where subsequently one of the two channels was band-pass filtered from 300 to 3400 Hz, thus simulating the degrading characteristic to telephone-transmission. Jitter-measurements were made using the fo from the channel containing the high quality recording, and the *third* harmonic (h3) from the filtered channel, because even in low male voices, it can be safely assumed to be within the range of telephone-transmission.

RESULTS

Because it was observed that jitter values vary with the duration of the measured portion of the vowel and because of the fact that in connected speech vowel durations of more than 120 ms are rare, portions of 120 ms were used for all measurements. Furthermore, it was found that the four different vowels /e/, /ɛ/, /a/, and /o/ did not yield any systematical differences in jitter values. Therefore, these vowels were pooled in the study.

For the purpose of comparing between high quality and degraded speech samples distributions of high and low jitter values were investigated on measure-

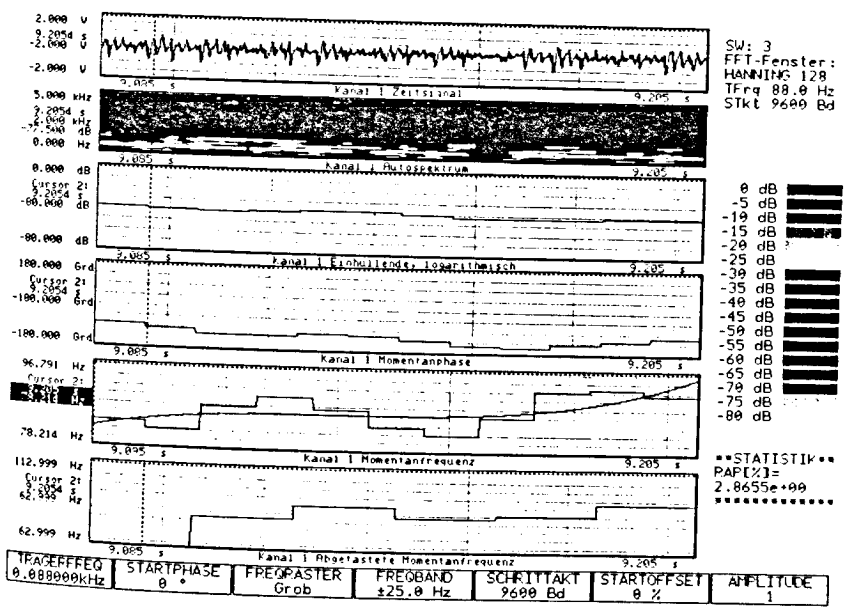


Figure 1. Procedure of the new jitter-algorithm working in the sustained production of the vowel /a:/

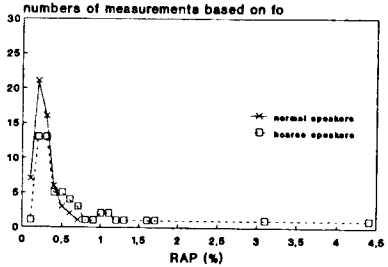


Figure 2a. Distribution of jitter values: measurements based on fo of sustained vowels

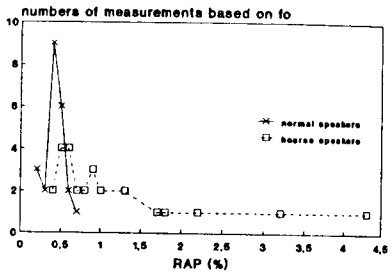


Figure 2b. Distribution of jitter values: measurements based on fo of vowels from connected speech

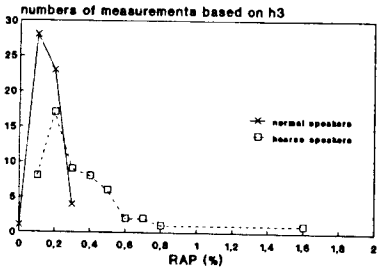


Figure 3a. Distribution of jitter values: measurements based on h3 of sustained vowels

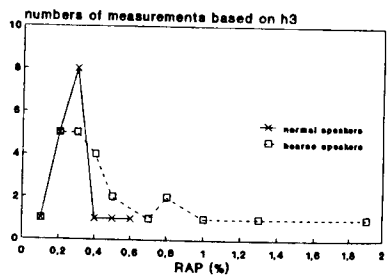


Figure 3b. Distribution of jitter values: measurements based on h3 of vowels from connected speech

ments based both on fo and h3 for the hoarse and normal speakers and for both vowel conditions separately. Mean jitter values, as well as minimum and maximum values were computed for the different types of measurements for the normal speakers and the two groups of hoarse speakers.

Jitter-measurements based on fo

Figure 2a shows the distribution of jitter values for measurements based on fo for the two *sustained vowel* productions. The full line represents the values for the normal speakers, the broken line the values for the hoarse speakers.

It was established that most of the values for the normal speakers range between 0.2 to 0.3% RAP. Values of more than 0.5% are rare. The means are 0.3%, the minimum values 0.1% and the maximum values 0.7% RAP for both of the two sustained productions. The values for the hoarse speakers, however, show a very large range of between 0.1 and 4.4% RAP. But, whereas the values for the hypofunctional group are well within the range of the normal speakers, the values for the hyperfunctional group are clearly higher, and values of less than 0.5% RAP are very rare. The mean values are 1.0 and 1.1% RAP for the two sustained productions, however, the variation within this group is large.

Figure 2b shows the distribution of jitter values for measurements based on fo of the vowels from *connected speech*. It emerged that jitter values were higher than in the sustained vowel productions for all groups of speakers, and both of the two groups of hoarse speakers differ from the group of the normal speakers. Whereas the range of the normal speakers' jitter values amounts to 0.2 to 0.7% RAP with a mean of 0.4%, the hoarse speakers exhibit jitter values ranging from 0.4 to 4.3% RAP with a mean of 0.7% for the hypofunctional and a mean of 1.6% for the hyperfunctional group.

Jitter-measurements based on h3

Figure 3a shows the distribution of jitter values for measurements based on h3 of the *sustained vowel* productions. It is evident that the jitter values of all groups of speakers are lower than the comparable values obtained from measurements on the basis of fo. Nearly all of the jitter values (51 out of 56) of the normal speakers are distributed between 0.1 and 0.2% RAP for the two sustained productions. The minimum value is 0.0%, the maximum value is 0.3% RAP. The hoarse speakers, however, exhibit larger variation in the jitter values ranging from 0.1 to 0.5% RAP for the hypofunctional and from 0.1 to 1.6% RAP for the hyperfunctional group. The means are 0.2 and 0.3%, respectively, for the hypo- and 0.4 and 0.5% RAP for the hyperfunctional group for the two sustained productions.

Figure 3b shows the distribution of jitter values for measurements based on h3 of the vowels from *connected speech*. As for the sustained vowels it was observed that all jitter values obtained from measurements based on h3 are lower than those based on fo. However, the values of vowels taken from connected speech are higher than those based on sustained vowels. The normal speakers show values from 0.1 to 0.6% RAP with a mean of 0.3%, and the hypofunctional group yields values which are well within the range of the normal speakers. The values of the hyperfunctional group, however, range from 0.2 to 1.9% RAP with a mean of 0.8%.

DISCUSSION

The jitter-algorithm was found to differentiate speakers with certain kinds of pathological hoarseness from normal speakers through measurements based on h3 as well as on fo and on both types of vowel production. Whereas speakers who suffer from hypofunctional dysphonia show jitter values which are more or less within the range of those of normal speakers, speakers who suffer from hyperfunctional dyphonia exhibit values which are much higher and more widely distributed. The boundary line between these hoarse speakers and the normal speakers is about 0.5% for the measurements based on fo and about 0.3% RAP for the measurements based on h3. Jitter values based on vowels in connected speech were found to be slightly higher than those based on sustained vowels. This can possibly be explained by the influence of coarticulation on the vowels in connected speech. All these findings correspond with the results of a previous study by the present author [7], where jitter-measurements were made on the basis of fo from the vowel /a/ only. The observation that jitter values obtained from measurements based on fo are systematically higher than those based on h3 is considered to be related to the working principle of the algorithm.

It can be concluded that the algorithm is in principle able to measure jitter in connected speech which is degraded by the particularly band-pass filtering of telephone-transmission. However, before the procedure can be established in forensic SI, more research is needed.

REFERENCES

[1] Baken, R.J. (1990), 'Irregularity of vocal period and amplitude: A first approach to the fractal analysis of voice', *Journal of Voice*, 4 (3), 185-197.

[2] Hollien, H.; Michel, J. & Doherty, E.T. (1973), 'A method for analysing vocal jitter in sustained phonation', *Journal of Phonetics*, 1, 85-91.

[3] Horii, Y. (1979), 'Fundamental frequency perturbation observed in sustained phonation', *Journal of Speech and Hearing Research*, 22, 5-19.

[4] Koike, Y. (1973), 'Application of some acoustic measures for the evaluation of laryngeal dysfunction', *Studia Phonologica*, 7, 17-23.

[5] Lieberman; P. (1963), 'Some acoustic measurements of the fundamental periodicity of normal and pathologic larynges', *The Journal of the Acoustical Society of America*, 35 (3), 344-353.

[6] Ludlow, C. Coulter, D. & Gentges, F. (1983), 'The differential sensivity of measures of fundamental frequency perturbation to laryngeal neoplasms and neuropathologies', in D. M. Bless and J. H. Abbs (eds), *Vocal fold physiology: Contemporary research and clinical issues*, San Diego: Colledge Hill, 381-392.

[7] Wagner, I. (1995), 'A new jitter algorithm to quantify hoarseness: an exploratory study', to be published in: *Forensic Linguistics*, 2 (1), 000-000.

EXPERIMENTS WITH AN OBJECT-ORIENTED SPEECH DATABASE

M. Vainio[1], T. Altosaar[2], M. Karjalainen[2] and Antti Iivonen[1]
[1]University of Helsinki, Dept. Phonetics
P.O. Box 35 (Vironkatu 1 B), 00014 Helsinki, Finland
[2]Helsinki University of Technology, Acoustics Lab.
Otakaari 5 A, 02150 Espoo, Finland

ABSTRACT

This paper describes a very flexible environment in which to perform speech related experiments along with some examples. A complete range of operations for the speech researcher is available from low-level signal processing to high-level phonetic analyses and semi-automatic transcription using neural networks. Entities such as signals and transcriptions are all represented as objects and can be stored between sessions via a transparent database system.

INTRODUCTION

Speech databases have been developed for the major European languages and have been available for use for several years already [1]. However, the lack of a comprehensive and transcribed speech database for Finnish has hindered research in speech analysis and recognition for this language. In 1991 we initiated a project in which Finnish speech material was collected into a database and transcribed manually by phoneticians [2].

Since we desired a very flexible environment in which to perform speech experiments we decided to implement the database on top of our object-oriented QuickSig signal processing system [3]. The entire system is written in Common Lisp and CLOS and provides for a seamless integration of all activities. The system includes input and output audio channels, graphical tools for the user to move around in the database, and transcription frames with semi-automatic transcription aids, such as diphone detectors implemented with neural networks. Loading, caching, and storing of signals is managed by the system automatically and transparent to the user. An advanced speech representation framework is used to represent phonetic and linguistic information and can be used in speech processing tasks such as analysis, synthesis and recognition. The framework allows for abstract, structural, specific, and fuzzy phonetic objects to exist over different scales, e.g. from sentences down to acoustic segments. Transcriptions are automatically transformed into these linked speech representation objects when accessed from persistent store.

Predicate functions can be designed and applied to search over all or part of the database. For example, a search over part of the database that includes 20,000 transcribed phonemes can be performed in a few seconds. The search returns the phonetic objects that matched the predicate, e.g., a set of phonemes. These phoneme objects are all linked to their original signals and thus can be used in a wide variety of signal processing methods and techniques found in QuickSig such as spectral averages calculated over specified regions of the speech signal, formant analysis, and duration analysis.

LABELING PROCESS

Labeling of speech signals is accomplished with the aid of transcription frames. The transcription frame serves as a graphical user interface between the technical aspects of the transcription process and the transcriber. Any number of different transcriptions may be linked to a signal since different interpretations may be required in some situations.

Segmentation

To ensure reliable and consistent segmentation it is important that the user can utilize decision-aiding transcription tools. Conventional tools include time-waveform, FFT-spectrogram, and energy contour displays.

In addition to these standard tools we have implemented more advanced methods that rely on auditory modeling. An auditory spectrogram on the Bark scale and a loudness contour usually permit more accurate segmentations to be made. A spectral change measure calculated from auditory spectra covering the span

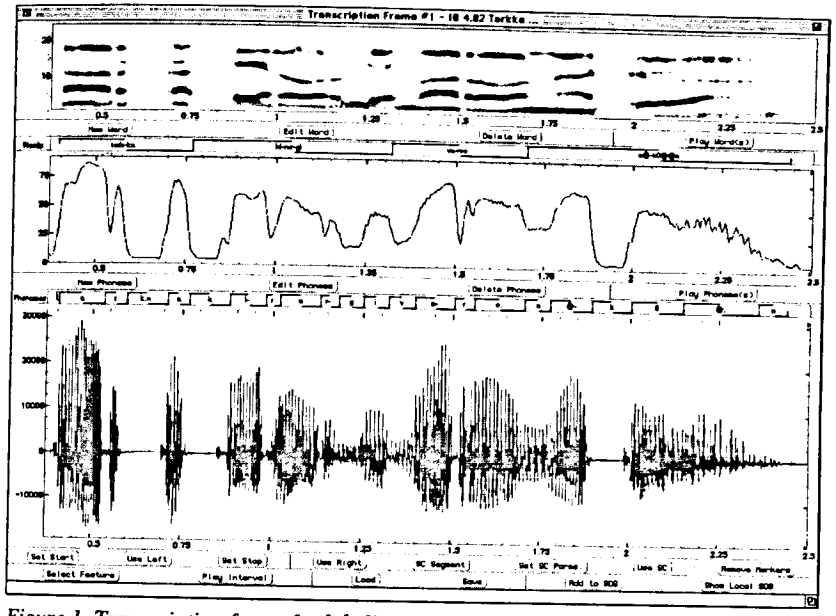


Figure 1. Transcription frame for labeling speech signals. An auditory spectrogram, a loudness contour, and the time waveform are shown along with a word and phonetic level transcriptions.

of the signal often indicate fairly accurate locations for segment boundaries.

Neural networks can also be used to determine segment boundaries. Diphone detecting networks can offer reliable semi-automatic hypotheses for phoneme boundaries.

Labeling

Signals are usually segmented according to their structure on a phonemic level. In addition to this phonemic representations it is possible to label material in several hierarchical levels ranging from low-level acoustical segments to syllables, words, and sentences.

The user can define the symbols used for labeling operations according to specific needs. For instance, a more narrow transcription may be required in some cases. This may be accomplished by adding a new representational level to the existing transcription, or, by creating an alternative transcription altogether.

Reading of other transcribed material from databases is also possible and can be viewed in the transcription frame.

Figure 1 shows a transcription frame for the Finnish sentence <tarkka kirurgi varoo näköään>. Besides the time-waveform, loudness contour, and auditory spectrogram, two different levels of transcription are visible: a phoneme and a word level. Frequently used operations have been assigned their own buttons and allow the user to perform different functions such as playing portions of the signal, assigning boundaries, and invoking different transcription aids.

DATABASE

The database system has also been implemented in an object-oriented programming fashion and is designed for simple and user transparent operation. Besides containing signals, the database can store transcriptions, speaker related information, and in general any user designed object. Links between different objects allow for deferred loading, i.e., an object is loaded into memory only when required for a calculation and discarded when not needed. This means that an entire speech database can exist in working memory simultaneously. Fast

and efficient analyses can thus be performed on large amounts of material.

Objects in the database are arranged in a hierarchical manner. When a user terminates a session the database system automatically checks for objects that have been created or changed and are transferred to permanent store.

Figure 2 shows a graphical representation of part of a Finnish speech database. Nodes in the tree are mouse-sensitive and allow for different operations to be performed, e.g., opening a transcription frame, playing signals, and inspecting the state of specific objects.

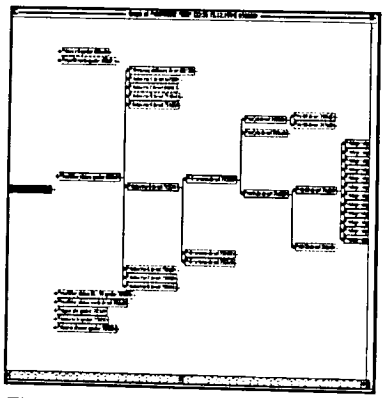


Figure 2. A graphical representation of part of a speech database. By default nodes are collapsed but may be opened with the mouse.

The database currently contains approximately two hours of labeled speech from two male and two female native Finnish speakers. At this stage most of the database consists of isolated words but sentence material is being added.

PHONETIC HIERARCHY

All of the phoneme symbols created during labeling are transformed into instances that have a implicit feature structure. Figure 3 shows part of this network seen from a phonological viewpoint. The use of object-oriented CLOS class hierarchies define the specific structural relations between different phonetic units. These relations can be used to represent phonetic and linguistic information when performing searches and analyses over the database.

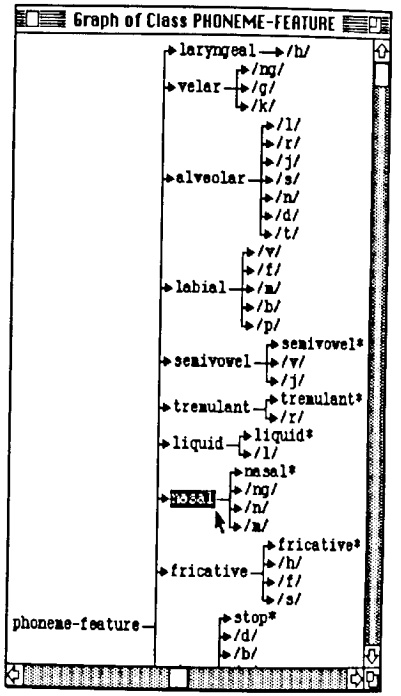


Figure 3. Part of the phoneme inheritance network graphed according to phonological features.

ANALYSES

In this section we briefly present some of the methods that are available for advanced phonetic analyses.

Duration Analysis

In this example the duration distribution of the vowel /i/ in a C/i/C context is to be calculated for a single speaker. First, a predicate is defined using Lisp syntax:

```
(define-predicate C/i/C
 (and
  (previous-phoneme is-not-a V)
  (x is-a /i/)
  (next-phoneme is-not-a V)
  (speaker-is "MV")))
```

Then a search over the entire database is performed and a set of phonemes matching the predicate is returned. Each phoneme object has its own internal state and can determine its duration in time. A histogram can be built from these phonemes and is shown in figure 4.

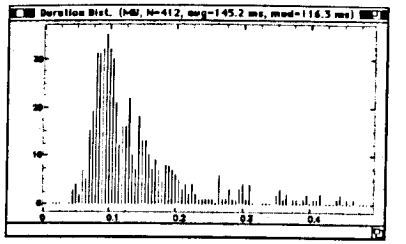


Figure 4. Histogram of the vowel /i/ in a consonant context (x-axis represents time in seconds).

Spectral Average

Each phoneme not only knows it duration but also its absolute position in time. This information may then be used by any analysis method. One such interesting analysis is to calculate the auditory spectrum at the midpoint of each phoneme's signal span. The average spectrum and the spectral distribution for all 412 /i/ vowels found in the previous example are shown in figure 5.

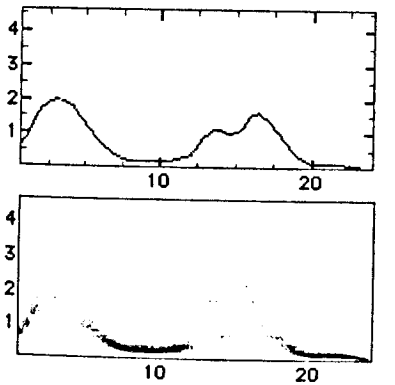


Figure 5. Auditory spectral average (top) and distribution (bottom) for 412 /i/ vowels (amplitude vs. frequency on the Bark scale).

Formant Analysis

Each individual phoneme's auditory spectrum can also be automatically analyzed for formant locations which can then be displayed on a F1/F2 chart. In figure 6 the same 412 /i/ vowels have been analyzed and displayed. Individual vowels have their formants represented by circles which are mouse sensitive. This allows the user to inspect each item separately and access information such as exact formant locations, the transcription and word in which the phoneme is situated, and other related information such as speaker identity and recording information. In this figure the actual spectrum for the vowel pointed to by the mouse is also shown at the bottom of the figure. This allows the user to interact with the analyzed data.

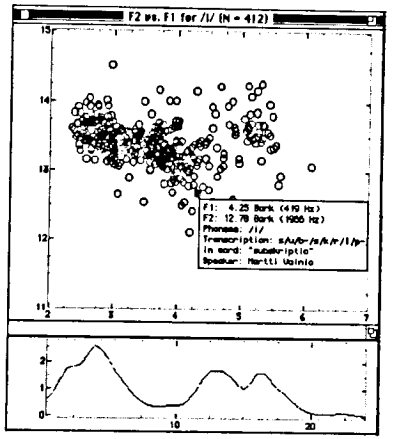


Figure 6. Interactive F1/F2 chart, information related to a specific /i/ vowel (pointed to by the mouse), and its corresponding auditory spectrum.

SUMMARY

This paper presented a powerful and flexible object-oriented speech database system. The tools used to transcribe signals, the database system, as well as the phonetic hierarchy were described. Finally, some speech analysis methods were presented. Since the system is built on top of the extendible Lisp, CLOS, and QuickSig DSP substrate, users are free to add new analysis methods according to their needs.

REFERENCES

[1] Esprit Project 2589 (SAM) Final Report.

[2] Karjalainen, M. and Altosaar, T. (1993) An Object-Oriented Database for Speech Processing. EuroSpeech-93, Berlin.

[3] Karjalainen, M. (1990) "DSP Software Integration by Object-Oriented Programming: A Case Study of QuickSig." IEEE ASSP Magazine.

AUTOMATIZED FORMING OF INDIVIDUAL SPEECH FILE FOR AUTOMATIC SPEECH RECOGNITION AND SYNTHESIS SYSTEM

Taras K. VINTSIUK
NAS Institute of Cybernetics, Kyiv, Ukraina

ABSTRACT

The automatized Individual Speech File (ISF) forming process is proposed. ISF is intended for introducing into computer to start automatic speech recognition and/or synthesis for a given person. For this the training procedures are used. They automatically calculate phoneme-threephone prototypes, temporal, prosodic, energetic and other parameters and characteristics of speaker voice which constitute so-called Individual Speaker File.

GENERAL PRESENTATIONS

ISF forming procedures appeal to Individual Speech Signal Archives (ISSA) and process all its signals.

ISSA is a set of original speech signals (e.g. speech realizations) expressed in amplitude-time domain. ISSA is considered as fulfilled then each speech realization is accompanied by such descriptions: 1) orthographical text; 2) phoneme (phonetical) transcription; 3) phoneme signal bounds (marks); 4) current pitch periods and bounds, and others.

ISSA forming is doing in laboratory conditions and controlled by an expert (e.g. researchers-phoneticians). The expert can correct the phoneme segment bounds using the mouse and by the way of hearing speech segments and regarding original and/or description speech signals (current autocorrelation, spectrum or cepstrum for example) through the computer audio-video-monitor. The marks accepted by the expert are then transferred automatically from one supported realization onto others ones. The expert can correct the realization transcription too. Each fixed speech segment (with fixed bounds) is considered as a realization of the phoneme-threephone, that is a base phoneme accompanied by both preceeded and followed phonemes accordingly with the phonetic context. The phoneme-threephone history that is word, sentence, realization number is conserved too. Further, fixed phoneme-threephone segments are used then as prototypes for automatic segmentation of speech realizations which correspond to other words and sentences. So, there are many procedures which copy expert actions and automatize the labelization process and ISSA forming.

ISSA and ISF for one supporting speaker are then used in ISSA fulfilling for a new speaker. The researcher-phonetician invasion into the process of forming of both Individual Speech Signal Archives and Individual Speech File expands the speech knowledge and improves the accuracy of individual automatic speech recognition as well as the quality of individual automatic speech synthesis.

ILLUSTRATIONS

As an example a speech signal realization of the Ukrainian word ОДИН (ONE in English) is presented in the Figure 1. The top graph is an amplitude-time speech signal. Then current autocorrelation, cepstr and spectr are shown in three lower graphs respectively. For these preprocessing presentations as well as for segment bounds the uniform discrete time with the step 15 *ms* was used. The phoneme-threephone bounds were arranged by the expert. Phoneme

segments are accompanied by phoneme symbols # (pause), О (О non-stressed), _Д_ (voiced stop phase of D), Д (plosive phase of D), И (I stressed), Н (N sonorant).

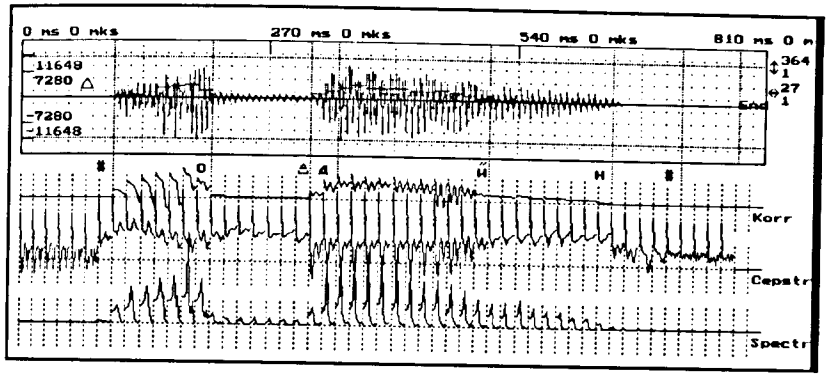


Figure 1. Monitor presentation of speech signal realization of the Ukrainian word ОДИН (ONE in English).

The result of automatic transferring of phoneme-threephone bounds accepted by the expert (see Figure 1) onto a new realization of the same word is shown in Figure 2.

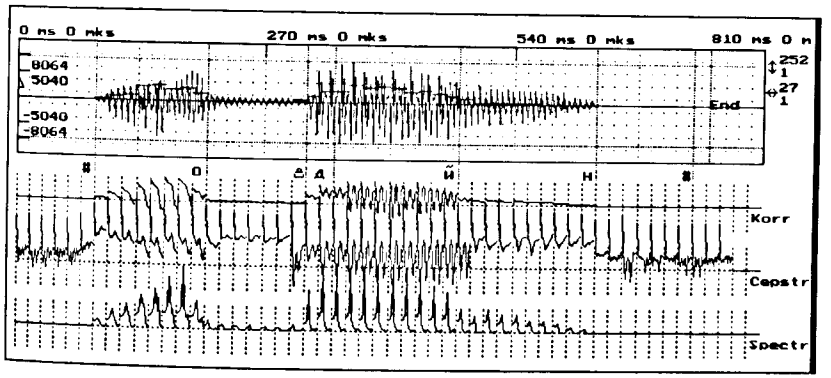


Figure 2. The result of automatic transferring of the phone bounds from the supported realization onto other one for the same word.

Figures 3, 4 and 5 present in detail the fragments of the speech signal realization shown in the Figure 1.

STATE OF THE ART

Now the Individual Speech Signal Archives and the Individual Speech File fulfiling technologies are being near completion. Mainly ISSA and ISF are made for Slavic languages especially Ukrainian and Russian.

The result of the study is used in the designing of Multilingual Speech Dialogue Systems [1], [2], Oral Dictation Machine, Oral Translation Machine.

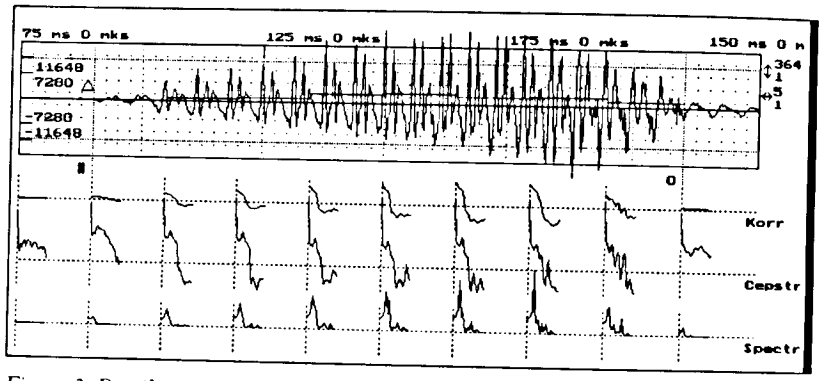


Figure 3. Detail monitor presentation of the supported speech signal realization of the Ukrainian word ОДИН (first fragment).

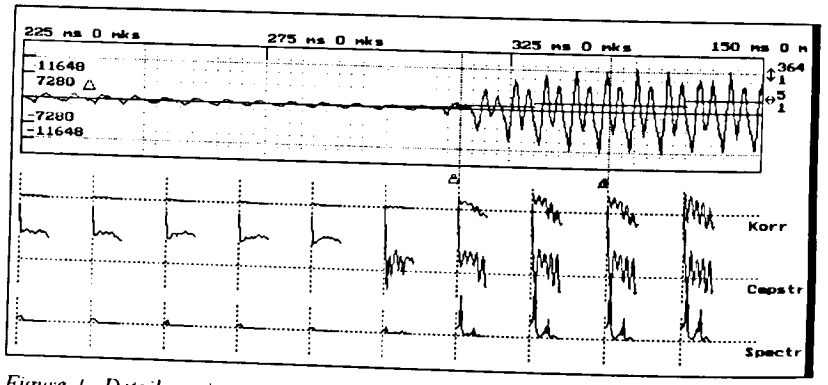


Figure 4. Detail monitor presentation of the supported speech signal realization of the Ukrainian word ОДИН (second fragment).

ACKNOWLEDGEMENT

This study have been inspired by the Research Programs of the Ukrainian State Committee on Science and Technology, and by the cooperation within ELSNet, and we want to express our gratitude to E. Klein and F. Neel and other people responsible for them.

REFERENCES:

[1] Final Report on the UNESCO Contract SC/RP 261060.8 "Development of Multilingual (including English and Russian Languages) Speech Dialogue System for a Microcomputer", Institute of Cybernetics of the AS of Ukraine, Kyiv, 1986, 97p.

[2] Final Report on the UNESCO Contract SC/RP 261377.9 "Advance of Multilingual Speech Dialogue System for a Microcomputer", Institute of Cybernetics of the AS of Ukraine, Kyiv, 1989, 33p.

PHONETIC CATEGORIES AND PHONOLOGICAL PROCESSES: VOWEL- GLIDE- CONSONANT ALTERNATION IN SPANISH

Lourdes Aguilar
Departament de Filologia Espanyola, Facultat de Lletres,
Universitat Autònoma de Barcelona, Bellaterra 08193, Barcelona, Spain

ABSTRACT

Problems around the interpretation of glides arise in a recurrent way in Spanish phonological descriptions. Nevertheless, phonetic properties are usually neglected: the term "semiconsonant" is based on the syllabicity but the property of syllabicity is not phonetically defined in a precise way. In this study, the aim is to characterise from a phonetic point of view the semiconsonantal elements versus the vocalic and consonantal elements related in the Spanish sound inventory and to observe their manifestation depending on communicative factors. Two groups of data have been used: speech samples excerpted from conversations between two speakers participating in the consecution of the map task and the reading of the same sequences in a neutral way.

The findings suggests that phonetic data obtained from the study of spontaneous speech provide additional information that should be included in the phonological descriptions.

1. INTRODUCTION

Spanish phonetic descriptions allude to different allophonic manifestations of /i, u/ and to some consonantal segments associated to them. With respect to /i, u/, syllabic and nonsyllabic vowels are described, and related to this distinction, differences between vocalic groups in hiatus -vowel+vowel sequences-, diphthongs -glide+vowel and vowel+glide sequences- are stablished. With respect to the consonant, approximant, fricative or affricate realisations of /j/ are found [1].

Nevertheless, phonetic properties are usually denied in the descriptions: the criteria determining the difference between glide and vowel are based in the syllabicity but the property of syllabicity is not phonetically defined in a precise way. The purpose of the present study is to provide additional data for Spanish showing which are the acoustic cues that distinguish between vowel, glide and consonant, and which modifications due to a change in speaking style are found.

2. PROCEDURE

A corpus of bisyllabic words with the combination of interest appearing in the first syllable has been constructed. Two variables have been taken account: the phonetic category -vowel, glide, consonant (only the approximant realisations have been taken into account)-, and the vowel environment. The sequences considered are the following: the syllables ['je, 'ja, 'jo, 'ju, 'wi, 'we, 'wa, 'ao], the hiatuses ['ie, 'ia, 'io, 'iu, 'ui, 'ue, 'ua, 'uo] and the diphthongs ['i̯e, 'i̯a, 'i̯o, 'i̯u, 'u̯i, 'u̯e, 'u̯a, 'u̯o]. All the sequences are stressed. In addition, the stressed vowels [i, u] in a consonantal environment have been included.

The occurrences of the items have been observed in two communicative situations: dialogues and reading.

In order to obtain a group of dialogues large and natural enough, the model of the HCRC corpus has been adopted [3]: speakers are required to collaborate to reproduce in the map of one of the participants a route which is printed in the other participant's map. In the maps, the toponyms correspond to the words of the corpus.

To carry out the reading task, the items of the corpus were inserted into carrier sentences. There are different types of carrier sentences so as to avoid the list effect, and the sentences are presented in separate sheets to regulate the elocution rate of the speaker.

Sixteen male speakers aged between 20 and 30, with medium and high-level studies, participated in the experiment. For each speaker, samples of both speech situations were obtained.

The recording sessions took place in a sound-treated room at the Phonetics Laboratory of the UAB using a Tascam 112 cassette recorder and a Sennheiser MKH20 microphone.

The items of the corpus were analysed by means of the speech analysis software Waves+. Waveform displays and broad-band spectrograms were plotted for each sequence, and the following measurements were taken: duration, F1 and F2 frequency.

Mean and standard deviation for duration and frequency values have been calculated, and significant differences between the conditions have been assessed by means of ANCVA tests.

3. RESULTS

From the obtained data, two questions can be highlighted: on one hand, it is possible to discriminate from an acoustical point of view between vowels, glides and consonants; on the other hand, phonetic reduction phenomena affecting vocalic groups have been found.

3.1. Vowel vs. vowel in hiatus vs. glide vs. consonant

Table II presents the mean values of data duration of a vowel in consonantal environment, a vowel in hiatus, a glide and a consonant in the two types of corpus. Data corresponding to palatal and velar quality have been pooled.

Table II. Number of cases (n), mean values (n) and standard deviation (sd) of data duration of a vowel (V), a vowel in a hiatus (H), a semiconsonant (SC) and a consonant (C) in the corpus of reading and in the dialogues.

	Reading		Dialogues	
	n	x (sd)	n	x (sd)
V	124	78 (16)	103	65 (16)
H	496	104 (27)	306	86 (23)
G	491	76 (22)	374	59 (16)
C	194	72 (23)	212	66 (14)

The glide, the consonant and the vowel in a consonantal environment present similar durations; only the vowel in a hiatus presents a longer duration. A one-way ANOVA analysis shows differences at a level of signification of 5%, in reading and in dialogues.

With respect to the consonant, a difference related to the style have been noted: in the corpus of reading, the consonant is longer than the glide whereas in the corpus of dialogues, it is shorter.

As far as frequential domain is concerned, any differences between the compared categories have been found in the F1 frequency values. On the contrary, the F2 frequency can be considered as an acoustic cue distinguishing the phonetic categories (see Table III).

Table III. Number of cases (n), mean values (x) and standard deviation (sd) of the F2 frequency data of the vowel [i, u], the vowel [i, u] in a hiatus, the semiconsonant [j, u̯] and the consonant [j, w] in the corpus of reading and in dialogues.

	Reading		Dialogues	
	n	x (sd)	n	x (sd)
[i]	62	2216 (108)	54	2104 (129)
[i]V	248	2155 (103)	152	2117 (148)
[i̯]	248	2116 (96)	171	2019 (163)
[j]	71	2124 (130)	137	2001 (190)
[u]	62	808 (92)	49	891 (129)
[u]V	124	875 (205)	154	892 (188)
[u̯]	124	903 (196)	206	872 (191)
[w]	123	739 (93)	116	737 (117)

In the palatal set, the glide and the consonant show frequency values lower than the corresponding to the vowels in a consonantal environment or in a hiatus, in both types of corpus. A one-way ANOVA analysis shows important differences due to a change in the phonetic category (p:.0001).

In the velar set, a difference due to the speaking style can be described. In the corpus of reading, two groups corresponding to vowel and vowel in hiatus, on one hand, and to glide and consonant in the other hand; on the contrary, in the corpus of dialogues, the

glide shows the highest F2 frequency. However, this difference doesn't arise as significative when a two-way ANOVA analysis (category x style) is applied on the data: there are significative variations related to a change in the category (p:.0001) but not to a change of style (p:.1638).

3.2. Phonetic reduction phenomena in vocalic sequences

The main characteristic of a communicative situation such as the participation in the map task, where subjects engaged in the achievement of a common aim lose the attention paid to their discourse, is the presence of segmental reductions.

Three types of phonetic reduction processes can be observed in the data: a) diphthongisation, where a hiatus is pronounced as a diphthong, b) deletion in a hiatus, and c) vocalisation of a diphthong.

From an acoustic point of view, it is considered that a hiatus have become a diphthong when the duration is reduced and the formant frequencies are displaced with respect to the ideal values of a hiatus; deletion and vocalisation is noted by the presence of a single segment.

3.2.1. Reduction of vocalic groups in hiatus

All the sequences observed in the dialogues show at least one case of reduction to a diphthong, as presented in Table IV.

Deletion is a less frequent process and it can affect the first element in the group, as in ['ie] found as [e], the second element in the group, as in ['io] which becomes [i], or it can result in an intermediate segment, as [o] coming from ['ui].

Globally considered, hiatuses are reduced to diphthongs a 9.5% of occurrences whereas the 2% of cases is pronounced as a single vowel.

Table IV. Number of cases analysed in the corpus of dialogues (n tot), number of sequences realised as diphthongs (n dip), number of vocalisations (n vowel) and vowel derived from the reduction process.

	n tot	Red. dipht	Reduction to a vowel	
		n dip	vowel	n vowel
['iu]	31	2		
['io]	28	3	[i]	3
['ia]	34	1	[i]	1
['ie]	20	2	[e]	1
['ui]	29	5	[o]	1
['ue]	31	3		
['ua]	19	3		
['uo]	17	1		

3.2.2. Reduction of diphthongs

Table V presents the number of deletion cases found in the diphthongs analysed in the corpus of dialogues and the adopted solution for each case.

Table V. Number of diphthongs analysed in the corpus of dialogues (n tot), number of cases appearing as a vowel (n) and observed vowel; when several solutions, number of cases of each solution.

	n tot	n red	vowel	n
['i̯u]	30	3	[u]	2
			[i]	1
['i̯o]	39	7	[i]	5
			[o]	1
			[e]	1
['i̯e]	34	8	[i]	6
			[e]	2
['u̯e]	41	14	[o]	6
			[u]	8
['u̯a]	30	4	[a]	1
			[o]	3
['u̯o]	28	1	[u]	

Diphthongs are reduced to a vowel in 18.31% of occurrences and there is not an important difference between the behaviour of palatal groups (17.47%) and velar ones (19.19%).

Focusing on the vowel derived from the reduction process, it has been observed a strong tendency to preserve the first element in the group: the 56.75 % of diphthongs shows an ellision of the final element in front of the 16.21% where deletion affects the initial element.

Table VI . Number of diphthongs reduced to a vowel in the corpus of dialogues (n red), number of cases of deletion of the first element (% first), number of cases of deletion of the second element (n second), number of cases in which the solution is a new element (% vow).

	n red.	Del. init. element	Del. final element	New element
['i̯u]	3	2	1	
['i̯o]	7	1	5	1
['i̯e]	8	2	6	
['u̯e]	14		8	6
['u̯a]	4	1		3
['u̯o]	1	1		

4. DISCUSSION

The analysis of speech samples obtained from different communicative situations provide data about the acoustic cues of the segments and about the modifications dues to a change in speaking style. The experiment presented here have shown that both in reading and in dialogues, duration and F2 are the primary acoustic cues distinguishing between the phonetic categories of vowel, glide and consonant. These properties shapes the minimal but sufficient contrast for the identification of the categories.

The duration of the vowel in hiatus, different in a large degree from the duration of the rest of segments, suggests the presence of a process of strenghthening of the vocalic quality that preserves the group as hiatus, stopping the diphthongisation, usual in Spanish.

The main acoustic cues are the same in reading and in dialogues, but some questions referred to the processes observed in relaxed speech should be taken into account. The behaviour of phonetic reduction processes -when they are applied, which are the affected elements- can provide additional data in the description of the segments.

Observing the solution adopted in a majority of cases as a result of the vocalisation of a diphthong, a clash between phonological intuitions and phonetic behaviour is found.

If phonetic reduction was related to a hierarchy of strengthness of the elements [4], a higher percentage of deletion of the glide would be expected.

On the other hand, if we focus on syllabic role, the result is similar: given that the vowel occupies the position of the nucleus, the element more susceptible to undergo a phonetic change should be the glide.

On the contrary, in presence of phonetic reestructurations, a strong trend to delete the vocalic element has been found. It could be hypothesized then that the position in the syllabic group exerts a stronger influence than the phonological nature of the element.

These findings suggests that information coming from phonetic analysis of speaking styles provide additional data in the phonetic and phonological description of the linguistic systems and should be integrated in the theoretical accounts [5].

REFERENCES

[1] NAVARRO TOMÁS, T. (1918) *Manual de pronunciación española*, Madrid: CSIC, 1989.
[2] ALARCOS, E. (1965) Fonología española, Madrid: Gredos.
[3] McALLISTER, J- SOTILLO, C.- BARD, E.G.-ANDERSON, A. (1990) *Using the Map Task to investigate variability in speech*, Ocassional Paper, Department of Linguistics, University of Edinburgh.
[4] HOOPER, J. (1976) *An Introduction to Natural Generative Phonology*, New York: Academic Press.
[5] KOHLER, K.J. (1991) "The Phonetics/ Phonology issue in the Study of Articulatory Reduction", *Phonetica*, 48 (2-4): 180-192.

SEGMENT FREQUENCY AND WORD STRUCTURE IN BRAZILIAN PORTUGUESE

Eleonora C. Albano, Agnaldo A. Moreira, Patricia A. Aquino, Adelaide H. P. Silva and Regis K. Kakinohana
Laboratório de Fonética Acústica & Psicolinguística Experimental
Universidade Estadual de Campinas, Campinas, S. P., Brazil

ABSTRACT

Segment frequencies in two representative *corpora* of Brazilian Portuguese are interpreted as evidence for a lexically-based phonetic pattern. The pattern consists in underexploiting the vowel inventory, allowing for much coarticulation to the right of the vowel and favoring auditorily salient segments to the left of the vowel.

INTRODUCTION

Maddieson [1] convincingly argues that cross-linguistic frequency studies of segments and segment sequences may help clarify the interplay of articulatory economy and acoustic/auditory distinctiveness in shaping sound patterns. At the same time, he regrets the insufficiency of the current statistical knowledge about the sound units of the world's languages. The study of a particular language may thus contribute directly to fill this empirical gap, shedding at least some indirect light on the theoretical issues.

There are nevertheless some more direct ways to gain theoretical insight from the statistical study of a single language. If the database is large, rich and carefully coded, segment frequencies may be taken as an estimate of the language's phonetic preferences, whatever their ultimate articulatory or acoustic interpretation may be. The question then arises whether such preferences are consistent with what is independently known about the language's sound pattern. For example, do preferred and avoided segments form phonetically consistent classes? Are such classes related to allophony and sound change trends? Do they agree with other aspects of the pattern such as syllable structure? If so, can segment preferences throw further light on the pattern itself? In particular, can they contribute to the phonetic interpretation of the segmental notation?

This paper is a progress report on an investigation of such questions in Brazilian Portuguese. Segment preferences were found to be stable in the language by comparing a dictionary to a large oral language database whose unity had been established through the comparison of several samples. The local differences between the dictionary and the oral language database, rather than obscuring their correlation, were useful in interpreting the segment preference pattern that emerged.

This pattern can be summarized as follows: Brazilian Portuguese underexploits its vowel contrasts and balances its consonant preferences between those which can be loosely coarticulated with the vowel (cf. [2]) and those which rank high in acoustic/auditory distinctiveness (cf. [3]). Such an interpretation is consistent not only with the historical and synchronic trends of the language's phonology but also with a deductive phonetic stance on the problem of sound organization and selection, as currently advocated by some leading phoneticians [4, 5, 6].

METHOD

The dictionary sample consists of the entire set of orthographic entries (27,074) from Ferreira [7]. The spoken language sample consists of 57 orthographically transcribed tapes from a public database originally collected for a nationwide oral language survey (*Projeto Norma Urbana Culta*, henceforth NURC). The NURC team had recorded a representative number of highly educated adults from five major Brazilian state capitals in various lecture and conversation settings. We have used all the available electronic data, which amounts to 72 hours of recording on 58 males and 43 females.

Both *corpora* were converted from official orthography to an abstract phonological code with the aid of a computer program which determines syllabification and lexical stress and resolves orthographic ambiguities by means of rules (including some grammatical ones) and exception lists. The initial version of this code (to be further discussed below) distinguishes a total of seven vowels [i, e, ɛ, a, ɔ, o, u] and twenty-four consonants [p, b, m, f, v, t, d, s, z, n, ɾ, l, ʎ, ɲ, j, ʃ, ʒ, k, g, w, w̃, ʁ, N]. Consonants are further classified as to whether they occur to the left (L) or to the right (R) of the vowel. All vowels are considered oral, since nasal vowels and diphthongs are dealt with by allowing three nasals (unspecified N, labiovelar w̃, and palatal j̃) in R position. For morphophonological coherence, orthographic e̲ and o̲ are always assigned to /e/ and /o/, though phonetic [i] and [u] tend to surface in most post-stressed and some pre-stressed contexts.

To further inquire into the differences between the lexicon and connected speech, two subset files were derived from the NURC *corpus*: GRAMNURC, containing all the occurrences of grammatical words, and CONTNURC, containing the remaining (content) words. A computer program was then elaborated to make the segment counts in the dictionary (henceforth DIC) and in the NURC files.

The lexicographers' convention of listing infinitives, which are rather infrequent relative to other verb forms in the language, introduced a bias for /ɪ/ in the DIC file, which was corrected by simply ignoring infinitive /ɪ/'s in the counts.

RESULTS

Figure 1 displays the segment percentages in the DIC and NURC data, with syllable positions shown in stacks. Spearman's ρ is 9.4 ($p<0.001$), which indicates a high correspondence between the two rankings. At this point, double stacks apply only to [s, ɾ, l], since the functional distinction between [i, u] and [j, w] is built into the segment code. It is already clear, in any case, that vowel selection does not conform to a "maximal contrast" [8] or "quantal" [2] principle, as the high vowel and semivowel bars generally add to less than the mid vowel bars.

Equally intriguing is the question of consonant selection. The ability to occupy R position clearly pushes [s, ɾ] into higher ranks, but the same cannot be said of [l], at least in the NURC data. As [N], in turn, ranks very high in both *corpora* without ever occurring in L position, a possible explanation for the high ranks of R consonants might be freedom to coarticulate with the vowel. This interpretation cannot, however, be pushed too far, since it leaves unexplained the lower rank of [l] as well as the general pattern of L consonants, where coronals rank higher than labials and velars.

On the other hand, the presence of a certain drive for distinctiveness is very clear in these data. Seven out of the ten most frequent consonants in both samples correspond to the "salient" segments which Stevens and Keyser [3] deduce from the combination of their "primary" features (sonorant, continuant, and coronal) with their "enhancing" features (voice, strident, nasal, etc.). This set consists of four voiceless obstruents [p, t, s, k], two consonantal nasals [m, n], and one lateral [l]. As should be expected, ranks are slightly higher in the DIC than in the NURC *corpus*, though the total percentages are about the same (26.4% *vs* 27.4%, respectively).

Let us now look at two rearrangements of the same data.

Figure 2 is a remake of Figure 1 with high vowels and semivowels collapsed into a single category (i.e., [j→i] and [w→u]). Note that the linearity deliberately adopted in the initial notation has been abandoned here: nasal semivowels are reassigned simultaneously to high vowels and [N] (i.e., [j̃,w̃→ j+N, w+N]).

Figure 3 displays the count made on the GRAMNURC and CONTNURC files, with the reduced (less linear) inventory.

It is now clear that high vowels pattern with R/L consonants in ranking between L consonants and [a, e, o]. This tendency is evident in the NURC data, and somewhat more dispersed in the DIC data, due to the high rates of [t, d]. In addition, the conformity of L consonants to Stevens & Keyser's predictions is much clearer in these graphs. Note, for example, that the set of less optimal L consonants [b, f, v, z, ʎ, ɲ, ʃ, ʒ, ʁ] - which amount to 10.5 and 7.5 % in the DIC and NURC data, respectively - is very small in the GRAMNURC file (1.17%), where prosodic weakness in-

creases the need for efficient encoding of selected distinctions.

The role of the other not so "salient" L segments [d] and [ɹ] is also clarified by comparing the GRAMNURC with the CONTNURC data. Their near complementarity in these graphs suggests that the crowded coronal region of the language's consonant space is sensitive to lexical strata, with grammatical words accounting for most of the [d]'s and content words, in turn, largely accounting for the rather conspicuous auditory contrast between [t] and [ɹ].

It follows from all of the above that the reason why Brazilian Portuguese does not fully exploit the basic vowel triangle in lexical distinctions is that acoustic distinctiveness and articulatory economy divide the spoils within the syllable: onsets tend to be auditorily salient while rhymes tend to be free from any restrictions on coarticulation. Non-high vowels are preferred underlyingly because they coarticulate easily with most R segments including high vowels while leaving enough room for surface raising, which does in fact occur very frequently.

CONCLUSION

A simple conceptual and graphic analysis of segment preferences in Brazilian Portuguese has shown that both an ideal lexicon (DIC) and the lexicon actually used in connected speech (NURC) exhibit trends which are consistent with processes observed in the language's synchronic and diachronic phonology. These trends can be summarized as a preference for loosely coordinated gestures to the right of the vowel, counterbalanced by a preference for auditorily salient gestures to the left of the vowel.

The tension generated by this asymmetry has been responsible for the historical loss of stops in syllable final position and for the emergence of syllable initial approximants in word medial position. It is also currently responsible for a number of phonological processes taking place in various dialects such as rhotacization of laterals in syllable initial clusters and loss of lateral, sibilant and rhotic gestures in some medial and final environments.

However preliminary, the results of this study support experimental approaches to phonology and point to the need for further quantitative studies of Brazilian Portuguese. In addition, they indirectly support theories which translate segments into articulatory gestures, particularly those which allow for the interaction of auditory and motor factors in their selection and distribution [9].

ACKNOWLEDGEMENTS

Support from CNPq (grant no.50.0400/90) and FAPESP (grant no.93/0565-2) is gratefully acknowledged.

REFERENCES

[1] Maddieson, I. (1993) "The structure of segment sequences", *UCLA Working Papers in Phonetics*, vol. 83, pp. 1-7.

[2] Sproat, R. & O. Fujimura (1993) "Allophonic variation in English /l/ and its implications for phonetic implementation", *Journal of Phonetics*, vol. 21, pp. 291-311.

[3] Stevens, K. & J. Keyser (1989) "Primary features and their enhancement in consonants", *Language*, vol. 65, pp. 81-106.

[4] Stevens, K. (1989) "On the quantal nature of speech", *Journal of Phonetics*, vol. 17, pp. 3-45.

[5] Ohala, J. (1990) "There is no interface between phonology and phonetics: a personal view", *Journal of Phonetics*, vol. 18, pp. 153-171.

[6] Lindblom, B. (1990a) "Phonetic content in phonology" *PERILUS*, vol. XI, pp. 101-118.

[7] Ferreira, A. (1977) *Minidicionário Aurélio*. Rio de Janeiro: Nova Fronteira.

[8] Jakobson, R. (1968 [1941]) *Child language, aphasia, and phonological universals*. The Hague: Mouton.

[9] Lindblom, B. (1990b) "Models of phonetic variation and selection", *PERILUS*, vol. XI, pp. 65-100.

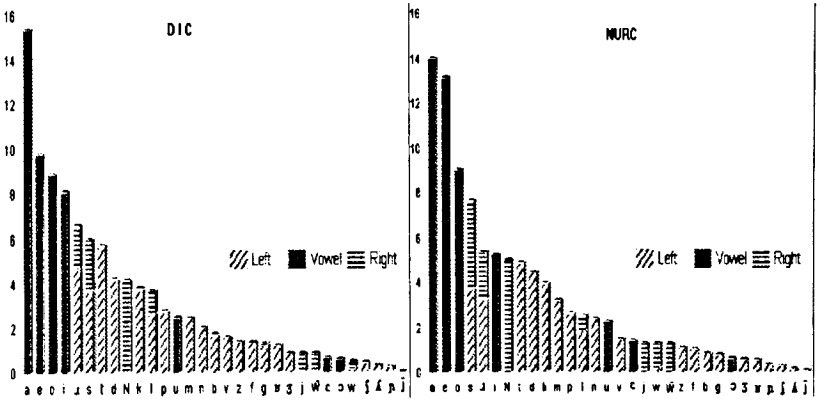


Fig. 1- Segment percentages in the DIC and the NURC samples (original inventory).

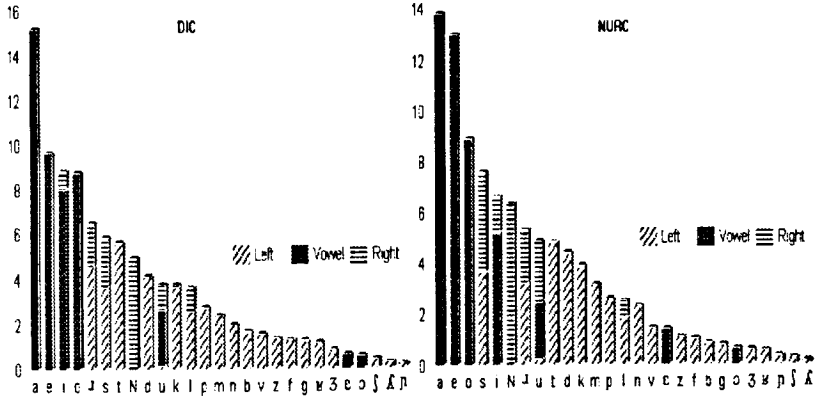


Fig. 2 - Segment percentages in the DIC and the NURC samples (reduced inventory).

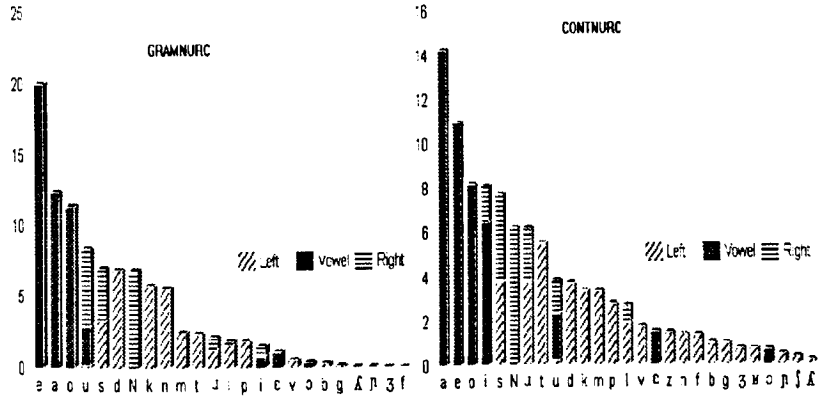


Fig. 3 - Segment percentages in the GRAMNURC and CONTNURC samples.

THE DELETION OF /d/ IN PREPOSITION 'DE' IN SOUTHERN FRENCH SPONTANEOUS SPEECH

Nadine Bagdassarian
Institut de Phonétique, URA CNRS 261 "Parole & Langage"
Université de Provence, 13621 Aix-en-Provence, France
email : phonetic@univ-aix.fr

ABSTRACT

In Southern French from Marseilles, a consonantal reduction phenomenon concerning the deletion of the /d/ of the French preposition "de" in noun phrase is described. An explanation of these segmental simplifications is proposed in the theoretical framework of generative phonology. Involved phonological processes are presented. The deletion of /d/ leads us to take into account the behaviour of its two adjacent vowels. The proposed analysis relies on both theoretical and empirical arguments, in particular on the general nature of the mechanisms of elision in French.

INTRODUCTION

Spontaneous speech is subject to various reduction phenomena general enough to be taken as such in the grammar of a particular language. The interest of such an option has considerable implications in the fields of speech synthesis and recognition, in particular, in the perspective of adapting systems to the dialect forms of specific communities of users. In the framework of an investigation of spoken Southern French, a morpho-phonological phenomenon concerning the deletion of /d/ in preposition "de" in noun phrases has been observed. An explanation of these segmental simplifications (which suppose the existence of non-reduced abstract forms) is proposed in the theoretical framework of standard generative phonology [4]. The deletion of /d/ leads us to take into account the behaviour of its two adjacent vowels.

INFORMANTS AND CORPUS

The corpus we have constituted for the general investigation of Southern French consists of tape-recorded conversations with 6 middle-class informants (30 minutes for each speaker). Since our purpose is to bring to light the linguistic regularities for a sociolect, we have to define at once a level of investigation. We can thus eliminate most sociological variability factors. For methodological reasons, the 6 male informants were chosen the same area in Marseilles. They share a number of common sociological references (sex, school level, socio-cultural category, birth place) but they are divided into three age groups (23-27 years, 58-64 years, more than 75 years). Taking into account these sociological indicators allows a better generalization of the observed phenomena.

A phonetic transcription, corroborated by a segmental acoustic analysis of the occurrences was carried out. This allows us to establish an inventory of the various realizations and to bring to light the different behaviours of the sequence.

ANALYSIS

At first, we briefly expose how the underlying representations are defined for the representation of final schwas of polysyllabic words which intervene in the subsequent devlopment. In this regional variant final schwas belong to the underlying representation of words. This position is justified for spoken French of Marseilles because of its validity at a very deep level. Brun [2] observed that in this dialect "L'*e* dit *muet*, n'est pas muet." (1978 : 31).Today, schwas are still pronounced by native speakers as described in recent studies [1, 3, 6]. Speakers also delete these final schwas ; thus the word *abeille* (bee) can also be realized [abɛj]. But by contrast to Northen French in which this deletion is obligatory [5], in Southern French it remains a variable process which is a function of different factors (social, phonetic, stylistic,...).

A last remark concerning the realization of schwa. We consider that schwa is phonetically present, only if its realization is as long as the other vowels.

We observe that /d/ is not realized when it is preceded by a polysyllabic lexical unit which ends with schwa. This does not depend on the nature, nor the number of consonants which precede the final schwa of the polysyllabic word of the noun phrase, or the one that follows the preposition, eg.: *les élèves de sa classe* (the students of his classroom) [lezelɛvəsaklasa], *le centre de gravité* (the center of gravity) [ləsɑ̃trəgravite], *la crosse de son fusil* (the butt-end of his riffle) [lakrɔsəsɔ̃fysi], la ville de Sanary (the town of Sanary) [laviləsanari], la place de l'église (the place of the church) [laplasəlegliza](see Figure 1 below).

This deletion is however optional, eg.: *la porte de l'Orient* (the door of the East) [laportədəlorjɑ̃], *une mine de charbon* (a mine of coal) [minədəʃarbɔ̃].

No deletion case has been found, when the preposition is preceded by a word ending with a consonant or a vowel different from schwa, eg.: *le gaz de ville* (the town gas) [gazdəvile], *le curé de Saint-Julien* (the priest of Saint-Julien) [kyredəsɛ̃ʒyljɛ̃].

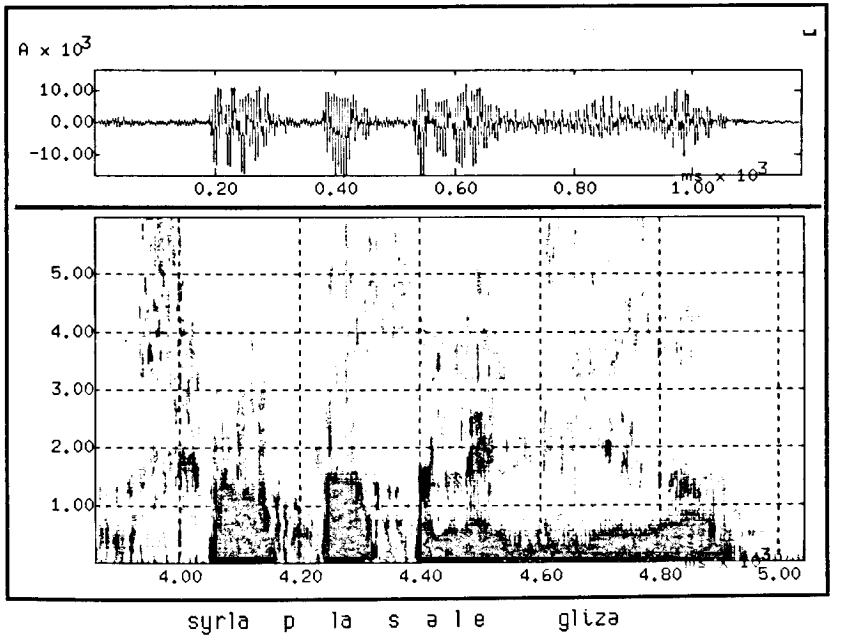


Figure 1. : Spectrogram of the phrase "sur la place de l'église"

When a pause or a hesitation mark is realized after the preposition, /d/ is never deleted, eg.: *une page de ... un cahier d'école* (a page of ... a school note-book).

In the same structural environments, the sequence /_ X#d#/ can also be realized in some different ways, in particular it can be pronounced with the deletion of the final schwa of the word that precedes "de". For example, the phrases *une bouteille de 'badoit'* (a bottle of 'Badoit'), and *l'école de la vie* (the school of life) are respectlly realized [ynəbutɛjdəbadwa], and [lekɔldəlavi].

The comparison of different realizations seems to suggest that the first way of analysing the reduced forms

consists of deleting the final schwa of the word that precedes the preposition, and then of truncating /d/. This solution presents a major drawback. When a /d/ deletion process takes place, it is likely to treat without distinction the forms derived from the phonological sequences /_Xə##də#/ whose final schwa could be deleted, and those derived from the sequences /_X##də#/, in which it must not apply in order not to generate false realizations, eg., *le gaz de ville* /gaz**də*vilə/ > *[gazəvilə].

The second alternative consists in admitting that the truncation process precedes the final schwa deletion process. So, when whatever segment, different from schwa, precedes "de", the consonantal deletion can not operate, and /d/ remains.

After /d/ deletion, the two adjacent schwas are now contiguous. However, only one of them appears in phonetic representations, the other one is deleted during the phonological derivation. Two obligatory phonological processes in French are traditionally considered responsible for schwa deletion in vocalic environment, if one admits that /ə/ can be defined like any vowels by the [+ syllabic] feature. These processes are schwa elision in pre-vocalic and post-vocalic environment, that can affect respectively the first and second schwa of the sequence. It seems a priori that there is no difference between one or the other of the processes taking in charge the deletion of one or the other of the schwas. In both cases, the required realization at the output of the phonological component is obtained. The two processes operate at a lexical-morphological level, i.e. their application domain is word bounded, in inter or intra-morphematic sequences, eg.: *une robe bleue* (a blue dress), /blø+ə/ > [blø], *princesse* (princess) /prɛ̃sə+ɛsə/ > [prɛ̃sɛsə]. But pre-vocalic schwa elision operates also, like the /d/ truncation process, at a post-lexical level, i.e. it applies between word boundaries, eg.: *l'ami* (the friend) /lə*ami/ > [lami]. The application domain of post-vocalic schwa deletion could be enlarged. But this procedure has to be set aside at once, because it has neither theoretical nor empirical scope. At the difference of pre-vocalic elision, it has only a weak application domain, limited to this particular phenomenon. In this sense, it does not satisfy the maximal generality condition required by linguistic theory. As Dell [5] underlines, "*Une grammaire doit exprimer le fait que l'existence de régularités linguistiques n'est pas fortuite, mais découle de certaines propriétés de la langue comme système stucturé. Elle doit prendre les faits dans un réseau de généralisations aussi serré que possible.*"(p. 85) On the contrary, this analysis is licit and satisfactory in the case of pre-vocalic schwa elision. Futhermore, this analysis is corroborated by an argument relying on the general nature of the mechanisms of elision in French, as confirmed by various linguistic facts.

1- The definite article or subject pronoun "la" has its vowel deleted before a feminine word begining with a vocalic initial, eg.: *l(a) armoire, je l(a) ai rangée* > *l'armoire, je l'ai rangée* [larmwara ʒalerãʒe] (I have put the wardrobe in order). In the same way, /i/ of the adverb "si" is deleted when it is followed by masculine subject pronouns "il, ils" (he, they (masc.pl.)), eg.: *s(i) il mange* > *s'il mange* [silmãʒə] (if he eats). In these two cases, the first vowel elision is represented by the symbol of the apostrophe in the orthography.

2- A frequent vocalic elision process has been observed in our corpora. Its application domain is limited to a few words that belong to non-lexical categories, like pronouns "tu" or "qui", eg.: *t(u) apprends* (you learn) > [taprã], *t(u) as* (you have) > [ta], *qu(i) était* (who was) > [kete]. In this very common use of casual French, the two successive vowels sequence is reduced again by the deletion of the first one.

3- Like the preposition "de", the /d/ of the contracted article "du" can be omitted under the same conditions, eg., *la traverse du Maroc* (the Morocco cross-bar) [latraversymarɔk], *la place du rapport* (the place of the report) [laplazyrapɔr]. The processes used are similar ones to those we have seen for "de". After the truncation of /d/, the two vowels are contiguous. The first of them, here a schwa, is deleted to the benefit of the second, /y/.

We illustrate our analysis with two derivation examples.

la ville de Sanary

/vilə**də*sanari/
d_Trunc vilə** ə*sanari
ə_elision vil ** ə*sanari
[laviləsanari]

la traverse du Maroc

/traversə**dy*marɔk/
d_Trunc traversə** y*marɔk
ə_elision travers ** y*marɔk
[latraversymarɔk]

CONCLUSION

This segmental reduction phenomenon which is not described in the literature of Southern French seems to be quite frequent among our informants. It is also generalisable to the whole group of sociolects of spoken French from Marseilles we have observed. However, nothing allows us to affirm here that these forms characterize this regional variant in particular. In other words, it could be simply specific forms of spoken French in general. Only complementary investigations can state precisely and evaluate the geographical, stylistic and sociological variability of these forms.

REFERENCES

[1] Blanchet, P. (1986), *Le Français régional de Provence, Analyse phonétique, phonologique, lexicale et syntaxique* (rôle du substrat provençal), Thèse de Doctorat de Linguistique, Centre d'Etudes et de Recherches d'Oc, Université Sorbonne-Paris IV.

[2] Brun A. (1931), *Le Français de Marseille*, Marseille, Institut historique de Provence, Laffitte Reprints, 1978.

[3]Carton, F., Rossi, M., Autesserre, D., Léon, P. (1983), *Les Accents des Français*, Paris, Hachette, collection "De Bouche à oreille".

[4] Chomsky, N. & Halle, M. (1968), *The Sound Pattern of English*, N.Y., Harper and Row.(Trad. française de P. Encrevé, Principes de Phonologie générative, Paris, Edition du Seuil).

[5] Dell, F. (1985), *Les Règles et les Sons*, Paris, Hermann, 2ème édition.

[6] Walter, H. (1982), *Enquête phonologique et variété régionale du Français*, Paris, PUF, collection "Le Linguiste".

ANOTHER MARGINAL PHONEME OF ENGLISH

Laurie Bauer

Victoria University of Wellington, New Zealand

There is a relatively large - and still growing - literature concerning a distinction in many varieties of English between a short [æ] and a longer and/or diphthongised [æ:] (Bernard 1963; Fudge 1977; Lass 1984: 34; Lawrence 1993; Trager 1930). The precise words in which this distinction is found vary from variety to variety, but may include minimal pairs such as *band* and *banned, madder* (the colour) and *madder* ('more mad'), *pans* ('pots') and *pans* ('criticises strongly'), *can* (modal) and *can* (noun or derived verb) and may give rise to pairs which do not rhyme properly such as *baddy* and *daddy, sad* and *glad, stag* and *slag, clammy* and *jammy, passion* and *fashion* and so on. What is particularly interesting about these two 'types of æ' is that most speakers appear to be unaware of them until specific attention is drawn to them, and even then may have difficulty in saying which occurs in any particular word; also, they do not behave like prototypical phonemes, most of the distribution being predictable, and the places where it is distinctive apparently differing from speaker to speaker or at least from dialect to dialect. Accordingly, people cannot use this distinction to stress which word they mean ('Did you say *madder* or *madder?'*). In all these senses, the contrast between [æ] and [æ:] is marginal in the English phonemic system.

In this paper I wish to consider what appears to be another equally marginal phoneme in my own speech. Some biographical details are thus in order. I was brought up in the north of England, from the age of seven in what is now North Yorkshire. My parents were both speakers of varieties of RP, though my mother had traces of both Scottish and Welsh English on occasions. At the age of 17 I went to university in Edinburgh, where I remained for eight years. I then went to live in Denmark for three and a half years, before moving to New Zealand, where I have now lived for 16 years. My wife is a New Zealander, as are our two children. My basic accent remains north of England, much modified towards a standard, but with influences of Scotland and New Zealand audible.

I discovered this extra phoneme in my speech by reading lists of homophones given by Carney (1994: 401-7). One of his pairs of homophones is *told* and *tolled,* and to my surprise I discovered that I did not pronounce these the same way. 'To my surprise' because I have been reading and writing transcription for many years, have always considered these two words to have the same phonemic structure, would take them to be good rhymes, and have even found myself making puns on these two words. I thus appear to have a phonemic distinction of which I was completely unaware. Moreover, having become aware that there is a distinction there, I was still not able, for a long time, to tell which phoneme occurred in which words. Within a psychological theory of the phoneme such as those proposed by Baudouin de Courtnay or Sapir, it is clear that these distinct vowels would not be regarded as different phonemes in my speech.

My first step was to attempt to characterise the two vowels in terms of their formant structure. Ten tokens of each of *told* and *tolled* and other words containing the GOAT vowel both before /l/ and in other environments were recorded in carrier sentences using the SoundScope software on a Macintosh LCIII. The carrier sentences were chosen so that the word under consideration would be in stressed but not tonic position: *He said he told the story* and *He said he tolled the bell* respectively, with a high head beginning on *said* in each case and the nucleus on *story* and *bell.* Similar sentences were used to embed other words which were considered. In each token, the onset of the diphthong was marked and its end, and five formant readings were taken (using SoundScope's LPC facility) at equal steps between these two points.[1] The first point noted showed clear influence of the preceding consonant, and is not significant; the last shows some of the structure of the /l/ in words like *told* and *tolled.* The middle three show the general trend of the diphthong. The diphthongs in the two words are indistinguishable from each other on this measure.

Next I considered the length of the diphthong + /l/ sequence in the two words. Although there was a fair amount of overlap in the lengths, nevertheless the length of the diphthong + /l/ in *tolled* was significantly greater than that for *told* ($p<0.05$ using a one-tailed t-test).

Finally I considered the quality of the /l/ in the two words: Where the two words were spoken in isolation adjacent to each other, the /l/ in *tolled* was noticeably darker than that in *told.* Using the formant structure of the /l/ derived from SoundScope and applying a simple sign test, this was easily shown to be significant ($p < 0.05$). But the fact that a significant difference can be documented here gives a new interpretation to the phenomenon. It appears that the distinction is not so much in the quality of the diphthong (though that is affected) but in the resonance (Kelly & Local 1989) of the whole syllable. The distinction is carried as much in the /l/ as in the diphthong, and it is the quality of the /l/ which makes the distinction most easily perceptible to me. Rather than saying there are two distinct vowel phonemes here, we might just as well say that there are two distinct /l/ phonemes here - though of course they would be just as marginal within the system of English as the vowels would be.

Various phenomena have been described in the literature which appear to be similar to the distribution I am describing here. In New Zealand English there is a marked difference in the phonetic qualities of the vowel in *code* and *cold,* though given the degree of /l/-vocalisation in New Zealand it might be better to say that /ʌu/ in *code* and /ɒu/ in *cold* are separate phonemes. But the element found in *cold is* found everywhere there is an underlying /l/, and so is not the same distinction as the one in my own speech. Harris (1994: 29) (following Wells 1982) talks of the *molar /roller* distinction in London English. In his data, there are two allophones of the GOAT vowel, one which occurs before /l/ followed by a #-boundary and the other of which occurs elsewhere. *Molar* and *roller* contain different variants because of the presence of the boundary in *roller.* My distinction is different from this one in that *molar* and *roller* are good rhymes, and that *toll* and *tolled* have different nuclei. Yet the presence of the #-boundary does seem to play a part in my *told* vs *tolled* distinction.

The lengthening of vowels before #d# is reminiscent of a process in other varieties of English which is usually

[1] I should like to thank Anita Easton for allowing me to use the application she had developed within SoundScope as part of her work towards an MA thesis.

described under the title of Aitken's Law. Aitken's Law, which applies to Scottish dialects of English, lengthens vowels before [v, ð, z, r] or a #-boundary. In my non-rhotic variety, [r] is not a relevant environment and [v, ð] do not appear to have any lengthening effect *(grebe* and *grieve, breed* and *breathe,* for example, having similarly long vowels). The effect of [z] is less clear in my variety, but the effect of a #-boundary is interesting, since I, like the Scots, make a difference between *brood* and *brewed.* If we extend this to *told* and *tolled,* it is arguable that precisely the same thing is happening. The case is not absolutely straightforward, because of the morphological structure of *told,* where the /d/ is presumably at least part of the marker of the past tense or past participle. However, most authorities seem to agree in seeing irregular morphology of this type being either lexical or introduced at Level I in a level-ordered model, so that there would be no #-boundary in *told* Where my variety differs from those varieties in which Aitken's Law has applied is that while Aitken's Law would predict the same nucleus in *toll* and *tolled* with a different one in *told,* I get the same variant in *told* and *toll,* and a different one in *tolled.* More accurately, I seem to get free variation between the two variants in *toll,* with the variant occurring in *told* the more common one. That is, while Aitken's Law is triggered by any #-boundary, the version in my speech is triggered by a single #-boundary, but not by a double ##-boundary. Even that is an oversimplification, since it has already been stated that *roller* and *molar* are a good rhyme for me, despite the #-boundary in *roller.* Rather it seems that my distinction is triggered by the sequence #C within the domain of the word. So *Rolls(-Royce)* and *rolls (a ball)* are different, but *roll* and *roller* are not.

At this point, there are two possibilities: either I have started to acquire Aitken's Law and have not acquired it fully, or this is a completely new rule. Pairs which I appear to keep apart (at least sporadically) by this rule include the following: *bald, bawled; band, banned; bruise, brews; choose, chews; Claude, clawed; clause, claws; find, fined; forth/Forth, fourth; furze, furs; grade, greyed; praise, prays/preys; road, rowed; seize, sees/seas; tide, tied;* Clearly, there is not just one marginal phoneme here; either there is a whole series, or we have to accept grammatical conditioning of allophones and none of these distinctions is phonemic. Such a conclusion might be strengthened by the sporadicness mentioned above. Several informants have independently suggested that it is possible to lengthen the short member of the pair under appropriate intonational conditions, but never to shorten the long member. The distinction, can, therefore, be masked even for speakers who make it.

If this is an improperly acquired rule, that is of itself interesting, since there are not many such cases attested. However, I think it more likely that this is a different process. One thing which leads me to believe that this is a completely separate rule is that it is generalised into areas where Aitken's Law does not apply (eg in making a distinction between *bruise* and *brews).* Having discovered this distinction in my own speech, I have found other speakers of English who appear to have similar distinctions. Some of these, speakers of New Zealand English, also appear to have Aitken's Law operative in their speech. If the two processes apply either independently or together, this seems like good evidence for their separateness. It is not clear to me how widespread this rule is: that remains a matter for further research.

Accordingly, I should like to postulate a new rule which I shall immodestly entitle Bauer's Law. In terms of distinctive features, this can be written

[+ syllabic] →

[+ long] / _ ([+ sonorant])#[- syllabic]

However, such a rule fails to capture the fact that where the sonorant is /l/, the /l/ (or perhaps better, the syllable) takes on a darker resonance as a result of the rule. This rule has effects which are similar to those of parts of Aitken's Law, and also reflects some distinctions between [æ] and [æ:] to which I referred at the beginning of this paper. Nevertheless, it appears to be a separate process, worthy of its own recognition.

REFERENCES

Bernard, John 1963. An extra phoneme of Australian English, *AUMLA 20,* 346-52.

Carney, Edward 1994. *A Survey of English Spelling.* London and New York: Routledge.

Fudge, Erik 1977. Long and short [æ] in one Southern British speaker's English, *Journal of the International Phonetic Association* 7, 55-65.

Harris, John 1994. *English Sound Structure.* Oxford and Cambridge, Mass.: Blackwell.

Kelly, John & John Local 1989. *Doing Phonology.* Manchester: Manchester University Press.

Lass, Roger 1984. *Phonology.* Cambridge: Cambridge University Press.

Lawrence, Wayne P. 1993. Vowel length in a New Zealand English dialect. In: *In Honor of Tokuichiro Matsuda,* Tokyo: Kenkyusha, pp. 148-161.

Trager, G.L. 1930. The pronunciation of "short A" in American Standard English, *American Speech 5,* 396-400.

Wells, John C. 1982. *Accents of English.* Cambridge: Cambridge University Press.

FEATURE GEOMETRY AND BRAZILIAN INDIGENOUS LANGUAGES (Macro-Je)

Wilmar da Rocha D'Angelis
Laboratório de Fonética Acústica e Psicolingüística Experimental
UNICAMP, Campinas (SP), Brasil

ABSTRACT

Kaingang, Xokleng and Maxakali are indigenous languages of Brazil belonging to the Macro-Je stock. This important South American linguistic stock includes more than 30 languages, all located only in the Brazilian territory. Research in these languages has shown a close relationship between the features [voiced], [nasal] and [sonorant] in phonological processes. A treatment of such processes in autosegmental phonology with the more recent "feature geometries" points to problems in the hierarchical structures attributed to such features.

PHONOLOGICAL PROCESSES of Kaingang, Xokleng and Maxakali

Kaingang is a language spoken by about 14 thousand persons living in the three southernmost states of Brazil. Xokleng is a very closely related language which is spoken by about 650 persons in the state of Santa Catarina. Finally, Maxakali is the language of an indigenous nation with about 600 persons living in Minas Gerais, a state in the southeast of Brazil.

Kaingang, Xokleng and Maxakali have two phonological processes involving voicing, nasalization and sonority: in the first, the nasal quality of a vowel in the syllable nucleus spreads to other elements in the same syllable marked with the value [+ sonorant] ; in the second, voiceless obstruents in initial word position affect nasal consonants in final position of the preceding word in relation to the features [voiced], [nasal] and [sonorant].

The first process

The first process of spreading of the feature [nasal] from the vowel in the syllable nucleus to other elements in the same syllable marked with the value [+ sonorant] reults in the nasalization of *approximants* { j, w, r } in syllables containing nasal vowels, as well as an oral contour to nasal consonants adjacent to oral vowels. Examples can be seen in Table 1.

Table 1 : First Phonological Process - Examples

Kaingang : me → ['m͡be] = "mother-in-law / aunt"
no → ['n͡do] = "arrow" nuŋ → ['n͡dug͡ŋ̃] = "stomach"
iɲ → [i:ɲ͡ʒ̃] =Pr. 1st p.sg. ŋa → ['ŋ͡ga] = "land"
meŋ → ['m͡beg͡ŋ̃] = "large" han → ['hadñ̥] = "to make"
Xokleng : kɔya + m => kɔyab͡m = "to requite"
m + a + n => m͡badn = "to kill"
m + lo => m͡blo = "to swim" pla + ŋ => plag͡ŋ = "to bite"
Maxakali : n + ay => n͡day = "clay pot"

The second process

The second process changes nasal consonants preceding voiceless obstruents (stops or fricatives) into [- voiced], [- nasal] and [- sonorant] and may occur in totality or in part, depending on whether they have been affected by the first process (Table 2). Notice that in the Kaingang examples, there are two distinct types of cases.

Table 2 : Second Phonological Process - Examples

Kaingang: *Group 1 :* ɲɔ̃n. 'kw => [ɲɔ̃n͡t'kw] = "mouth"
'mĩŋ + 'ʃĩ => [mĩŋ͡k'ʃĩ] = "cat"
kaʃĩn + 'fa => [kaʃĩn͡t'fa] = "leg of *preá*[1]"
1= kind of rodent
Group 2 : fɔg͡ŋ̃. 'fɛj => [fɔk'fɛj] = "otter"
kɔ'ʃidn + 'ʃĩ => [kɔʃit'ʃĩ] = "small son"
'kɔb͡m + 'kɔb͡m => [kɔp'kɔb͡m̥̃] = "to lighten/flash"
Xokleng : čag͡ŋ̃ . ča => čaKča = "forked"
Maxakali : mĩm . koy => mĩm͡pkoy = "canoe"
ʃo + ŋ . ʃo + m => ʃog͡ŋ̃ . ʃob͡m => ʃokʃob͡m = "animal"

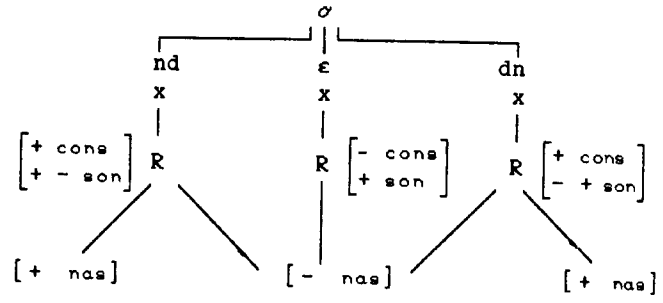
THE MANNER FEATURES

In the past fifteen years, phonological theory has advanced in the analysis of the central issue that processes often operate on consistent subsets of distinctive features within a segment [1]. The attempt to overcome the unsatisfaction of very powerful models based on a feature matrix has led to some fundamental claims of more recent models in phonology, such as the autonomy of "tiers" in the phonological representation, on the one hand, and the hierarchical characteristics of the distinctive feature structure which attempts to express the actual relations among them, on the other. The results of this search are a number of different "geometries" reflecting different analyses of the hierarchical relations of specific features.

A critical review of these "feature geometries" - from Mohanan 1983 to Clements & Hume 1993 [2] - reveals an inconsistent treatment of the so called "manner features", among them the features [nasal] and [sonorant], which are central in the two processes involved in these indigenous languages. Phonological processes involving "manner features" have thus presented difficulties for an adequate representation using such models. By way of example, in Table 3 we present a possible description of the first phonological process of Kaingang, Xokleng and Maxakali adopting the Clements & Hume (1993) geometry.

Table 3 shows the example of the Kaingang word /nɛn/ = "thing". The *spreading* of the [nasal] feature from the vowel to the nasal consonants in the syllable provokes a change in these consonants resulting in a contour [– nasal] and [– sonorant], so they become, respectively, [nd] e [dn]. This

Table 3 - Representation of the first process



requires a simultaneous change in the specification of the feature [sonorant], but, in the model of geometry used, this feature, since it is inert, is placed close to the Root node. Though change is possible in the Root, the result is a feature with double values - i.e, [+ and – Feature] - instead of a segment with double marking for the same feature, i.e, [+ Feature] [– Feature]. The difference is very important : in the first instance, it represents the abandonment of the gains in the auto-segmental view and a return at linear phonology, in line which Anderson's solution for the problem of the pre-nasalized stops [3].

For the second phonological process from Kaingang, Xokleng and Maxakali, the Clements & Hume (1993) geometry seems to provide an adequate treatment (Table 4). The solution - very similar to that given by Clements (1987) for the "intrusive stops" in English [4] - consists of a spreading of the class node "oral cavity" from the nasal consonant to the following obstruent. This simple solution, however, appears acceptable as a proof of the fitness model only if the same geometry can explain other processes involving the same features, [sonorant], [voiced] and [nasal], but it was unable to account for the first phonological process as seen in Table 3.

Table 4 - Representation of the second process

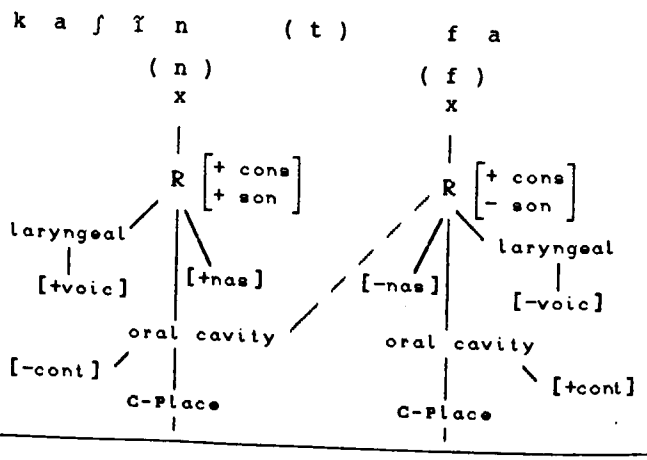
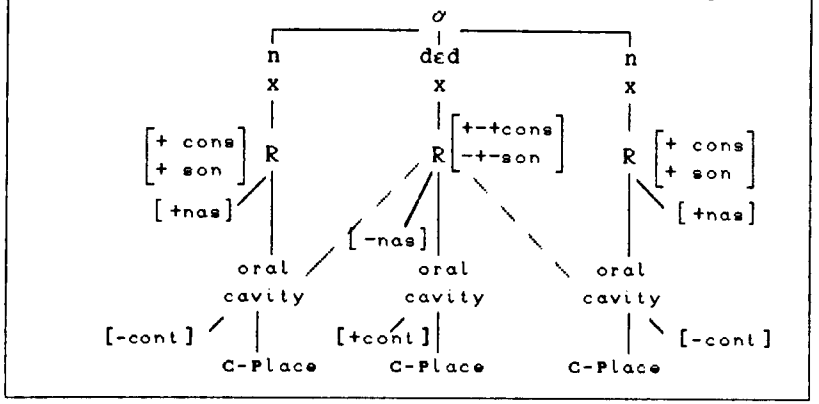


Table 5 - Alternative representation for the first process



There are other ways to attempt a solution for the first process (also inspirated by Clements 1987), such as that in Table 5, but the result is counter-intuitive.

CONCLUSION

The failure of Clements & Hume geometry to provide a solution for a fundamental phonological process in some indigenous languages of Brazil, presents practically unsurmountable difficulties for geometries which emphasize the inert or not-active characteristics of features such as [sonorant], even though they provide adequate solutions for other processes. The phonological processes of indigenous languages discussed here thus point out the need for more research on relationships among the features [sonorant], [voiced] and [nasal], and further, about how to treat features of manner in the feature geometries.

REFERENCES

[1] Cf. McCARTHY, J. J. (1988). Feature geometry and dependency: a review. **Phonetica**, 43, p.84.

[2] MOHANAN, K. P. (1983). **The structure of the melody**. Ms. Cambridge/ Mass, MIT.

CLEMENTS, G.N. & HUME, E. (1993). The internal organization of speech sounds. Version 2 (12/05/1993). Unpublished. p.52.

[3] ANDERSON, S. R. (1976). Nasal consonants and the internal structure of segments. **Language**, 52, (2):326-44.

[4] CLEMENTS, G. N. (1987) Phonological feature representation and the description of intrusive stops. In BOSCH,A. et al. (eds). **Parasession on Autosegmental and Metrical Phonology**. Chicago Linguistic Society, p.29-50.

CONCEPT FORMATION AS A TOOL FOR THE INVESTIGATION OF PHONOLOGICAL UNITS IN TAIWANESE

Bruce L. Derwing
University of Alberta, Canada

and

H. Samuel Wang
Nat'l Tsing Hua University, Taiwan

ABSTRACT

Using a concept formation task, 140 subjects were exposed to one of eight different target sets of real Taiwanese words, with each target defined in terms of a different subcomponent of a CVC syllable. Results showed that the onset targets were easier to learn than the coda targets, while the rime and body targets were about equally difficult. The general difficulty of all tasks, however, suggested that none of the words were readily analyzed into subcomponents.

THEORETICAL BACKGROUND

Since the invention of the International Phonetic Alphabet over a century ago, western linguists have tacitly assumed that the speech streams of all languages were naturally analyzable into segment- or phoneme-sized units, i.e., into individual consonant (C) and vowel (V) elements. Thus westerners have traditionally treated syllables like /na/ to be segmented into a C plus a V unit (i.e., as CV) and /tun/ into CVC, with both of examples sharing a common /n/ element in this case. Although this is a seemingly obvious conclusion to be drawn by speakers of languages that use a segment-based alphabet, such as the Latin or the Greek, the universality of this assumption may not be valid. The Chinese languages pose a particularly interesting case for testing this assumption, in that the syllable patterns they exhibit are relatively simple (often with CVC as the most complicated case), with the result that none of these languages involve more than about 1000 distinct syllable types, a number that could in principle be readily perceived and stored by speakers as unanalyzed wholes.

Though experimental psycholinguistic research in the Chinese languages is lamentably rare, one recent study has provided some hard evidence in favor of the non-segmentation or "whole syllable" hypothesis for these languages. Specifically, in an experiment performed on the Chinese Mainland, Read et al. [1] demonstrated that some native speakers of Mandarin had great difficulty in performing a seemingly simple substitution task, namely, to replace one initial consonant in a Chinese word or syllable by another consonant (e.g., to change a word like /san/ into /nan/). Interestingly, the speakers who had this difficulty were all older ones, who, though literate in the use of Chinese characters, had not been exposed in school to the Latin-based *pinyin* transliteration scheme, which is, of course, segment-based. The authors concluded from this that the segment may not be a natural phonological unit for Chinese, but only one that is learned through exposure to a writing system that represents each segment as a distinct orthographic unit.

The present study is an attempt to extend this line of research to another language in the Chinese family and to a wider range of potential syllable subcomponents. Taiwanese was chosen as the specific vehicle for this extension both because of its structural properties (specifically, because it exhibits a wider range of coda consonants than is typical of the family) and especially because no segment-based writing system is directly associated with it in the Taiwanese educational system.[1]

THE CONCEPT FORMATION TASK

The particular experimental task that we have selected for this investigation is the so-called "concept formation" technique. This is a widely used technique in experimental psychology, where it has been applied to test subjects' categorization ability with a variety of complex stimuli [2,3,4]. It has also been fairly widely used to test a variety of linguistic concepts, including word meanings [5], sentence types [6], phonemes [7], phonetic features [8] and even vowel alternation sets linked to hypothesized morphophonemic rules [9]; see [10] for an overview.

The particular version of the CF technique that is employed here is called the *identification* method [2], in which subjects are exposed to one stimulus at a time. In this task, subjects are trained to recognize a target set defined on some structural property of interest (e.g., words containing the onset /t/, which would include words like /ta51/, /tun24/, /tap33/, etc. but exclude words like /pa51/, /kun24/, /ap11/, etc.[2]). As the training proceeds, the subjects' best guesses as to the defining property are reinforced through feedback to responses of either "Yes" (used when a subject thinks that a particular test item *is* included in the set) and "No" (when he/she thinks that a particular test item is *not* included). Success in mastering a target can be measured in terms of the number of subjects who satisfy some fixed response criterion (e.g., ten consecutive correct responses), the average number of trials to reach this criterion and/or by the total number of correct responses on the test.

The main assumption that lies behind the use of this technique is that a target set will be more readily mastered if its defining property is already available to subjects (i.e., if they readily recognize it as a salient property on the basis of their prior experience) than if it is not (which means that the subject would presumably have to work it out for the first time during the course of the experiment itself). Thus for a language in which C_1VC_2 syllables were naturally broken into onset (C_1) plus rime (VC_2) subcomponents, targets defined in terms of these elements ought to be more readily identified than targets defined in terms of an arbitrary body (C_1V) vs. coda (C_2) subdivision.

TAIWANESE PHONOLOGICAL UNITS AS TARGET CATEGORIES

Our study focussed on the following eight different potential subdivisions of the Taiwanese C_1VC_2 syllable:
(1) the onset, C_1 (e.g., words beginning in /t-/);
(2) the coda, C_2 (e.g., words ending in /-n/);
(3) the nucleus V, with a fixed falling tone (e.g., words containing /a51/);
(4) the nucleus V, with varied tones (e.g., words containing any /a/ vowel, regardless of tone);
(5) tone alone (e.g., words containing a falling 51 tone, regardless of the vowel or either of the consonants, i.e., V51);
(6) the rime, VC_2 (e.g., words containing both the vowel with varied tones of (4) and the coda of (2), i.e., /-an/);
(7) the body, C_1V (e.g., words containing both the onset of (1) and the vowel with varied tones of (4), i.e., /ta-/); and
(8) the margins $C_1...C_2$ (e.g., words containing both the onset of (1) and the coda of (2), regardless of the vowel or the tone, i.e., /t-...-n/).

Since attention was focussed on potentially competing elements at both the monosegmental (onset vs. coda vs. nucleus) and bisegmental (rime vs. body vs. margins) levels, distractor sets within each level were carefully selected to maintain a balance in terms of number and range of variation from targets, and the reinforcement schedule was exactly the same across all eight target categories. The final stimulus set consisted of 100 words for each target category, half of them conforming to the defining characteristics of the target and the other half not (distractors).[3] After the master lists were organized to maximize parallelisms across target categories, the stimuli were then placed in a fixed, random order for recording and presentation to subjects.

SUBJECTS AND PROCEDURE

Subjects were 160 army recruits stationed at a military base near Hsinchu, Taiwan. Soldiers were chosen as subjects because of their representativeness in terms of educational background, and also because of their relatively limited exposure to English, especially in comparison with university students. Twenty of these soldiers were randomly assigned to each of the eight target categories. All subjects were male and each was paid NT$100 for his participation in the experiment.

Before the administration of the main experiment, subjects were given two short pretests, one to assess their ability to discriminate a sample of the stimuli and a second to illustrate the CF experimental procedure. This latter practice session consisted of 50 Taiwanese com-

pound words, 24 of which were targets defined in terms of the syllable /ki33/ as the first element of the compound. (Performance on the practice test showed an understanding of the task and varied only slightly across the subject groups. There were no discrimination problems.)

In both the practice session and the main experiment, stimuli were recorded on individual tapes for presentation in the language laboratory at the National Tsing Hua University. Each stimulus was presented twice, with a pause of 4 seconds to allow the subject to record his "Yes" or "No" response before the correct answer was provided. Subjects were tested together in groups of 40.

RESULTS AND CONCLUSIONS

Responses were recorded as 2 ("Yes"), 1 ("No") and 0 ("no response"). Because of a large number of "no response" answers (over 30% of the total), 20 subjects were eliminated from the analysis; for the remaining 140 subjects the "no response" rate averaged a respectable 2.8%. For each of these 140 subjects, a "percent correct" (%C) figure was calculated, based on total number of matched responses to the key (i.e., all mismatched and null responses were tabulated as errors).

Since less than 20% of the subjects managed to supply ten consecutive correct responses, this criterion for mastery was abandoned as too rigid under the time constraints that were imposed on responses in this particular experiment. For the purpose of making trials to criterion calculations, therefore, the criterion was set at 12 correct out of 14 consecutive responses, which does not require a long string of error-free performance but is still very difficult to achieve by chance ($p = .0065$ for a binomial distribution). For subjects who satisfied this criterion, the number of trials to reach it was tabulated directly; for the remaining subjects, a figure of 101 was arbitrarily assigned for purposes of the analysis (i.e., one more trial than the total number of stimuli provided on the test).

Table 1 below shows the number of subjects in each target group, the subset of these who reached criterion (#S), the average number of trials required to reach it (#T) and the overall mean

Table 1. Total number of subjects number of subjects reaching criterion, average trials to criterion and overall percent correct responses for each target category.

Target	N	#S	#T	%C
(1) Onset	18	7	80.4	56.9
(2) Coda	17	1	96.3	49.6
(3) /a51/	16	8	78.8	57.6
(4) /a/	20	2	96.7	55.1
(5) /V51/	19	4	86.7	53.9
(6) Rime	16	7	80.0	56.6
(7) Body	17	8	74.5	58.5
(8) Mar	17	4	88.3	51.9

percent correct (%C) for each of the eight target categories. These results indicate that the coda category (2) was more difficult to identify than either the onset (1) or the toned nucleus (3), by all three measures: far fewer subjects reached criterion (1 vs. 7 and 8), the average trials to criterion was higher (over 96 vs. about 80 for the other two), and the overall percent correct was significantly lower (49.6 vs. 56.9 and 57.6; $p < .05$). There was no significant difference in percent correct between the rime and the body, nor, in fact, between any of the other %C means that did not involve the coda. (This special status of the coda is consistent with our earlier finding [11], based on global sound similarity judgments, that the coda is the least salient component of a Taiwanese syllable.) The #S and #T figures also show a big difference between the /a51/ and /a/ targets, suggesting a problem in separating a vowel from its tone (cf. [12]); this difficulty is not reflected by a significant difference in the %C results, however.

Perhaps the most outstanding feature of the results summarized in Table 1, though, is that performance on all eight of the tasks was relatively poor. Even in the quantitatively best case (Body), fewer than half of the subjects (8/17) reached criterion over the course of 100 trials, with an average value of about 75 trials. (Ignoring the subjects who did not reach criterion, the average trials for those eight subjects who did was still a rather unimpressive 44.6.[4]) Moreover, even in this best case, the average percent correct was less than 59% (with 50% to be expected by chance for overt responses).

In sum, therefore, we conclude on the basis of the present data, at least, that while Taiwanese syllables can be subdivided into smaller units by some subjects, in response to task demands of the kind that were imposed by this experiment, most subjects find this to be a very difficult and unnatural thing to do. For most purposes, therefore, we see the syllable as the smallest viable phonological unit in Taiwanese, and probably in Chinese generally. At least this is true for speakers who have not been trained in *pinyin* or any other segment-based alphabetic scheme.

NOTES

[1]Although elementary school children in Taiwan are exposed to an onset-rime based transliteration scheme for Chinese called *chuyinfuhao*, this system is used only for the teaching of Mandarin in the early grades and is never used for Taiwanese. The same is true of the Latin alphabet, which is only used in basic English classes.

[2]The numbers in these and other examples represent tones.

[3]Since all target words had to be repeated at least once in order to achieve 50 items, 10 distractors were also repeated to discourage the strategy of focussing on repeated words.

[4]Overall, the mean #T figure is 51.3 trials for the 25 out of 140 subjects who satisfied the more rigid criterion of ten consecutive correct responses; this compares to an average of 38 trials in an earlier study [9], where 25 out of 35 high school students satisfied the same criterion on the quite abstract task of categorizing a variety of English vowel alternation sets.

REFERENCES

[1] Read, C, Y-F. Zhang, H-Y. Nie, & B-Q. Ding (1986), "The ability to manipulate speech sounds depends on knowing alphabetic writing," *Cognition*, vol. 24, pp. 31-44.

[2] Deese, J. & S. Hulse (1967), *The psychology of learning*, New York: McGraw-Hill.

[3] Dominowski, R.L. (1970), "Concept attainment," in M.H. Marx (ed.), *Learning: Interactions*, New York: Macmillan, pp. 152-191.

[4] Bolton, N. (1977), *Concept formation*, Oxford: Pergamon Press.

[5] Rosch, E. (1973), "Natural categories," *Cognitive Psychology*, vol. 4, pp. 328-350.

[6] Baker, W.J., G.D. Prideaux & B.L. Derwing (1973), "Grammatical properties of sentences as a basis for concept formation," *J. of Psycholinguistic Research*, vol. 2, pp. 201-220.

[7] Jaeger, J.J. (1980), "Testing the psychological reality of phonemes," *Language and Speech*, vol. 23, pp. 233-253.

[8] Jaeger, J.J. & J.J. Ohala (1984), "On the structure of phonetic categories," *Proc. of Tenth Berkeley Linguistics Society*, pp. 15-26.

[9] Wang, H.S. & B.L. Derwing (1986), "More on English vowel shift: the back vowel question," *Phonology Yearbook*, vol. 3, pp. 99-116.

[10] Jaeger, J.J. (1986), "Concept formation as a tool for linguistic research," in J.J. Ohala & J.J. Jaeger (eds.), *Experimental phonology*, Orlando: Academic Press, pp. 211-237.

[11] Wang, H.S. & B.L. Derwing (1993), "Is Taiwanese a 'body' language?", *Toronto Working Papers in Linguistics*, pp. 679-694.

[12] Hombert, J-M. (1986), "Word games: some implications for analysis of tone and other phonological constructs," in J.J. Ohala & J.J. Jaeger (eds.), *Experimental phonology*, Orlando: Academic Press, pp. 175-186.

ACKNOWLEDGMENTS

We express our sincere thanks to the Taiwan Ministry of Defense for permission to test the soldiers and to General Li, the Head Military Instructor of National Tsing Hua University, and Major Deng of Division 206, for coordinating the experiment. We also thank Bao-chin Hsiao for her assistance in preparing and recording the stimuli and in testing the subjects, and to Yeo Bom Yoon and Grace Wiebe for their help in tabulating and analyzing the data. This work was supported in part by a Chinese Studies Fellowship from the Taiwanese Ministry of Education awarded to the first author, and in part by a research grant from the Taiwanese National Science Council to the second author (No. NSC 83-0301-H-007-018).

THE PHONETIC FATE OF FOREIGN WORDS

P. Durand
Institut de Phonétique, URA 261 CNRS, Aix, FRANCE

ABSTRACT

A set of loanwords from American-English are examined in two languages Polish and French. Assuming that speakers of both communities tend to preserve the communicatively important aspects of speech, even with monolingual listeners, the fitting of borrowed words into their new surroundings seems to be a promising tool for isolating the main phonic aspects of a specific language.

INTRODUCTION

The increasing number of languages borrowing words from American English [1] allows the parallel study of the phonic fate of foreign words. The phonic adaptation that they undergo is restrained by linguistic factors which have to do with the listeners' perceptual process, which is language-dependant [2]. In order to point out the main aspects of this problem, a classification is proposed [3]which allows further acoustic comparisons.

THE PROBLEMS

Since the end of the XIXth century[4] the borrowings have attracted the attention of linguists, especially that of structuralists who posed the question of coexisting phonemic systems[5]. Among the various studies on the subject, we give a particular importance to those of M.Cling[6] who gives evidence that borrowings are not, at the present time, the product of monolingual subjects. But all of these works do not attach sufficient importance to the way these exolinguistic elements enter their new surrounding (the processus and its result).

The choice of languages

A-E was chosen as the source language for its importance in current international communication. Target languages ought to be sufficiently distinct to show differences both between them, and between each of them and the source language. They ought to share the same way of adapting the phonic form of foreign words instead of translating its component lexical morphemes in morphemes of its own. For this purpose, the two selected borrowing languages are French and Polish. At the stress level, as opposed to A.-E., Polish has a fixed lexical accent while French has a syntagmatic accent. At the vocalic level, Polish shows a very simple system and French a more complex one (Fig. 1 and 2). For the consonant system, Polish offers enough consonants to convey nearly all the consonant information of the source, but French does not have this possibility. Finally, if Polish gives foreign words a new spelling with respect to its own system, French keeps them in their original written form.

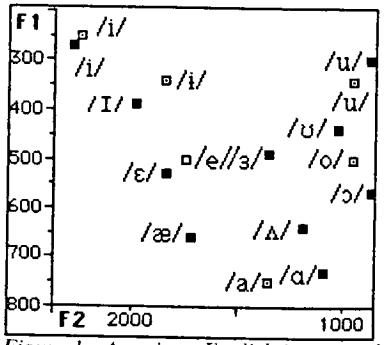


Figure 1 : American-English (■) and Polish (▫) vowels .

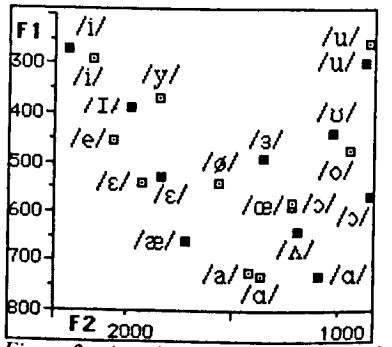


Figure 2 : American-English (■) and French (▫) vowels.

The method

A set of words that the two languages borrowed from A.-E. was selected, and rules for their adaptation were looked for. They were recorded in an sound proofed room by French and Polish speakers in order to look at the phonic fate of these exolinguistic elements. As it was not possible to find in both languages speakers with a total ignorance of the source language, it was decided to choose subjects with only an academic knowledge, i.e, nearer to the situation of a large part of the population of both countries. The borrowings were inserted in a carrying sentence with three syllables preceding the word, the word, and three or four syllables following it. The meaning and the grammatical structure of this sentence was identical. This procedure was chosen to integrate the word in its new linguistic surroundings in order to allow a pronunciation matching that of the target language.The sentence form to read were :

" Powiedział ... i poszedł"

[po'v'ɛdƶ'aw... ipɔ'ʃɛdw] in Polish

and" Il a dit ... et il est sorti"

[ila'di...eilesɔʁ'ti] in French .

Further experiment in running speech will be also necessary to collect more useful data. As dictionnaries do not mention some of the borrowings, perceptual interpretation and acoustical measurements were made, a part of the interpretation has not yet been achieved.

THE HYPOTHESIS

On the basis of the data collected a set of hypotheses was made concerning various aspects of the fitting of borrowings to their new linguistic surroundings.

- The first of them is on the stress regulation by the target language of the original one. The former is considered as an emphatic one, intended by the speaker to valorize this new significant unit, for the listener. As a consequence, it will be suppressed and will give place to, in one case the lexical penultimate stress, in the other, to the syntagmatic final accent.

- The second is that low-level rules of production and phonotactical regularization will apply to the new word.

- The third is that the transformation of the original word is largely dependent on the way - oral or written - in which the word is entering the new language. If the original spelling is kept, the graphotactical rules of the target language in the conversion grapheme-phoneme will apply unless another set of specific rules take precedence.

- The original phonemes will be transformed not only by reference to the acoustic distance from the sound of the other language but also by reference to its distribution

THE RESULTS

- The hypothesis that the stress pattern of the original word would be replaced by the Polish penultimate lexical accent and the French syntagmatic final accent was confirmed in acoustic data for trisyllabic words. Measurements of intensity, length and Fo variations were made and show stress changes. The distribution of loanwords shows small number of trisyllabic (< 16%), more monosyllabic (24,5%) and many bisyllabic words (61%). This patterns descriptive findings but not statistical interpretation, because of the small number of subjects in the corpus.

- The hypothesis of the application of low-level rules is confirmed both on the segmental and suprasegmental aspects. As A.Cutler says "listener characteristics can determine aspects of the message form"[7]. Among them, linguistic ones are important and can increase or reduce the use and the acceptance of a word.

- The phonic form of a loanword depends largely on the way a loanword enters its new language. As quoted J.B.Carroll[8], a native American imitating the French phonic word "pain" |pɛ̃| will say |pʰæn| and reading the same word will say |pejn|.As the lack of isomorphism between the orthographic code and the phonology is well known [9], if in a language the original graphy is kept, the phonic shape of the word will be different from the one of another language which has interpreted the sound sequence with the use of its phonic system and transcribed it on its own.

From the analysis of the corpus, three kinds of borrowings can be identified.

I - The original written word is kept

In this case, which is frequent in French, the pronunciation of the loanword may follow one of the graphotactic rules of the target language, keeping to the graphic unit one of the values it has in that language. So, for the "er" of "joker,charter" the graphotactic conversion will give [ɛ:ʁ] one of the values of this sequence (fer, cher, hiver) This value will be the same under stress, in unstressed or in reduced syllables. So t "a" will be a [a] in all position as in "apache".When the written element does not have the same phonetic value as the "u" of "trolleybus", it will be spoken [y] in French and [u] in Polish.The same letter may correspond to a sound or a zero as the "h", zero in French or [x] in Polish as in "hockey".

II - Phonic and graphic adaptation

This kind of adaptation is more usual in Polish, where the graphic code is phonologically regular, even if it does not deal with the palatalization oppositions.

In French, if we find "kidnappeur", "fioul", and sometime "ticheurte", this processus is not usual.

In Polish, the graphic adaptation will be conditioned by the phonic system of the language. As the vocalic system has six elements (if we do not take into account the nazalised vowels), with only one central vowel [ɨ] acoustically not very distant from [I], the central [ə], the anterior [æ] will be interpreted as /e/ phonetically [ɛ].The graphic "u" of "business" will be rewritten as "i". The English long vowels and diphthongs that have not counterparts in Polish will be interpreted when possible by two elements a vowel followed by an approximant as "crawl" in "kraul" [krawl], "hokey" in "hokey"[xokɛj]. The graphic adaptation is also systematic with [ʧ,ʤ,ʃ] which are written with their graphic Polish equivalent "cz,dż,sz". For phonotactic rules [10] /d,t,r,ʒ,ʤ/ ought to be followed by [ɨ] that will be noted "y" as in "brydż, szeryf, dżyn" for "bridge, sheriff et gin"

III - Phony and graphy are kept

In some cases, when the spelling is kept, in the target language a new graphotactic rule may appear. So for "er",or "an" often pronounced as [ɛʁ] or [ɑ̃] in French, it is possible to find [œʁ] or [an] as in "leader, speaker" or "gentleman", or [ɛn] in Polish as in "big-band". In front of this new rule, may appear alternate prononciation as in "ketchup" with [œ] ou [ɔ] in French and [a] or [u] in Polish. In this late category, some loanwords do not follow phonotactical rule of the target language as in Polish for the sequence [d,t,r,ʒ,ʤ]+[ɨ] for "sherry, jury, tee-shirt...", and sometime graphy do confirm this change "dżinsy" "jeans" instead of *"dżynsy"("dżyn" for "gin").The English affricates that have no equivalence in French are interpreted as biphonematic elements in a word "budget" [byd+ʒɛ] are simplified as word initial or final elements "jazz" [ʒaz:] get a new alternate pronunciation [ʤa:z].In this last category, we find phenomena that are missing in the two others. For these words, the linguistic fitting is not yet achieved, and it is possible to say that we see in them the process of borrowing, and in the two others, the result of this process.

CONCLUSION

From the study of loanwords, we can see firstly that the phonic shape that they adopt is a function of the target language. This divergence, as the acoustic data show, will be largely the result of the phonologic and prosodic systems of the borrowing language but also of the phonotactic rules which apply and may transform the phonic aspect of the word, even if the same phoneme exists in both languages, but not with identical distribution. Acoustic cues of the phoneme in source language are changed for the ones of the new language. Hearers will hear their input according to phonotactic rules of their language and modify both their listening and their reproduction of foreign word to fit the linguistic surrounding where it takes place.

As loanwords are often written words, it is necessary to know what kind of graphy the source language has and the target language has, and the way in which people adapts foreign words. The level of adaptation will be different, and the role of the phonic input will be more or less important.

From our data, it is possible to classify loanwords in three categories that explain the phonic fate of foreign words.

The first deals with the words for which no graphic adaptation is made. In this case, the grapheme-phoneme rules of the target language apply, when specific new rules do not operate.

The second category unites the words for which a graphic adaptation is made. In this case, the adaptation is near phonic interference, which is frequent in second language learning, with the difference that the new pronunciation is not sanctioned in learning.

The third category is the one in which, with the conserving of the original graphy speakers want to follow the original pronunciation. The study of data shows that the borrowings are recent ones and that for them two or more pronunciations coexist. In this category also, it is possible to find examples that do not follow phonotactical rules of target languages or show consonantic groups which do not exist in the target language. In this case, we have the borrowing process, and not the result of the process, as in the two former categories. Therefore, the replica is not yet assimilated, and still appears as a foreign word to the speaker.

BIBLIOGRAPHY

[1] Filipovic, R.(1982),"Phonologization and activation of latent phonemes in linguistic borrowing", *Journal of the I.P.A.*, 12:1, pp.36-47

[2] Polivanov, E.(1931),"La perception des sons d'une langue étrangère", *T.C.L.P.*, IV, pp.79-96.

[3] Durand-Deska, A., Durand, P., 1994 La forme sonore des emprunts : les mots anglais en polonais et en Français, Travaux du Claix, 12, pp.79-105

[4] Paul, H.(1886), *Prinzipien der Sprachgeschichte*.

[5] Fries,Ch., Pike,K., (1949), "Coexistent phonemic Systems", *Language*, 25, pp.29-50.

[6]Cling,M.(1980), "Aspects méthodologiques d'une expérience sur l'intégration phonématique de l'anglais par l'enfant francophone naïf", *Encrages*, N° spécial Acquisition d'une langue étrangère, Université Paris VIII, pp.41-45.

[7] Cutler,A.(1987),"Speaking for listening" *Language Perception and Production*, A.Allport, D.Mackay, W.Prinz, and E.Scheerer ed., Cognitive science series n°3, London, Academic Press, pp.23-40.

[8] Carrol,J.B.(1968),"Contrastive analysis and interference", *19th Round Table Meeting on Linguistics and Language Studies*, J.E.Alatis ed., Georgetown Univerity Press, pp.113-122.

[9] Kay,J.,(1987),"Phonological codes in reading, *Language Perception and Production*, A.Allport, D.Mackay, W.Prinz, and E.Scheerer ed., Cognitive science series n°3, London, Academic Press, pp.181-196.

[10] Bellot-Deska,A.(1989) *Aspects et temps en polonais: contribution à la traduction des formes aspectuo-temporelles de l'indicatif polonais en français*, Thèse de Doctorat, Aix en Provence

LEXICAL STRESS IN GERMAN MONOMORPHEMES

Caroline Féry
University of Tübingen, Germany

ABSTRACT

Lexical stress of German monomorphemes has been systematically studied with the help of CELEX, a large lexical database developed at the Max-Plank-Institute for Psycholinguistics in Nijmegen. The investigation has revealed that German is a quantity-sensitive language. In the second part of this paper, a theoretical account of word stress in the Optimality Theory (OT) framework is sketched. German has no exhaustive footing, but only a final trochee and an initial foot, preferably also trochaic.

GERMAN STRESS

In German, lexical stress, which we only consider on di- and trisyllabic monomorphemes, has been systematically examined by means of words listed in CELEX which contains 11,900 disyllabic words and 19,227 trisyllabics. If one eliminates compounds, proper names, derived words (most of which have an unstressed suffix), and some redundancies (like words listed with two orthographies), there remain about 2,500 monomorphemic disyllabic words and ca. 1,300 trisyllabic ones.

Disyllabic words

Some examples of initially stressed disyllabic words: (μ = mora)
2μ2μ: Gecko, Pudding, Mammut, Drama, Judo, Efeu, Scharlach, Firma, Kürbis,
2μ3μ: Pharynx, Gepard, Demut, Platin, Index, Schicksal, Turban, Kleinod
3μ2μ: extra, Arktis, Müesli, Plankton
3μ3μ: Leutnant, Sandwich, Labskaus
Some examples of finally stressed disyllabic words:
2μ2μ: Kopie, Partei, Schafott, Pardon, Büro, April
2μ3μ: Figur, Fasan, Student, Alaun, Menthol, Reptil, kompakt, Kamel
3μ3μ Symptom, extrem

Table 1. Disyllabic words

	full vowel in second syllable	schwa in second syllable
stress on the 1st syllable	577	approx. 1930
stress on the 2nd syllable	918	0

Table 2. Moraic count in initially stressed words

2μ2μ	2μ3μ	3μ2μ	3μ3μ
472	83	17	5

Table 3. Moraic count in finally stressed words

2μ2μ	2μ3μ	3μ3μ
195	706	17

Trisyllabic words

Examples of initially stressed trisyllabics:
2μ2μ2μ: (pronounced as disyllabics) Prämie Linie Stadion Thymian
2μ2μ2μ: (true trisyllabics) Embryo Ozean Pinguin Februar stereo Exodus Kolibri
2μ2μ0μ: Roboter Araber
Examples of medially stressed trisyllabics:
2μ2μ0μ: Schimpanse Oktober Forelle Lavendel
2μ2μ2μ: Inferno Albino Veranda Pyjama Meniskus Bazillus Inspektor Professor
Examples of finally stressed trisyllabics:
2μ2μ3μ Appetit Kormoran Vitamin Diamant Katafalk Vagabund
2μ2μ2μ Omelett Karusell Kompromiß Theorie

Table 4. Trisyllabic words

stress on the first syllable	255 (38 with final schwa, 217 others)
stress on the second syllable	664 (528 with final schwa, 136 others)
stress on the third syllable	393

Table 5. Moraic count of the stressed syllable

	2μ	3μ
Stress on the first syllable	254	1
Stress on the second syllable	663	1
Stress on the third syllable	94	299
Total	1011	301

OPTIMALITY ACCOUNT

German has nonmoraic syllables (with a schwa or a syllabic sonorant as nucleus), bimoraic syllables (with a tense vowel, or a lax vowel plus a consonant), and (usually final) trimoraic syllables (tense vowel plus consonant or lax vowel plus two consonants), but no monomoraic syllables (consisting of a lax vowel). Final trimoraic syllables attract stress; otherwise the default stress location is on the penult. The antepenult can be stressed if the final syllable is bimoraic and the penult is open (*Páprika, Braütigam*, but *Veránda, Forélle*).

Stress can also be prespecified: some final bimoraic syllables are stressed (*Theorie, Spinétt*). Optimality Theory (presented in [1] and [2]) accounts for these intricate data in an elegant way. The following constraints are needed: FOOT-BINARITY (Feet must be binary under syllabic or moraic analysis), FOOT-FORM (TROCHAIC) (Ft –> σ_s σ_w or Ft –> σ_s), ALIGN-TROCHEE-RIGHT (Every Prosodic Word ends with a syllabic trochee), ALIGN-FOOT-LEFT (Every Prosodic Word begins with a foot).

Figure1. Tableau of the word 'Mámmut'

σ σ /\ /\ μμ μ μ | | | m a m ʊ t *Mammut*	FOOT-BINARITY	ALIGN-TROCHEE-RIGHT	ALIGN-FOOT-LEFT	FOOT-FORM (TROCHAIC)
a. (x .) ☞ Mammut				
b. (x) (x) Mam mut		*!		**
c. (. x) Mammut		*!		*
d. (x) Mammut		*		*

Figure 2. Tableau of the words 'Sekúnde'.(On the bottom of the tableau, the stress pattern of the quadrisyllabic *'Apotheóse'* and *'Marsupilámi'* is shown.)

σ σ σ /\ /\ μμ μμ \/ | | ze ku n də *Sekunde*	FOOT-BINARITY	ALIGN-TROCHEE-RIGHT	ALIGN-FOOT-LEFT	FOOT-FORM (TROCHAIC)
a. (x) (x .) ☞ Se kunde				*
b. (x .) (x) Sekun de	*!	*		*
(x .) (x .) ☞ A p o the óse ☞ Marsu pi lámi				

Further relevant constraints are: ALIGN-HEAD (The right edge of every PrWd coincides with its head, THREEMORAS = TWOGRIDPOSITIONS. This last constraint forces a trimoraic syllable to project two grid positions, as proposed in [3], thus to form a trochee.

Figure 3. Tableau of the words 'Kamel'

σ σ /\ /\ \ μμ μμ μ | | V | k a m e l *Kamel*	3μ = 2x	FOOT-BINARITY	ALIGN-TROCHEE-RIGHT	ALIGN-FOOT-LEFT	ALIGN-HEAD	FOOT-FORM (TROCHAIC)
a. x (x) (x.) ☞ Ka mel						*
b. x (x .) Kamel	*!					
c. x (. x .) Kamel		*!	*			*
d. x (x) (x.) Ka mel					*!	*

Prespecified stresses are accounted by a constraint called FAITH (All input grid positions are kept) which, like FOOT-BINARITY and THREEMORAS = TWO GRIDPOSITIONS, is unviolated.

Figure 4. Tableau of the words 'Spinett

σ σ /\ /\ μμ μμ | | | Sp i n e tt *Spinett*	FAITH	FOOT-BIN	ALIGN-TRO-RIGHT	ALIGN -FT-LEFT	ALIGN-HEAD	FOOT-FORM (TROCHAIC)
a. x (x) (x) ☞ Spinett			*			**
b. x (x.) Spinett	*!					

REFERENCES

[1] McCarthy, J. & A. Prince (1993) *Prosodic Morphology I: Constraint interaction and satisfaction.* To appear, MIT Press. (also: Technical Report #3, Rutgers University Center for Cognitive Science.)

[2] Prince, A. & P. Smolensky (1993) *Optimality Theory: Constraint interaction in generative grammar.* University of Colorado, Boulder: Ms.

[3] Prince, A. (1983) Relating to the grid. *Linguistic Inquiry 14.* 19-100.

SOME IMPLICATIONS FOR GESTURAL UNDERSPECIFICATION AS A RESULT OF THE ANALYSIS OF GERMAN /t/ ASSIMILATION

Anja Geumann and Bernd Kröger
Institute of Phonetics, Cologne, Germany

ABSTRACT

As can be shown for English data, the assimilation of the alveolar stop can result from an increased gestural overlap of the following oral closure gesture. Our experiment with German synthetic speech showed similar results. Further, it suggests that it is neccessary to complete the gestural specification of the glottal state. A voiced stop should be represented not only by an oral gesture, but by a glottal one as well.

1. INTRODUCTION

In German as well as in English, reduction or assimilation of word final alveolar stops in the case of a following labial or velar stop can be found. These phenomena have been described as speechrate, -style dependent. In a gestural approach [1], [2], they can be explained by increasing gestural overlap of the following oral closure gesture.

It is not yet clear what types of gestures are explicitly given for the glottal state, e.g. opening, critical closing. At first approximation, a total underspecification of voicing is found to be underlying, i.e. the default is an critical closed glottis (phonation), even in the case of an oral closure (a stop) [3], [4].

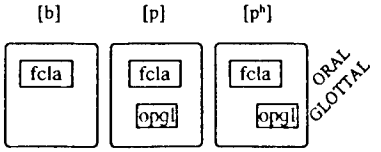


Figure 1. Gestural representation of voicing contrasts. (To be revised). Vertical length of the boxes represents gestural duration.

The sound thus results from the activity of gestures in a certain time interval and in a particular temporal relation, e.g. between labial closure gesture and glottal opening.

In our gestural model, the /b/ vs. /p/ distinction in German is represented as the leftmost vs. the rightmost gestural configuration in Figure 1.

Our goal is to model the assimilation in the sequences /t#p/ and /t#b/, where it is that an intermediate morpheme or word boundary is found. As proposed in segmental descriptions, e.g. [5], /t/ should be assimilated into the place of the following labial stop, as in [tʰb] > [p̪] and [tʰpʰ] > [pʰ].

2. EXPERIMENT

In our experiment, synthetic speech stimuli were used containing /t#p/ and /t#b/ stop sequences (as a minimal pair contrast) in three-word phrases: <Er geht packen> ('He is going to pack'), <Er geht backen> ('He is going to bake'). The stimuli were synthesized using a gestural based articulatory speech synthesizer at the Cologne Phonetics Institute [6]. In this model, a self-oscillating glottis model is implemented [7]. In both contexts the gestural overlap of the following labial gesture was increased in 7 equidistant steps from 0% to 100%, as could be seen in Figure 2. The glottal gestures are temporal associated to the respective oral gesture. Thus, the overlap of the glottal gestures increases in the same way.

In a second part of the experiment, in addition to the overlap of the oral gestures, the first glottal opening gesture (associated with the apical gesture) decreases in temporal and spatial extension (see Figure 3). Thereby, the offset of the glottal gesture is synchronized with the onset of the labial gesture. The stimuli were randomized and presented to 32 native listeners, without experience in synthetic speech. In a forced choice situation they were asked to decide whether or not there was /t/, and if there was /p/ or /b/.

In this way, listener judgements should show if place assimilation was perceived at all, and, whether there was - as proposed - a perceptual salient /b/ vs. /p/ discrimination even at full overlap.

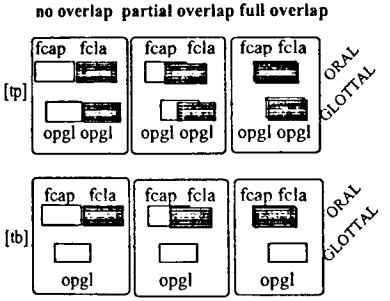


Figure 2. Increasing gestural overlap in the /tp/ (upper row) and /tb/ (lower row) sequence, with unreduced first glottal opening gesture. Gestures: fcap - apical full closure; fcla - labial full closure; opgl - glottal opening.

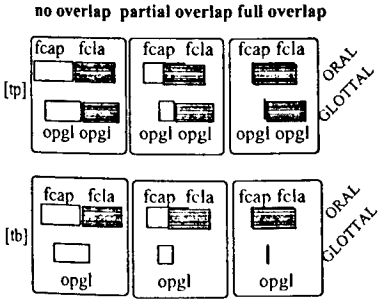


Figure 3. Increasing gestural overlap and decreasing glottal opening gesture, with a total reduced opgl at full overlap. In fact, at full overlap, the (first) opgl has disappeared.

Firstly, it was found in the experiment that gestural overlap is perceived as /t/ assimilation or reduction. The results for /t/ assimilation were significant in all four contexts (/tb/, /tp/; with/without glottal reduction). In Figure 4 a general survey is given. (It should be noted, that the data in Figure 4 differ from those of Byrd [2]. This might result from the somewhat differing gestural specification. Within the model used by Byrd, after 67% of gestural inherent time, the full closure was reached; in our configuration, this occured after 45% of inherent time. This might have led to assimilation judgements at a lower percentage of overlap in our experiment.)

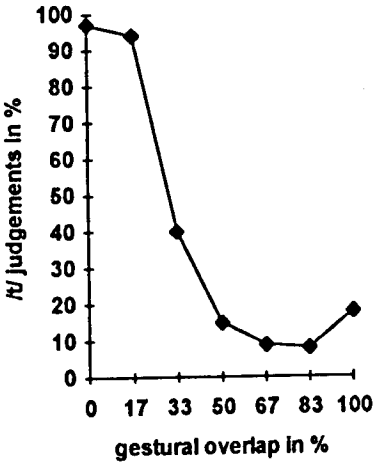


Figure 4. /t/ judgements for all four contexts. (non-/t/ = assimilation)

Secondly, the results suggest that there has to be a reduction (in space & time) of the glottal opening gesture in the /t#b/ context. In the /t#p/ sequence, there was no significant difference in whether the first glottal opening gesture was reduced or not. As seen in Figure 5, listeners tend, with increasing gestural overlap, to perceive a /p/ in the /t#b/ sequence if the preceding glottal opening

gesture is not reduced. The acoustic data in Table 1 confirm this.

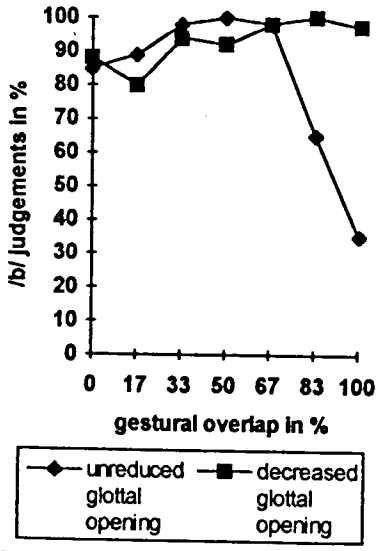


Figure 5. /b/ judgements for the /tb/ sequence. (non-/b/ = /p/ judgement)

Table 1. Duration of aspiration of the /tb/ sequence, without reduction of opgl.

Overlap in %	Aspiration in ms
0	-
17	-
33	-
50	5
67	5
83	20
100	25

3. CONCLUSION

There seems to be a kind of asymmetry in the behavior of glottal gestures for /tp/ vs. /tb/. For /tp/, we assume that it is unnecessary to reduce the first glottal opening gesture as in Figure 2. In the /tb/ sequence, on the other hand, it was shown necessary to decrease the glottal opening gesture, as in Figure 3.

The temporal association of the glottal gesture with the onset of the following oral gesture, certainly, is a first attempt. If one does not want to give up the autonomy of the oral and glottal articulators, the problem should be solved within the glottal articulator. It seems reasonable to secure the voicing of the labial stop by an additional glottal gesture.

For this reason, we suggest introducing **a glottal critical closing gesture** associated with the '/b/' oral closing gesture in German, which dominates (blends with) a preceding glottal opening gesture. This means one must have contrastive gestural specification for the glottal state of stops.

Further, if the glottal critical closing gesture (ncgl), as in Figure 6, has only a short duration, it allows at least partial devoicing, as might often be the case in German [p̥], see [5].

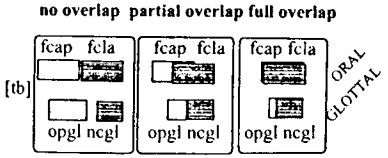


Figure 6. Contrastive gestural specification of the glottis.

The introduction of this additional gesture also seems to be appropriate from a phonological markedness point of view. This means, in a gestural model, a voiced stop should be specified with no fewer gestures than a (unmarked) voiceless stop.

ACKNOWLEDGEMENTS

Many thanks to Louisa Schafer for checking our English.

REFERENCES

[1] Browman, C.P. & L.M. Goldstein (1990), Tiers in articulatory phonology, with some implications for casual speech. In: J.H. Kingston & M.E. Beckman, eds. *Papers in Laboratory Phonology. Vol. I.: Between the Grammar and Physics of Speech.* Cambridge: CUP, pp. 341-376.

[2] Byrd, D. (1992), Perception of Assimilation in Consonant Clusters: A Gestural Model. *Phonetica* 49, pp. 1-24.

[3] Goldstein, L.M. & C.P. Browman (1986), Representation of voicing contrasts using articulatory gestures. *Haskins Laboratories: Status Report on Speech Research* SR-85, pp. 251-254.

[4] Schiefer, L. (1989), 'Voiced aspirated' or 'breathy voiced' and the case for articulatory phonology. *Forschungsberichte des Instituts für Phonetik und sprachliche Kommunikation der Universität München* (FIPKM) 27, pp. 257-278.

[5] Kohler, K.J. (1990), Segmental Reduction in Connected Speech in German: Phonological Facts and Phonetic Explanations. In: W.J. Hardcastle & A. Marchal, eds. *Speech Production and Speech Modelling.* Dordrecht: Kluwer, pp. 69-92.

[6] Kröger, B.J. (1993), A Gestural Production Model and Its Application to Reduction in German. *Phonetica* 50, pp. 213-233.

[7] Kröger, B.J. (1990), Three glottal models with different degrees of glottal source - vocal tract interaction. *IP Köln Berichte* 16, pp. 43-58.

CODA CONDITION IN ITALIAN AND UNDERSPECIFICATION THEORY

Giovanna Marotta
Department of Philology, University of Siena, Italy

ABSTRACT

In Italian syllable coda licenses only a skeletal slot, associated with a small set of consonants. The coda condition which filters the segments admitted in this syllable constituent has to be sensitive to the manner as well as to the place of articulation. According to Underspecification Theory, [coronal] is assumed to be the unmarked place of articulation. Therefore, the coda condition would simply mark as [cont] the segment associated with that position.

CONSONANTS IN CODA

Like many other natural languages, Italian shows asymmetric distribution of segments in Coda with respect to the Onset, since the latter is much richer than the former, both regarding the number and kind of segments involved. The reasons for such asymmetry are both phonetic and phonological. Phonetically, we should consider at least the different degree of muscle tension as well as the strength of articulation. Phonologically, the typological studies carried out since Jakobson have documented the greater diffusion of the onset constituent over the coda (the well-known universality of the CV syllable).

In the syllable structure of Italian, the onset can be empty (e.g. *a.mico* "friend") or filled by one or two consonants (e.g. *pa.ne* "bread", *tre.no* "train"). The coda also can either be empty (e.g. *fi.ni* "(he) finished") or it projects on the skeletal tier a position, which may be filled by a glide, in case of a falling diphthong (e.g. *pau.sa* "pause") or by a small set of consonants (see Nespor [1] for a general sketch of the Italian syllable template). We will consider here only the consonantal codas, in order to find an autosegmental condition which could represent the constraints available in the language.

The consonants admitted in coda are the following:

a. dental liquids; e.g. ['salto] "jump", ['pɛrla] "pearl", ['kɔrpo] "body"
b. dental sibilant; e.g. ['asta] "pole", ['vɛspa] "wasp";
c. homorganic nasals; e.g. ['kampo] "field", ['dɛnte] "tooth";
d. the first half of a geminate; e.g. ['fet:a] "slice", ['pɛl:e] "skin".

To capture significant generalizations from this picture is not simple, since the segments occurring in coda do not seem to form a natural class in terms of traditional phonological features.

A FIRST CODA CONDITION

Considering the articulation manner, we find that a formulation of the Coda Condition in terms of the features [snrt] and [cont] cannot exaustively account for the linguistic data.

Itô [2] proposes the following filter for the coda in Italian:

1) * C $
|
{[- snrt]
[- cont]}

But such a filter is incorrect, since it allows all liquids and nasals, which are [+snrt], while in actual fact only the dental [l r] are found, and not even the palatal /λ ɲ/. On the other hand, the occurrence of /s/ is predicted, since it is [+cont], but also that of the other fricatives of Italian, i.e. /f v ʃ/, which do not occur in coda position. Moreover, it is unclear why the nasal in coda has to be homorganic with the subsequent consonant; e.g. ['kampo] "field", not *['kanpo], ['baŋka] "bank", not *['banka]).

As to /ʃ ɲ λ/, there are arguments for claiming that these consonants are underlyingly long. Firstly, they cannot form consonant clusters with other segments (e.g. *-VrɲV/, *-VɲrV/. Secondly, /ʃ ɲ/ select the allomorphs ending with vowel of the article and other determiners (e.g. [lo 'ʃ:ɛ:mo], *[il 'ʃɛ:mo] "the fool", [,kwel:o 'ɲ:ɔ:mo] "that gnome"); the palatal liquid can occur only in internal position of the word (e.g. ['fiλ:o] "son", but *[λ:o-]). Therefore, the treatment of these segments should be considered part of the more general case of geminates.

The intrinsecally heterosyllabic nature of a geminate allows it to occupy the coda constituent. In fact, in non linear phonology, a geminate projects two skeletal slots, the first one associated with the coda and the second one with the following onset; i.e.:

2) C:
/ \
x x
| |
Cd O

Given this representation, each geminate segment occurring in the language can fill the syllable coda, since in this case the phonological content of the coda constituent is totally governed by the onset.

As far as the rule order is concerned, a filter like 1) should follow the sillabification by default of a long consonant as in 2). However, with a negative condition such as that formulated in 1), it is still impossible to explain the lack of fricatives other than /s/ in coda or the constraint on homorganicity holding for nasals.

UNDERSPECIFICATION

If we once again consider the inventory of segments allowed in coda, we observe that, apart from geminates and nasals, all other segments share the same point of articulation, i.e. they all are dental. In terms of articulatory features, they are marked [+coronal] and [+anterior]. Inside the feature geometry, the Coronal node is assigned a special status by the Underspecification Theory, both Radical and Contrastive: Coronal is assumed as the unmarked Place of Articulation. This hypothesis is supported by different arguments: a) coronal segments are normally present in the inventory of natural languages; b) the number of coronals in a language is normally higher than that of other points of articulation; c) in language acquisition anterior coronals appear early; d) they show a special behaviour in phonological processes, as they are often assimilated and, at the same time, transparent to vowel harmony; e) in linguistic errors, especially in the case of substitutions, coronals are involved more often than segments in other articulation places. All these data together indicate that the coronal place is special, i.e. it is unmarked (see [2] and [3]).

The basic idea of Underspecification Theory is that the underlying representation (UR) shows only the relevant information, while redundant values and unmarked features are excluded. Among coronal segments, only the anterior ones (i.e., dental or alveolar consonants) are recognized as unmarked for Place at the level of UR.

In Italian, there are two series of coronal segments: the dental /t d s ts dz l r n/ and the palatal /tʃ dʒ ʃ ɲ λ/. Only the first ones are anterior and therefore they can be assumed as being unmarked for Place of Articulation.

According to the Underspecification Theory, we could assume that in Italian only segments unmarked for Place in UR are licensed in coda. The corresponding filter would have the following form:

3) * C $
|
Pl Art

With such a negative condition, all segments of the anterior coronal series would be possible in coda position. But, again, this would not be correct, since dental plosives and affricates, both [+cor] and [+ant], are not allowed in coda position, unless they are geminates. Something else is then necessary in order to exaustively capture the data of the language.

A SECOND CODA CONDITION

Baroni [4] has recently proposed a Coda Condition which, assuming the underspecification for Place of anterior coronals, prevents the occurrence of dental stops and affricates by the addition of the feature [stop]; formally,

4) * C$
|
Pl Art
[+stop]

Baroni rejects the traditional feature [cont] and instead introduces two distints features, [stop] and [fricative].

This hypothesis is grounded partly on the analysis by Lombardi [5], who refu-

ses a representation of the affricates as contour segments with two ordered values of the feature [cont], since this representation would not be able to interpret some relevant phonological processes occurring in natural languages. Lombardi proposes to replace [-cont] with [stop], maintaining the [+cont] feature, while Baroni claims also the substitution of [+cont] with [fricative].

The introduction of two features instead of the more traditional [cont] allows to keep separate the different obstruents of the dental series: /t d/ will be marked by [stop], /s/ by [fricative] and /ts dz/ by both these features. However, such an hypothesis, which appears to be supported only by the represention of affricates as complex segments, is rather expensive in the phonological analysis. Since Baroni follows the framework of Radical Underspecification, he is obliged to represent the affricates as marked by two different features; but, in Contrastive Underspecification, a same feature can receive a different value, even at the UR level too.

On the other hand, apart from obstruents, the new features could not be applied, while [cont] is relevant even in the interpretation of other segments. Moreover, the representation of affricates as complex segments with two unordered features, [stop] and [fricative], does not hold for the palatal series, where, besides /tʃ dʒ/, there are not palatal stops in Italian. For palatal affricates, it would be necessary to postulate - as Baroni does - a redundancy rule which assigns the values [+stop, +fricative] to the coronal segments [-ant] which are unmarked by other features in UR, with a conflation of the component relative to the derivation.

Finally, with the introduction of [stop] and [fricative], a coronal plosive like /t/ must be marked by [stop] in the UR, in order to be opposed to the fricative /s/ on one hand and to the affricates on the other; at the same time, the other plosives, /p/ and /k/, have to be marked for place, by [labial] and [dorsal], respectively. In this way, since all the stops are marked by at least one feature in UR, coronals risk losing their special status of being consonants more underspecified than the others.

As a consequence of the afore mentioned arguments, we do not agree with the proposal of rejecting [cont] in favour of two different features, [stop] and [fricative].

A NEW CODA CONDITION

Another Coda Condition for Italian can be formulated, a condition which takes into basic account the feature discussed up to now. Sharing the hypothesis of anterior coronals as unmarked for Place in UR, we propose that the segment associated with the coda position has to be marked [+continuant]. The underspecification for Place of anterior coronals on one hand and the constraint on [cont] on the other lead to the following filter:

5) * C $
  |
  Pl Art
  [- cont]

Such a negative condition predicts that the consonants admitted in coda have to be marked by the positive value of [cont], but, at the same time, the point of articulation in UR must be absent.

To assume [continuant] as the feature constraining the segments licensed in coda position entails the marking of this feature at the level of UR. As we saw, the adoption of [cont] is problematic for the representation of the affricates in Radical Underspecification, where only one value for each feature is admitted. This is not the case in Contrastive Underspecification, where the same feature can assume different values in UR in relation to the specific contrasts occuring in the language.

In the framework of Contrastive Underspecification, an affricate like /ts/ can be represented with reference to [cont] either as a contour segment (see 6a) or as a complex segment (see 6b):

6a) /ts/ 6b) [-cont]
  | |
  x x
  / \ |
[-cont] [+cont] [+cont]

The dental affricates will be kept distinct from the fricative /s/, which is only [+cont] and from the stops /t d/, which are only [-cont].

LICENSING

However, the Coda Condition formulated in 5) as a filter is able to license directly /s l r n/ as possible codas. It should be underlined that we assume that nasals segments are marked as [+cont], like liquids and fricatives. Such a mark does not imply the absence of some occlusion in their articulation.

We have now to interpret the homorganicity constraint on nasals as well as the occurrence of geminates in coda position. In both cases, we believe it is relevant to make reference to the notion of licensing as formulated in Goldsmith [7]. The basic assumption is the recognition of the coda as a secondary licenser with respect to the other syllable constituents, i.e. onset and nucleus. The reduced autosegmental potential assigned to this syllable constituent accounts for the fewer contrasts available in coda, as to those possible in onset. As regards the topic under examination, the coda crucially does not licence a point of articulation autosegment.

In other words, in Italian, as in many other languages, the coda licenses only a skeletal slot, associable with a small subset of features, which do not include those relative to place. Thus, consonants that appear in coda position may receive by default an unmarked point of articulation, i.e. they must be coronal or can be filled by the autosegmental content of a following onset.

In the case of the nasals, we assume that a nasal licensed in coda is underlyingly coronal. For Italian, such an interpretation is supported by the occurrence only of /n/, and not of other nasals, in word final position in functional, clitic elements, such as prepositions (e.g. *in* "in", *con* "with") as well as in determiners, like articles and adjectives (e.g. *un* "a, an", *nessun* "no, not any"). The homorganicity constraint on nasals may be directly derived from their syllable position: in coda, a nasal cannot license an autonomuos point of articulation. Therefore, it is either coronal or it assumes the place of the following consonant in onset. Even homorganic nasals thus appear to respect the Coda Condition formulated in 5).

Regarding geminates (palatals /ʃ ɲ ʎ/ included), we saw already that they project two skeletal slots, which are associated by default with different syllable constituents: the first slot goes with the coda, while the second one with the following onset (see 2) for the formal representation). Since in the case of a geminate the coda is licensed by the following onset, its position is basically empty, i.e. without phonological content. The coda becomes filled by the autosegmental spreading of all the features from the onset position.

The syllabification by default of geminates entails the possibility for these segments to occupy the coda position. At the same time, their special licensing from the onset position allows segmental features normally not permitted in coda to be present in this position. For instance, in Italian the feature [-cont] is prevented from occuring in coda by the Coda Condition we proposed (see 5). However, it may mark a coda position in the case of a long consonant (e.g. ['fet:a] "slice", like ['faʃ:a] "band").

In such a perspective, geminates once again must be syllabificated before the Coda Condition, which in fact does not apply to long consonants. After the syllabification of the geminates by default, as in 2), the filter given in 5) will apply to the segmental strings.

REFERENCES

[1] Nespor, M. (1993), *Fonologia*, Bologna: il Mulino.
[2] Itô, J. (1986), *Syllable theory in prosodic phonology*, Ph.D. Thesis, University of Massachussets.
[3] Paradis, C. & J.F. Prunet (1991), (eds.), *The special status of coronals: internal and external evidence*, New York: Academic Press.
[4] Marotta, G. (1993), "Dental stops in Latin: a special class", *Rivista di Linguistica* , vol. 5, pp. 55-101.
[5] Baroni, M. (1993), "Teorie della sottospecificazione e restrizioni sulle code consonantiche in italiano", *Rivista di Grammatica Generativa*, vol. 18, pp. 3-59.
[6] Lombardi, L. (1990), "The nonlinear organization of the affricate", *Natural Language and Linguistic Theory*, vol. 8, pp. 375-425.
[7] Goldsmith, J. (1990), *Autosegmental and metrical phonology*, London: Blackwell.

A PROSODIC ACCOUNT OF ENGLISH VOWEL LENGTHENING

Geoffrey S. Nathan
Southern Illinois University at Carbondale

ABSTRACT

The famous rule of English Vowel Lengthening may not be directly attributable to the voicing of the final consonant, but may rather be related to the rhythmic organization induced by stress-timing in English.

INTRODUCTION

One of the classic problems in English phonology/phonetics has been the rule lengthening vowels before syllable-final voiced consonants. As is well-known, at least in American English, vowels are much longer before syllable final voiced consonants than they are before syllable-final voiceless consonants. The reason that this is a problem is that the lengthening is far too much to be accounted for by the universal phonetic effect found in other languages. For example, French shows lengthening, but the difference is on the order of 10% or so, while American English lengthening may approach 100%, at least in utterance-final position. For example, the Klatt synthesizer calculates a value for /ai/ before a voiced stop as 286ms, and before a voiceless stop as 167, a ratio of 1.71:1 [1]. So the question is, why does English have this rule?

Some have suggested that this is an instance of phonologization—the exaggeration of a pre-existing tendency for phonological purposes. The problem with this proposal is that voicing lengthening is a purely allophonic, or post-lexical process. It is a typical instance of something that native speakers are not consciously aware of, but which can be brought to consciousness in a phonetics classroom. As numerous phonologists have said, rules at this level are not normally available for conscious manipulation. Consequently, it is unlikely to have been 'seized upon' by the language for exaggeration.

In addition, no other instance of a phonetically-motivated allophonic rule that I am aware of has this property—namely that a universal, speech implementation tendency is exaggerated, stretched or otherwise distorted, resulting in an allophonic rule. For example, languages normally front velar stops before front vowels: the point of articulation of the stop in 'key' is different from that of 'caw'. However, I know of no language in which this fronting has been extended to front velars to, say palato-alveolars allophonically (although subsequent language change may make alternations between velars and palatals a morphologically-conditioned rule in the language). Similarly, we find that aspiration is longer after velars than after labials, but no language makes velar aspiration longer still (or conversely, deaspirates labials). In languages that have aspiration it is generally the same length cross-linguistically. It appears to be only English vowel-lengthening that is so extreme. This leads us to wonder whether the length alternations found connected with voicing contrasts in English are in fact caused by the voicing of the following consonants at all, and are in stead due to other features of English.

BISYLLABIC SHORTENINGS

There are other principles governing vowel length in English, but they are not related to segmental factors at all, but rather deal with metrical structure. Syllable length in English is sensitive to foot type. I am suggesting here, instead, that vowel lengthening is a rhythmic phenomenon, and is somehow related to the mapping of syllables onto timing beats in speech production.

It has been argued for a long time that English is a stress-timed language (Classé, [2] is the first definitive discussion). In a stress-timed language the same amount of time is assigned to every foot, where a foot consists of a stressed syllable and optionally one or more unstressed syllables. Some have said ([3]) that, based on measurements of spoken English, this dichotomy is an auditory illusion. It goes beyond the scope of this paper, but I believe that the reason Dauer and others have been unable to find stress timing is related to their definition of 'foot', which they normally define independent of the words being measured. The data to be reported below, both that found in the literature ([4],[5]) and collected for this paper, show that the stress-timing effect is found at least for isolated words.

LENTHENING IS RHYTHMIC

Let us suppose that every stressed syllable in English is associated with at least one beat. Unstressed syllables form, with the preceding stressed syllable, a single beat as well, being roughly equivalent to a trochaic foot in contemporary Metric Phonology (see, e.g. [6]). Let us also assume that the real time length of the beat can vary depending on such extralinguistic factors as speed of speech, but that the ratio of stressed to unstressed syllables will remain relatively constant under variation for tempo and other extralinguistic factors.

Let us suppose, further, that segments are mapped onto syllables following language specific implementations of universal principles, governed overall by something very much like the traditional sonority hierarchy. Thus, a full vowel will receive a single beat, but (for English) a coda consonant will not. Thus English will differ from, say, Japanese, where coda consonants do in fact receive beats.

If we assume that beats receive roughly the same amount of time, given a similar rate of speech, there should be rough isochrony in English among one and two syllable feet. Thus stead and steady should occupy roughly the same amount of time. [4], [7] investigated this with words like stick:sticky, sleep:sleepy, speed:speedy, shade:shady.

If it were the case that every foot received an identical amount of speech time (which is what we mean by assigning a beat to each foot), then we should expect that tight, tied, tie, and tidy should each receive the same amount of time. This is however, not what we find. Specifically, the stressed vowel lengths differ, and not in the way that one would expect. The vowel length in tight and tidy are (roughly) the same, and short, while those in tie and tide are also roughly the same, but much longer than in the preceding pair. Given these facts it is surprising that the voicing of the syllable-final consonant should be posited as the cause of the differences in vowel length: This can be seen in the following chart:

Short Vowel	Long Vowel
tight	tied
tidy	tie

We can explain the difference, however, if we assume that vowel length is determined by foot structure—if every foot gets an equal measure of time, tie and tidy should receive an equal measure. Since the latter word has two syllables, each must be much shorter than any single syllable by itself. Borrowing from musical principles, if we assume that each beat is worth a quarter note, tie would be assigned a quarter note, while tidy would be assigned two eighth notes. As a consequence, the /tai/ of the former should be much longer than the /tai/ of the latter. For example, in [7], Lehiste found the following average values for 'sleep', 'sleepy', 'speed' and 'speedy':

Table 1. Mean values of the nucleus /i/ for sleep, sleepy, speed, speedy, *expressed in ms.*

sleep	sleepy	speed	speedy
180.3	131.45	297.85	163.3

But now we must ask, why is the vowel in tight so short, and the vowel in tied so long? If we continue our assumption that we are dealing with rhythmic principles here, perhaps we can make the same assumption as with the preceding pair. Suppose that there is something special about voiced consonants specifically that they are extrametrical. This is an assumption that is made about all final consonants in English nouns (see, e.g [8]) for early discussion). However, in the Metrical phonology literature the extrametricality assumption is made solely in order to place stress on the correct syllable in words like cannon, whose final syllable must be light. In these cases the extrametricality is posited solely to make stress assignment rules either regular (in some cases) or simpler (in others, such as penultimate stress). I am here assuming that extrametricality is a real rhythmic

phenomenon, and that final voiced consonants do not 'count' for vowel length assignment, but that final voiceless consonants do.

The result of this set of assumptions is that, given the word tight, the entire syllable will be assigned a beat. Since final consonants take up real time, the vowel must shorten to permit the entire assemblage to occupy only a beat's worth of time. On the other hand, since final voiced consonants are extrametrical, only the vowel will be mapped onto the beat, and as a consequence, the vowels in tied and tie will be of roughly the same length.

Now, what justification, other than the fact that the results come out right, can we find for making voiced consonants extrametrical while voiceless ones are not? Overall I have no definitive answer. However, if we consider only stops, we can note that syllable final voiceless stops in English are normally accompanied by simultaneous glottal closure, while voiced stops are, of course, not. Thus it is often the case that voiced stops are released, while voiceless ones are not. It may be the case that this is somehow tied in to the timing relationships we have been discussing.

In any case, whatever the justification for assigning extrametricality to voiced obstruents, it ought to be possible to experimentally test this rhythmic account of VL in a number of ways. One would be simply to closely examine the differences between vowel allophones in final position and those before voiced vs. voiceless consonants. There are, of course, obvious cases like Canadian English where the prediction seems to be confirmed. Canadian English has an allophonic rule relating higher and lower nuclei of the diphthongs /ai/ and /au/, with the higher nucleus (normally close to [ə] occuring exclusively before syllable-final voiceless obstruents and the lower nuclei occurring elsewhere. The facts are sufficiently well known not to need rehearsal here, but it is the case that we find the raised vowels not only in classic cases such as *write* (vs. *ride*), but also in the shorter vowel contexts discussed above (such as *writer*). Exactly what happens with rider seems to be a matter of conjecture at this point, and the raising rule seems to be generalizing at this point to include not only voiceless stops but also /n/, which poses problems for any theory of phonology (including mine, incidentally) that believes that features rather than arbitrary classes of phonemes condition phonological rules. Other possible cases that bear investigation would include those dialects of English where only long allophones are diphthongs. For example, Northern Central US English seems to have monophthongal /e/ and /o/ in short contexts, with diphthongizing [ei] and [ou] only when either final or before voiced sounds.

In sum, while I have no definitive proof that word-final voiced consonants behave as if they were not located in the syllable they close, the length of the vowels preceding them indicates that they are not. As a consequence, we could also conclude that the supposedly 'un'natural rule of English voicing lengthening might be somewhat more natural than was previously thought.

ACKNOWLEDGEMENT

I am indebted to David Stampe for the genesis of this idea.

REFERENCES

[1] Klatt, D.H. (1979), "Synthesis by rule of segmental durations in English sentences." *Proceedings of the 9th international congress of Phonetic Sciences.* Copenhagen.

[2] Classé, A. (1939), *The rhythm of English prose.* Oxford: Basil Blackwell

[3] Dauer, R.M. (1983). "Stress-timing and syllable-timing re-analyzed." *Journal of Phonetics* , vol. 11, pp. 51-62.

de Jong, Kenneth. 1991. An articulatory study of consonant-induced vowel duration changes in English. Phonetica 48:1-17

[4] Lehiste, I. (1971), "Temporal organization of spoken language." in *Form and Substance: Phonetic and Linguistic papers presented to Eli Fischer-Jorgensen.* L.L. Hammerich, R. Jakobson and E. Zwirner, eds. Copenhagen: Akademisk Forlag.

[5] Fowler, C.; K. Munhall, E. Salzman, and S. Hawkins. (1983), "Acoustic and articulatory evidence for consonant-vowel interactions." *Journal of the Acoustic Society of America*, vol. 80 (suppl. 1): S96

[6] Hayes, B., (1995) *Metrical stress theory* : principles and case studies Chicago : University of Chicago Press,

[7]Lehiste, I. (1972), "The timing of utterances and linguistic boundaries," *Journal of the Acoustic Society of America*, vol. 51 pp. 2018-2024.

[8] Hogg, R.M., and C.B. McCully, (1987), *Metrical Stress Theory: A coursebook.* Cambridge: Cambridge University Press.

THE BALKAN SPRACHBUND IN THE LIGHT OF PHONETIC FEATURES

Irena Sawicka
Nicolas Copernicus University, Toruń, Poland

ABSTRACT

In the paper the division of the Balkan Sprachbund according to the phonetic features is shown. The main areas are: 1. the Eastern one, 2. the Central Balkanic, 3. the Mediterranean one.

INTRODUCTION

The Balkan Sprachbund is a linguistic community defined by morphosyntactic features. As far as phonetics is concerned it presents a completely different picture as that resulting from morphosyntax. In fact, we cannot speak anymore about the Balkan community. There are, however, several relatively compact areas characterized by a considerable similarity. Borders of these areas often extend beyond the territory of the Balkans.

A certain convergence characterizes: 1. the Eastern microregion, 2. the so-called Mediterranean microregion, and, the less compact 3. the Central Balkanic region.

THE EASTERN BALKANIC AREA

This area consists of the territory of Rumania and Bulgaria, especially their Eastern parts, and the North-Eastern Greece. Apart from Rumanian and Bulgarian it comprises other dialects occurring within its frames, such as Arumanian or Turkish. As in the case of other regions the borders of the Eastern area cannot be precisely defined; particular features have various extentions with central areas overlapping. This area has been considered the center of the phonetic Balkan Sprachbund [1]. In my opinion, it is rather an extention of the Eastern Slavic strip which presents the same phonotactic type. Its pronunciation is characterized by a number of vocalic, as well as consonantal, assimilations — it can be described as an accomodative type of phonotactics.

First of all, this is a kind of syllabic harmony, requiring an adjustment of segments in a syllable with regard to palatalization and/or labialization. However, the realization of this property varies from dialect to dialect. Besides, it regards either phonemics or only phonetics. This feature is connected with a rich consonantal system. Apart from that, reductions of vowels in unstressed positions appear in the Eastern area. Generally speaking, unstressed vowels become higher. In Northern Greek, in addition, high unstressed vowels are lost.

The presence of a centralized vowel, functioning as a separate phoneme, is usually considered a Balkanic feature. One or even two centralized vocalic phonemes occur in the languages of Eastern Balkanic area, however with the exception of Greek. On the other hand such a phoneme is also found in the Southern Albanian — a Western Balkanic region.

THE CENTRAL BALKANIC AREA

This is the less compact region, although it is characterized by the most typical Balkanic features. They are concentrated on the strip of land containing Macedonian, Albanian, Greek. However, particular features do not have identical extentions, some of them occur also in Bulgarian, Serbian or Italian. These features are: merger of the palatal affricates, among others resulting from the tendency eliminating palatalization from consonantal systems (this takes place in Albanian, Macedonian and Serbian); the loss of [x] (in the same dialects), a tendency towards proclisis of the clitical forms of personal pronouns and the most characteristic feature of the Balkans — a set of phenomena regarding clusters of a nasal sonorant and homorganic stop obstruent. These facts are linked by the tendency towards functional equivalence of these clusters and corresponding voiced stops. Phenomena contributing to this equivalence are: voicing of stops after nasal sonorants (Greek, Northern Albanian, Southern Italian), a very fused, monosegmental pronunciation of these clusters (Northern Albanian), simplification of the ND clusters to N (Albanian; in Southern Italian the change of ND into NN is found), or to D (Albanian, Northern Greek), appearance of unmotivated voiced stops after nasal sonorants, especially the change of *mr, ml* into *mbr, mbl* (Albanian), appearance of the unmotivated nasal sonorants before voiced stops (Albanian, Greek), the preservation of the motivated nasal sounds (reflexes of the Old Slavic nasal vowels) before voiced stops, whereas in other contexts nasality is lost (Southern Macedonian dialects). As a result of the tendency towards functional equivalence of the clusters in question and corresponding voiced stops, in some dialects ND and D can replace each other, and generally, options and hesitations in pronunciation of ND clusters are quite frequent in the Balkans.

The most unusual feature, connected with these clusters is that they can appear at the beginning of the word (Albanian, colloquial Greek, dialectal Italian) [2].

In my opinion, this very area should be regarded the center of the phonetic Balkan Sprachbund, and not the Eastern part of the Balkans.

THE MEDITERRANEAN AREA

The Balkan Sprachbund partially overlaps with another linguistic community characterized by certain convergence in the field of phonotactics. Languages and dialects located on the peninsulas of the Mediterranean Sea share, first of all, a similar syllable and word pattern with the last syllable open or closed by a single consonant [3]. Usually it is a sonant or [s]. Consonant clusters in Mediterranean languages have simple acoustic patterns; in final positions they occur quite exceptionally, mainly in loanwords.

Features concerning syllable patterns cut the Greek area into two parts — in the Northern Greek dialecs, as a result of vowel reductions, word final syllables can be closed.

Another Mediterranean feature is the restriction on intervocalic voiced stops. In the Romance languages this restriction is solved by fricativization — obligatory (Spanish, Catalan), optional (Portuguese), or it occurs only in colloquial and dialectal speech (Italian). Although fricativization is known in the history of Greek phonetics, today the restriction on intervocalic voiced stops is being solved by the shift of VDV into VNDV. Again this feature does not exist in Northern Greek where the ND clusters have been simplified into D.

These two phonotactic features oppose the Mediterranean languages to other European languages, including the Balkan languages.

The Mediterranean languages share with the Balkanic languages one property, namely in these languages sentences can begin with pronominal clitics. They also have similar question intonation contours [4].

REFERENCES

[1] Ivić, P. (1968), *Liens phonetiques entre les langues balkaniques,* Actes du Premier congres international des etudes balkaniques et sud-est europeennes, Sofia, p.133–144.

[2] Sawicka I. (1991), *The problem of the prenasalization of stops in Southern Slavic,* Studies in the phonetic typology of the Slavic languages, Warszawa: SOW, p. 113–125.

[3] Perlin, J., Sawicka, I. (1991), *Is there a Mediterranean phonotactic community?,* Studies in the phonetic typology of the Slavic languages, Warszawa: SOW, p. 51–63.

[4] Lehiste, I., Ivić, P. (1980), *The intonation of yes-or-no questions — a new Balkanism?,* Balkanistica VI, p. 45–53.

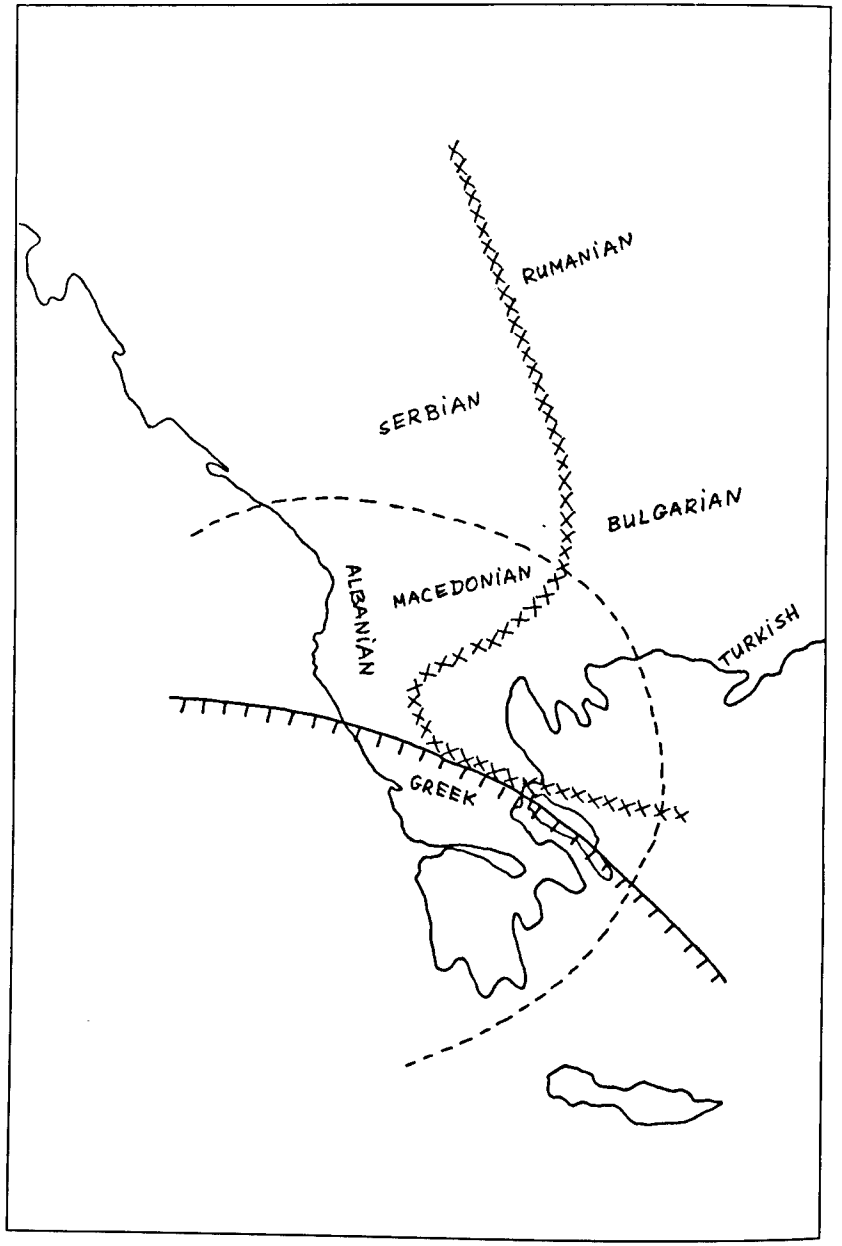


Figure 1. Approximate extensions of particular areas:

- – – – – *The Central Balkanic Area*
- xxxxxxx *The Eastern Balkanic Area*
- ⊤⊤⊤⊤⊤ *The Mediterranean Area*

ACOUSTIC CORRELATES OF VOWEL QUANTITY CONTRASTS IN AN ITALIAN DIALECT

A. Uguzzoni and M. G. Busà***
**Istituto di Glottologia , Bologna, Italy*
*** Dipartimento LLSM, Bologna, Italy*

ABSTRACT

This paper examines the issue of vowel quantity contrast in a northern Italian dialect in the light of a typology of so-called quantity languages. The results show that in this Italo-Romance variety the quantity contrast has the stressed vowel segment as its domain. The short vowels are less than half the duration of the long vowels, which indicates that this ratio is the main cue for the contrast. Some spectral variation is dependent on the long-short distinction.

INTRODUCTION

It is a widespread belief that Italo-Romance languages lack vowel quantity contrasts. Our previous studies, though limited to the northern Italian dialects of Emilia, led us to a different opinion. A preliminary phonological interpretation of the large vowel inventory found in a dialect of the Frignano area was proposed. In stressed syllables thirteen vowels are used distinctively: the nine long vowels /i:, y:, u:, e:, ø:, o:, ɛ:, ɔ:, a:/, and the four short vowels /e, ø, ɔ, a/. This analysis received later some instrumental support. A first research was focussed on the durational differences between the long segments /e:, ø:, ɔ:, a:/, and the short segments /e, ø, ɔ, a/ [1]. In a following research the spectral properties of all the thirteen phonemes of the system were examined, but the paper discussed only the formant frequency patterns of the nine long phonemes [2].

These early studies raised several questions which require further experimental investigation and discussion. Firstly, it seems important to establish whether and to what extent quantity contrasts exist in an Italo-Romance variety. Another issue concerns the position of this northern Italian dialect within a typology of so-called quantity languages. A further problem is to propose a hierarchy among the multiple factors serving as cues for the Frignanese speaker-listener.

To this purpose, the present study investigates the acoustic correlates of the quantity contrasts in the dialect, with particular regard to the temporal and spectral parameters. The first aim is to determine the magnitude of the duration of long and short stressed vowels and the possible durational differences in the consonant following the long/short stressed vowel, to establish whether the domain of the quantity contrast is the vowel segment or the VC sequence. A second objective is the analysis of the spectral differences between the long and corresponding short vowels, and their location in the vowel space of the dialect.

DATA AND METHODS

The speech material presented here includes both monosyllabic and disyllabic meaningful words stressed on the first syllable; the target vowels are /i:, y:, u:, e:, ø:, o:, ɛ:, ɔ:, a:, e, ø, ɔ, a/ and occur in the context of different prevocalic consonant and constant postvocalic consonant, i.e., an alveolar lateral /l/ of the common voiced approximant type. The choice of this corpus was due to the decision of using only meaningful words and reducing the number of variables to simplify the present analysis. The limitations on the distribution of certain vowels in the dialect makes it difficult to find minimal and/or quasi-minimal pairs containing all the vowels under study and forced us to choose the lateral consonant as a postvocalic context. On the other hand, the context which was kept constant was that of the postvocalic consonant, since the duration of the stressed vowel is commonly held to be more affected by the following than the preceding consonant.

The words were put in a carrier sentence and spoken two times by three male native speakers of the dialect. The recordings took place in the subjects' homes in reduced noise conditions, using a Uher professional tape recorder and monodirectional microphone. The data were analyzed on the 14 bit Sensimetrics SpeechStation system, sampled at 10kHZ. Measurements were taken of the stressed vowels and postvocalic consonants durations from both spectrograms and waveforms. For the formant values for F1, F2 and F3, LPC spectra with order 10 were computed in the middle of each vowel.

RESULTS

Durations

The measurements on vowel durations indicate that the speakers of the Frignanese speakers make a very clear distinction between the long vowels /e:, ø:, ɔ:, a:/, and the short vowels /e, ø, ɔ, a/. As can be seen from Table 1, the mean short vowel durations are less than half the mean long vowel durations in both CVCV and CVC structures. Vowels have shorter durations in disyllabic than in monosyllabic words, conforming to a common tendency in world languages [3]. Note that this shortening occurs to a similar extent for both short and long vowels, and therefore the V/V: ratio is unaffected.

Table 1. Mean durations in ms and V/V: ratio of the short and long vowel pairs.

	Short	Long	V/V:
CVCV	100.3	209.6	46.13
CVC	119.7	245.9	46.33

Differences in the mean values of the V/V: ratio by vowel type and word structure were found. These are reported in Table 2. In general, the V/V: ratios vary with degree of vowel height. It can be observed that in CVCV words the ratios are smaller for the mid high than for the mid low and low vowels and in CVC words the ratio for the low vowel is the largest. This is explainable if one considers the absolute vowel duration values in relation to the height dimension. As an example, in CVCV words, the vowel having the shortest intrinsic duration ranges from 78 ms (in /e/) to 220 ms (in /e:/), while the vowel with the longest intrinsic duration ranges from 122 ms (in /a/) to 225 ms (in /a:/), so that the V/V: ratio increases along the high-low dimension.

Table 2. Mean values of V/V: ratio by vowel type and word structure.

	e/e:	ø/ø:	ɔ/ɔ:	a/a:
CVCV	36.56	40.00	55.17	55.56
CVC	44.79	43.39	43.23	64.47

We looked for possible correlations between the absolute duration values of all the thirteen vowels and the quality differences along the high-low, front-back and round non-round dimensions. Systematic correlations were found only for the high-low dimension, exemplified in Table 3 only for disyllables. The vowel duration values, grouped in sets by degree of height, show that, for both long and short vowels, vowel duration increases from the high to low series. These data are in agreement with the well established phenomenon of intrinsic vowel duration [3, 4].

Table 3. Mean durations of all vowels, in sets, in CVCV structure.

Long		Short	
/i: y:u:/	201.8		
/e: ø:o:/	206.3	/e ø/	81.3
/ɛ :ɔ: /	211.4	/ɔ/	116.7
/a:/	225.2	/a/	122.0

The data of the single subjects show some interspeaker variability in the absolute vowel duration values. In short, there are greater temporal differences between our three subjects in their production of long than of short vowels. Though limited, these data seem to suggest that the short vowels display a relative stability with respect to the long vowels.

With regard to the duration of the consonant following the four pairs of long/short stressed vowels, our previous study [1], based on one subject, showed different duration values for the target consonant in the CVCV vs. the CVC structure: the C/C: ratio resulted negligeable in disyllables while it was

.77.75 in monosyllables, so that long vowels were followed by relatively short consonants, and short vowels by relatively long consonants. The present data confirm the observation for the consonants occurring in the CVCV structure, where the C/C: ratio is around .100. For the consonants in the CVC structure the C/C: ratios for the three subjects are very different: for one subject it is .81.29, while the other two subjects have larger ratios. In this case no generalization can be made.

Vowel quality

The analysis of the spectral properties of the Frignanese vowels has revealed a certain amount of variation in formant frequencies between the long and short vowels. The data relating to the mean differences in F1 and F2 are visualized in Fig. 1. It can be observed that, for F1 of all the vowels, the three subjects exhibit, though to a different extent, the same tendencies. For F2, subjects SG and RI show a similar pattern, while subject GB behaves in a very different way for the vowel pair /ø:, ø/, as can be seen from the direction of the bars in Fig. 1. The figure also illustrates how the spectral characteristics differ systematically from vowel to vowel. For F1, the subtraction value is positive for /a:, a/ and negative for the other pairs. This indicates that the durational distinction affects the F1 values of short vowels so that they are smaller in /a/ and larger in /e, ø, ɔ/ than in the corresponding long vowels. For F2, the variation from long to short is more remarkable in /e, ø, ɔ/ than in /a/: with the mentioned exception of GB there is a a decrease in frequency for the front vowels /e, ø/ and an increase for the back vowel /ɔ/. The effects of duration on the quality of the four vowel pairs can be seen also in Fig. 2.

For representing the data in the formant chart, the F1 and F2 values were converted into Mel using the formula given by Fant [5]. Fig. 2 shows the location of the four vowel pairs in the thirteen-vowel system of the dialect for two of the subjects. The quality variation due to duration is found on both the F1 and F2 axes. For the high-low dimension, a comparison of the short and long members of the pairs shows a lowering of the vowels /e, ø, ɔ/ and a raising of /a/. It is interesting to note that short /e/ lowers to such an extent that it approaches the quality area of long /ɛ:/. As concerns the front-back dimension, the short vowels /e, ɔ/ centralize with respect to their long counterparts; the shift in F2 for the pair /a:, a/ is minimal. Subject GB's divergent behavior regarding the vowel pair /ø:, ø/ can be observed also in the formant chart. While for SG short /ø/ has a smaller F2 value than long /ø:/, for GB the formants shift in the opposite direction.

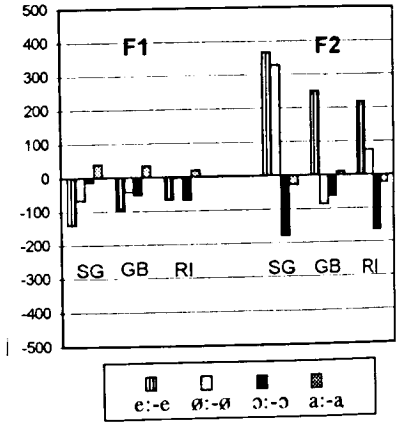


Figure 1 Differences in mean formant frequencies for the pairs of long/short vowels by the three subjects. Values of vowels in CVCV and CVC structures are averaged and expressed in Hz.

DISCUSSION

The main findings of our research on the acoustic correlates of the vowel quantity contrasts in the Frignanese dialect examined are the following. There is a considerable distinction between the long and the short vowels, which is constant and independent of speakers, vowel type and word structure. With regard to the postvocalic consonants, their duration appears to be unvaried in disyllables, but somewhat variable and speaker dependent in monosyllables. The vowel durational differences are accompanied by spectral differences, which vary in extent according to vowel type.

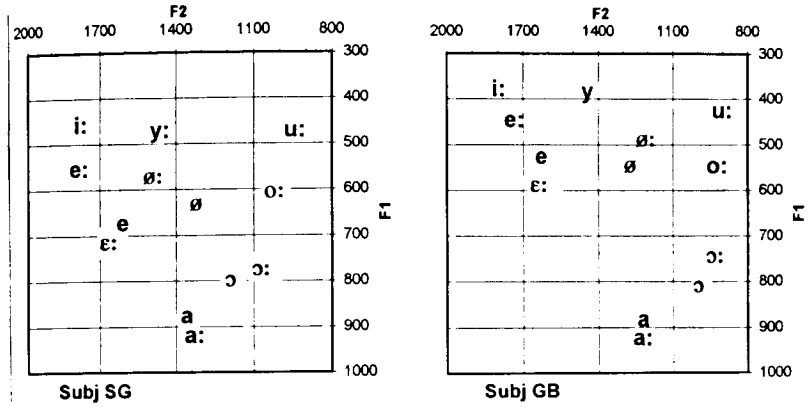


Figure 2. Mean F1 and F2 formant frequencies for the thirteen-vowel system (two subjects). Values of the vowels in CVCV and CVC structures are averaged and expressed in Mel.

On the basis of our production data, a tentative conclusion can be drawn as to which acoustic correlate is most important for the quantity contrast. It seems unquestionable that the crucial factor is the vowel duration distinction, the other two factors being additional. Qualitative vowel variation is dependent on the long/short distinction, like in many languages [4]. As for consonant duration, the fact that two different patterns were found for the two word structures poses a problem of interpretation. Of course we are aware that, to determine the actual hierarchy of importance among the phonetic factors serving as cues for the Frignanese speaker-listener, perceptual experiments are needed.

Finally, at a more general level of discussion, it is worth adding some concise remarks. As far as the data analyzed are representative of this dialect of the Frignano area, the results of the present study support our early hypothesis that this Italo-Romance variety is a quantity language. We attempt a brief definition of the Frignano-specific characteristics in the framework of the languages which make contrastive use of durational differences. Our analysis seems to suggest that the domain of the quantity contrast in the dialect is the vowel segment. There is a two-way quantity contrast which is found only between vowels occurring in stressed syllables. The durational distinction applies only to some phonemes of the thirteen-vowel system, that is, the four pairs /e:, ø:, ɔ:, a:/ and /e, ø, ɔ, a/. So, the qualities of the four short vowels are a subset of the qualities of the nine long vowels [6]. The short-to-long ratio is about .46 (i.e., 1:2.1) in both CVCV and CVC structures, which classifies this dialect among the languages having strong vowel quantity contrast.

REFERENCES

[1] Uguzzoni, A. (1971), "Quantità fonetica e quantità fonematica nell'area dialettale frignanese", *L'Italia Dialettale*, vol. 24, pp. 115-136.

[2] Uguzzoni, A. (1988), "Verso un'analisi parametrica dell vocalismo di un dialetto frignanese", *L'Italia Dialettale*, vol. 51, pp. 87-110.

[3] Elert, C. (1964), *Phonologic studies of quantity in Swedish*, Stockholm: Almqvist & Wiksell.

[4] Lehiste, I. (1970), *Suprasegmentals*, Cambridge MA: MIT Press.

[5] Fant, G. (1973), *Speech sounds and features*, Cambridge MA: MIT Press.

[6] Maddieson, I. (1984), *Patterns of sounds*, Cambridge: C.U.P. Press.

AN EXPERIMENTAL STUDY ON THE SEGMENTATION OF TAIWANESE SYLLABLES

H. Samuel Wang
National Tsing Hua University, Taiwan

ABSTRACT

This paper explores whether the syllable in Taiwanese, a language with about 800 distinct syllable types, is divided into phoneme-sized units. An experiment was conducted in which the subjects were asked to delete a segment of a syllable, to add a segment to a syllable, and to replace a segment of a syllable with another. It was found that the subjects who were less exposed to an alphabetic language were poor in performing the task, while those who were explicitly taught to use alphabet to spell the language performed excellently. It was thus concluded that the ability to segment the syllable into phoneme sized units arose through training in alphabetic orthography.

THE PHONEMIC ANALYSIS

Traditional linguistic analyses have proceeded with the assumption that the speech stream is analyzable into phoneme-sized units. However, such an assumption has been seriously questioned in recent years. Experiments in segmentation have shown that such ability is achieved mostly via training in orthography [1], [2].

Read et al.'s study [2] found that Mandarin speakers were not able to add or delete segments in the syllable if they were not trained in alphabetic writing. In this study we ask the same question of another Chinese language, viz. Taiwanese. Because Taiwanese is normally not used as a means of education, the native speakers' phonological knowledge is typically not influenced by orthography. On the other hand, most of the native speakers are educated in Mandarin and English. This gives us a chance to see whether their phonological knowledge is influence by Mandarin and/or English. In this study we are interested in finding out whether phoneme-sized segments are operating units in Taiwanese, and if they are, whether such ability is influenced by the alphabetic language, i.e. English.

THE EXPERIMENT

Three groups of subjects were tested in this experiment. The first group of subjects (N=21) were vocational high school students who worked at daytime and went to school at night. These students spoke a lot of Taiwanese at work, and because of the nature of their study, they did not have as much experience in English as the normal high school students. The language ability of this group of subjects is regarded as representative of ordinary Taiwanese speakers. The second group of subjects (N=20) comprised university students. Because they were going to the university, and because most of the university textbooks are in English, they were more exposed to English than the first group subjects. We are interested in seeing whether their ability to manipulate the Taiwanese syllables in terms of phoneme-sized segments is influenced by their training in English. The third group of subjects (N=19) were also university students, but these students had been self-taught to read, and sometimes to write, Taiwanese using Roman alphabet (the so-called Church Romanization). It is expected that because of their experience in using Roman letters for Taiwanese, the third group subjects will perform much better than the other two groups.

The Procedures

The subjects were first introduced to the concept of sound similarities through a popular Taiwanese folk song. The subjects were shown that a word can be converted to another word by adding, deleting or changing part of the sound of the word. Then the subjects were presented with pairs of words showing different phonetic relationships. Specifically, the following relationships were shown:

(1) Initial consonant deletion
e.g. tun33 → un33
(2) Initial consonant addition
e.g. an24 → tan24
(3) Final consonant deletion
e.g. kim55 → ki55
(4) Final consonant addition
e.g. ca51 → can51
(5) Initial consonant replacement
e.g. pi51 → ki51
(6) Vowel replacement
e.g. si55 → su55
(7) Final consonant replacement
e.g. t'an55 → t'am55

In each case, four examples were given to the subject to show that by adding (such as [t] in (2)), deleting ([t] in (1)) or replacing ([k] in (5)) a particular segment in the syllable, one can derive another syllable. Then in each category, two test items followed these examples to see whether the subjects could perform the task as exemplified. Among these two test items, one of the predicted correct answer would be a real word while the other would be a nonce word.

Results

Correct scores were tabulated for each subject. Mean percentages were calculated for each group (Group 1=6.89%, Group 2=49.17%, Group 3=71.67%). Mann-Whitney *U* tests were run comparing the ranks of correct scores among the groups. The results showed that the differences were all great. The percentages of correct answers in each item by groups are shown in Table 1.

A second set of the same tests was run. This time all answers containing expected changes were considered correct, with or without concomitant other changes; that is, when the subject made the correct change but at the same time made changes in other parts of the syllable, the answer was still considered correct. Gradations of group means were as above (Group 1=12.17%, Group 2=61.39%, Group 3=81%), and Mann-Whitney tests were significant. Table 2 shows the percentages of correct answers under such considerations.

Table 1. Percentages of correct answers (r and n in item numbers refer to real word and nonce word correct responses)

	Grp1	Grp2	Grp3
1r.	19.0	15.0	63.0
1n.	4.8	45.0	100.0
2r.	4.8	80.0	94.7
2n.	9.5	90.0	78.9
3r.	4.8	35.0	57.9
3n.	0.0	10.0	36.8
4r.	4.8	60.0	78.9
4n.	9.5	55.0	84.2
5r.[1]	9.5	45.0	68.4
5n.[1]	4.8	35.0	57.9
6r.[1]	4.8	55.0	84.2
6n.[1]	19.0	65.0	89.5
5r.[2]	0.0	70.0	78.9
5n.[2]	0.0	20.0	68.4
6r.[2]	4.8	20.0	31.6
6n.[2]	0.0	35.0	36.8
7r.	23.8	85.0	94.7
7n.	4.8	65.0	84.2

As can be seen from Table 1, Group 3 subjects did the best in every item except in (2n), and Group 2 did better than Group 1 in every item except (1r). No such exceptions were found in Table 2.

If we list these percentages by their magnitudes, we find more consistency between two items in a category in Table 2, which we do not find in Table 1. Still, no obvious patterns emerge. Category (3), which required the subjects to delete

[1]These replacement items are with CV syllables.
[2]These replacement items are with CVC syllables.

Table 2. Percentages of correct answers with concomitant changes (r and n in item numbers refer to real and nonce word correct responses)

	Grp1	Grp2	Grp3
1r	28.5	65.0	100.0
1n	33.4	75.0	100.0
2r.	4.8	90.0	94.7
2n.	14.3	90.0	94.7
3r.	9.6	40.0	63.2
3n.	0.0	10.0	47.4
4r.	4.8	75.0	84.2
4n.	9.5	55.0	84.2
5r.[1]	9.5	55.0	68.4
5n.[1]	4.8	40.0	57.9
6r.[1]	9.6	60.0	94.1
6n.[1]	19.0	65.0	89.5
5r.[2]	0.0	70.0	78.9
5n.[2]	0.0	45.0	84.2
6r.[2]	23.9	55.0	63.2
6n.[2]	0.0	45.0	63.2
7r.	33.3	85.0	100.0
7n.	14.3	75.0	100.0

the final consonants, scored the lowest. This is expected if we consider the final consonant as part of the rhyme. But Category (7), which required the subjects to replace the final consonants, was among the easiest. It is rather difficult to understand why a segment is hard to delete but easy to replace in the same position. It is not that segments are easier to replace than to delete, for Category (1) was to delete the initial consonant, and was among the easiest categories, while its replacement counterpart, Category (5), was rather hard.

In the case of real vs. nonce word responses, none of the comparisons (correct only or with concomitant changes) showed significant differences except for Group 2, where the comparison of correct answers with concomitant changes between real and nonce words showed significant difference (Wilcoxon z=2.79, p< .01).

DISCUSSION

From these results, it is clear that education played an important role in making the speakers aware of the segments. Group 1 subjects did not have as much contact with alphabetic language as Group 2 subjects, and as expected, their performances were much poorer than Group 2 subjects. Group 2 subjects and Group 3 subjects were all university students, and had similar education backgrounds except that Group 3 subjects explicitly learned to read Taiwanese in Roman letters. Presumably they were more able to manipulate speech sounds in terms of the orthography based on Roman alphabet. This is evident when we consider the answers given to item (1r) by some Group 3 subjects, where they were required to change [mĩ33] into [ĩ33]. The subjects made 37% expected answers with concomitant changes, and the concomitant changes were all [i33] instead of the expected [ĩ33]. One of the reasons for such response is the fact that in Church Romanization, the stimulus [mĩ] is written as *mi*, with the nasalization in the vowel left unspecified, while in other cases the nasal vowel is marked with a raised 'n', as in [tiⁿ55] 'sweet'. This is because in Taiwanese nasal consonants occur only before nasal vowels [3] When the initial consonant *m* is taken away, there is only *i* left in the orthography. In fact this is precisely what was pointed out to the experimenter by one of the subjects in Group 3.

As reported by the subjects after the experiment, many of the Group 2 and Group 3 subjects explicitly made use of the orthographic symbols. Some Group 2 subjects reported that they used the Mandarin Phonetic Alphabet (*Juyin Fuhao*), as did some Group 3 subjects. Some other Group 3 subjects reported that they used Church Romanization. It is interesting to find that there was not much evidence that Group 1 subjects made use of Mandarin Phonetic Alphabet, although they also learned it in school. One of the possibilities for such difference may be that Group 2 subjects use Taiwanese a lot more frequently than the other two groups. They are more used to operating the language without recourse to Mandarin, unlike the other two groups (especially Group 2). Perhaps primarily because of this, they were not inclined to use the Mandarin Phonetic Alphabet in doing the task.

As mentioned above, even if we admit concomitant changes, Group 1 subjects only achieved around 12%. With this percentage we can hardly say that the subjects could operate the language in terms of segments. The success with the other two groups was mostly based on their familiarity with alphabetic writing. Since the education backgrounds of the second and the third groups were similar except that the third group learned to read Taiwanese in Roman alphabet, the better performances by the third group subjects ought to have been due to their explicitly applying this knowledge in the segmentation tasks.

We have also noted above that final consonants are hard to delete but easy to replace, while initial consonants are easy to delete but hard to replace. These facts are hard to explain if these elements are individually considered. However, there is a possible explanation from the point of view of the distinguishability of the syllables. There are only three possible final consonants in non-entering tone syllables, but there are 14 possible consonants in the syllable initial position. Final consonants contribute a lot less than the initial consonants in distinguishing syllables.[4] Replacing a final consonant only means replacing a nasal consonant with another nasal consonant. The distance between these two syllables is rather small. But replacing an initial consonant can be a major operation, as replacing it would result in a rather different syllable type. On the other hand, deleting a final consonant would also result in a very different syllable type. It is a change from a closed syllable to an open syllable, and the syllable types are completely different. In the case of initial consonant deletion, the two syllables still rhyme after the deletion, and the contrast between the two syllables is not as great as that in initial consonant replacement, because in this case one syllable is with initial consonant while the other is without. What this indicates is that the subjects evaluated the effect the changes had on the whole syllable, rather than just changing part of the syllable.

We therefore conclude that explicit orthographic knowledge plays a significant role in realizing the segmental relationships among Taiwanese syllables. It seems that segmenting the Taiwanese syllable is a superfluous operation.

REFERENCES

[1] Morais, J., L. Cary, J. Alegria, and P. Bertelson (1979), "Does awareness of speech as a sequence of phones arise spontaneously?", *Cognition* 7: 323-331.

[2] Read, C., Y.F. Zhang, H.Y. Nie, and B.Q. Ding (1986), "The ability to manipulate speech sounds depends on knowing alphabetic writing." *Cognition* 24: 31-44.

[3] Wang, H.S. (1995), "Nasality as an autosegment in Taiwanese." To appear in F.F. Tsao and M.H. Tsai (eds.), *Proceedings of the First International Symposium on Languages in Taiwan*. Taipei: Crane.

[4] Wang, H.S., and B.L. Derwing (1993), "Is Taiwanese a 'body' language?" Paper read at the Canadian Linguistic Society Annual Meeting, Carleton University, Ottawa, May 1993.

PROSODIC VARIATION IN PARENTAL SPEECH IN SWEDISH

Anne-Christine Bredvad-Jensen
Department of Linguistics, Lund University, Sweden

ABSTRACT

The purpose of this study is to describe the prosodic adjustments made when going from child-directed speech to adult-directed speech in Swedish. 12 parent–child dyads are studied with children aged between one and four. Fo, amplitude and duration will be analysed in a sentence perspective and discussed in relation to generality for the adult and to function for the child's language acquisition and learning.

GENERAL PRESENTATION

Adjustment in speech directed to children up to three years of age has been considered to be a special speech register called BT (Baby Talk), [1] Snow & Ferguson 1977, or CA (child-adjusted communication), [2] Junefelt 1987. This paper will highlight the following questions: To what extent is child directed speech adjusted in the prosodic domain? Will the same physical adjustments occur in speech irrespective of the parent's speaking style and the child's age? Could there be a possible interaction between the child's acquisition and the parent's prosody?

MATERIAL

Parental speech directed to children between one and four years old was collected through bookreading which turned out to be a good combination of eliciting both spontaneous and "semi-spontaneous" speech. The parents did not only read the text but used the text material (as well as the illustrations) to engage the child in conversation. For comparison, adult-directed material was collected for each parent in adult–adult dyads. In order to facilitate the analysis of Fo, which in Swedish is used to signal sentence intonation, phrase-, sentence-, and word accent as well as boundary phenomena carefully prepared texts were used, (see [3] Bredvad-Jensen 1991).

RESULTS

Data show that all speakers adjust irrespective of their intention /not/ to talk Baby Talk to their child. All three parameters (Fo, amplitude and duration) are affected, but to different extents for the different speakers. More specific results will be presented at the congress. Individual speaking styles are preserved within the child-adjusted register which also varies according to the child's age. Typical features of child-adjustment are a) defocusing which seems to be used in order to give extra prominence to the sentence accented word, b) an added phrase accent even in very short sentences and c) pauses which are used more frequently than in adult-directed speech.

DISCUSSION

A developmental line can be seen in parental speech starting with the use of special tonal contours which are used to attract the baby's attention already during the preverbal period, (see for example [4] Bruner 1983). Once the child's acquisition has started, the role of child-adjustment is more complex; it will still serve as an attention-getting device, and it may at the same time facilitate the child's understanding and learning. Functional aspects of the physical parameters used in child-adjusted speech are threefold: pedagogical, affective and communicative, (see [5] Junefelt 1987). These aspects may of course characterize any dialogue, but their dominance in the child-adjusted speech is noticeable. As the child gradually will grow in verbal and communicative skill, the need for the parent to catch the child's attention will decline to a more adultlike level. This will result in a gradual decline in the use of the physical parameters. The question may remain as the last child-adjusted speech device which is interesting as the question itself is a verbal attention-getter.

[6] Bolinger 1978 pointed out that questions may be interrogative to different degrees, strong or weak questions (my terms) depending on the interrogative load. It is argued here that this appproach will also hold for child-adjusted attention-getters which then may be strong or weak to different degrees, depending on the attention-getting load. For both child-adjusted speech and for questions in general the same physical parameters may be used in similar ways. The dividing line here can be seen between the more "neutral" questions and the more emotionally loaded child-adjusted speech. The important role of affection in mother–child interaction has been emphasized by, among others, [7] Trevarthen 1988.

Even if child-adjustment is regarded as a special speech register with unique qualities it is at the same time part of man's universal capability of attracting somebody's attention and it will be performed with much the same physical means as in other situations. This might then be one explanation for the gradual decline from child-adjustment which can be seen in this study.

REFERENCES

[1] Snow, C. & C. Ferguson, (eds) (1977), *Talking to children, language input and acquisition.* Cambridge university press, Cambridge.

[2] [5] Junefelt, K. (1987), *Blindness and child-adjusted communication.* Stockholms universitet, institutionen för nordiska språk, (MINS 25), Stockholm.

[3] Bredvad-Jensen, A-C. (1991), "Remarks on question intonation in child-directed speech in Swedish", *Colloquium paedolinguisticum lundensis 1989,* paper no 6, Child language research institute, Department of linguistics, Lund university, Lund, pp. 33-42.

[4] Bruner, J. (1983), *Child's talk. Learning to use language.* W. W.Norton & Company Inc, New York, pp. 71-73.

[6] Bolinger, D. (1978), "Intonation across languages", *Universals of human language, vol 2, Phonology,* ed J. Greenberg, Stanford university press, Stanford, California, pp. 471-524.

[7] Trevarthen, C. (1988), "Infants trying to talk: How a child invites communication from the human world", *Children's creative communication,* ed R. Söderbergh, Lund university press, Lund, pp. 9-31

CHARACTERIZING THE ADULT TARGET: ACOUSTIC STUDIES OF SWEDISH AND AMERICAN ENGLISH /t/ AND /p/.

E. H. Buder[a], K. Williams[b], and C. Stoel-Gammon[a]
[a]University of Washington, Seattle, WA, USA
[b]Stockholm University, Stockholm, Sweden

ABSTRACT

Work by our group has investigated phonetic development of language-specific segments, including study of differences in place of /t/ articulation (Swedish: dental, American English: alveolar). We have reported acoustic measures showing language-specific spectral shapes for these bursts in both adults and 30-month-old children [1]. In this paper we examine factors that support the interpretation that these differences are indeed due to place.

INTRODUCTION

A recent study [1] revealed perceptual and acoustic differences between Swedish (S) dental and American English (AE) alveolar /t/ bursts. Listeners were able to categorize 15 ms burst portions of word-initial /t/s in both adults and 30-month-old children as dental and alveolar. Acoustic measurements indicated that spectral diffuseness (measured as std. deviation in Hz of the burst spectrum (SD)) and burst intensity (measured in dB difference from the following vowel) were significantly shorter across languages. Bursts were more diffuse and less intense in S than in AE. Shorter VOTs were also typical of Swedish /t/-initial tokens in both adults and children.

This paper reports similar differences in S and AE adults' word initial /p/ bursts from the two languages, suggesting that some language-specific acoustic features are shared among the stop consonants in our data. The central question guiding the present investigation is whether the spectral shape measures we have applied to /t/ bursts are uniquely associated with the alveolar/dental place distinction. Before continuing to investigate this topic in development, we need to characterize the adult targets using measures that uniquely capture this distinction. We therefore present data addressing two factors that may have affected earlier results: 1) that some of the differences in spectral shape may be due to differences in recording (e.g. equipment and room acoustic differences), and 2) that some of the spectral shape differences in both /t/ and /p/ are related to lower intensities and shorter VOTs of Swedish word-initial bursts.

RECORDING EFFECTS ON SPECTRAL SHAPE

Spectral shape measures of stop bursts based on a "moments" analysis of Fourier spectra have been of central interest in our work, following earlier work on the technique [2], [3]. Briefly, the spectrum can be characterized by its mean (M) and std. deviation (SD) in Hz and also by the higher moments-based dimensionless coefficients of skewness and kurtosis. Stoel-Gammon et al. [1] reported differences in all these measures when comparing adults' and children's /t/ bursts in S and AE, but acknowledged that some differences may have been due to recording effects such as room acoustics, microphone types, and the different standards of videotape recording media used (PAL in S, and NTSC in AE). Informal calibration efforts led us to suspect artifactual influences on the absolute validity of spectral M and the higher moments, but the effects in SD seemed too large to be artifactual. Further acoustic investigations of /p/ bursts revealed that SD differences were in the same direction as for the /t/ bursts.

In order to better calibrate recording effects, synthetic burst tokens were developed using filtered and dynamically shaped white noise to create 15 ms transients centering at four frequencies (1, 2, 3, and 4 kHz) and with two different bandwidths to emulate the alveolar/dental contrast in diffuseness. These burst tokens were recorded by playback over identical versions of Kay CSL software and the same speaker (JBL Pro-III) in both recording environments, with the same equipment used to record the actual speech data under investigation. Analysis of these data then proceeded using the same methods as Stoel-Gammon et al. Results indicated small yet systematic differences between S and AE recording environments in spectral M and SD, and larger differences in skewness and kurtosis. Because M and SD are the most powerful and interpretable in our speech data we focused on calibrations of these measures (see [3] for further discussion of difficulties with interpretation and statistical analysis of the higher moments). See Table 1 for adjustment values obtained from the calibration data.

Table 1. Adjustments to spectral mean and SD due to recording differences.

	Spec. M	Spec. SD
S-AE difference	+ 143 Hz	+ 73 Hz

Adjusting our adult /t/ and /p/ burst measures accordingly, the calibrated data were used to measure and statistically analyze (by *t*-tests) the language differences reported in Table 2. As can be seen, VOT and Burst intensity are significantly different across languages for both /t/ and /p/. However, the cross-language differences between spectral shape measures change when the measures are calibrated: an apparently non-significant difference between /t/ burst spectral Ms becomes significant, and a nearly significant difference between /p/ burst spectral Ms becomes non-significant. Regarding spectral SDs, the significant language difference in /t/s remains significant, and a significant difference in /p/ SDs is reduced, but still significant, when the measures are calibrated. The next section of this report investigates this further by assessing relations among VOT, burst intensity, and calibrated spectral shape measures.

VOT, BURST INTENSITY, AND SPECTRAL SHAPE MEASURES

It is possible that some degree of spectral shaping is related not to place of articulation but instead to burst intensity and VOT. In terms of intensity, the turbulence noise of a /p/ burst may become lower in central frequency and more compact when the stop is released with greater pressure as it appears to be in AE. In terms of VOT, a release burst may become higher in frequency and more compact when the longer VOT of AE yields a stop release that is followed by more aspiration. These possibilities seem supported by regression analyses investigating the effects of VOT and intensity on our spectral shape data. Burst intensity is a significant predictor of SD in S /p/s ($p<0.001$) and in AE /p/s ($p<0.05$), lower SDs correlating with higher intensities. Higher intensity is also a significant predictor of lower spectral means in S /p/s ($p<0.01$) and in AE /p/s ($p<0.001$). In AE /t/s, longer VOT is correlated with higher spectral means ($p<0.05$) and with lower spectral SDs ($p<0.05$).

Table 2. Acoustic measures of Swedish (S) and American English (AE) adult /t/ and /p/ bursts, with t-test comparisons across languages.

	S average	AE average	*t*	*p*
/t/				
VOT, ms	49	74	7.163	<.001
Intensity, dB below vowel	17.7	12.8	-6.376	<.001
Uncalibrated, kHz				
Burst Mean	5.158	5.501	1.793	n.s.
Burst SD	2.127	1.194	-9.477	<0.001
Calibrated, kHz				
Burst Mean	5.015	5.501	2.555	<0.05
Burst SD	2.054	1.149	-8.747	<0.001
/p/				
VOT, ms	41	66	7.061	<0.001
Intensity, dB below vowel	20.5	18.1	2.287	<0.05
Uncalibrated, kHz				
Burst Mean	2.957	2.604	-1.833	<0.10
Burst SD	2.015	1.689	-3.794	<0.001
Calibrated, kHz				
Burst Mean	2.814	2.604	-1.090	n.s.
Burst SD	1.942	1.689	-2.945	<.01

To investigate the effects among all these variables in a multiple regression framework, a logistic model can be used [4]. This type of regression model must be used when the dependent variable is categorical; here the continuous acoustic variables of VOT, intensity, burst spectral M and burst spectral SD can be entered as continuous predictors of the categorical language variable (S/AE). The logistic regression model is also appropriate because the model assesses the percentage of observations that are successfully classified according to language. By comparing the successful classification of a model incorporating VOT and intensity alone with the improvement in classification of a model that also incorporates the calibrated spectral shape measures, we see the extent to which spectral shape measures uniquely improve language discrimination above and beyond VOT and intensity alone. Table 3 lists these models and their classification success in the /t/ and /p/ bursts of our adult dataset. In the /p/ bursts, the marginal increment of classification success in the model incorporating spectral shape measures indicates that language differences in spectral shape do not contribute much additional predictiveness. In the /t/ bursts however, the addition of spectral shape measures contributes predictiveness that clearly goes beyond VOT and intensity. This result is consistent with the hypothesis that our spectral shape measures of /t/ bursts correlate with the alveolar/dental place distinction, and helps to explain the spectral shape differences in /p/ bursts as epiphenomenal to VOT and intensity differences.

Table 3. Language classification success percentages for logistic regression models with and without calibrated spectral shape measures for S and AE /t/ and /p/ bursts.

/t/	Classification Success
VOT and intensity alone	75%
VOT, intensity, burst mean and SD	85%
/p/	
VOT and intensity alone	71%
VOT, intensity, burst mean and SD	73%

CONCLUSIONS

After 1) calibrating spectral shape measures for effects of recording environment and 2) demonstrating effects of VOT and burst intensity on spectral shape, we conclude that burst spectral M and SD can be used as measures for language specific aspects of /t/, presumably correlating with place of articulation. Based on these demonstrations with adult speech samples, our future research will continue to use spectral shape measures to examine the development of place of articulation in children's speech.

REFERENCES

[1] Stoel-Gammon, C., Williams, K., and Buder, E.H. (1994). Cross-language differences in phonological acquisition: Swedish and American /t/. *Phonetica*, vol. 51, pp. 146-158.

[2] Forrest, K., Weismer, G., Milenkovic, P., and Dougall, R.N. (1988). Statistical analysis of word-initial voiceless obstruents: preliminary data. *Journal of the Acoustical Society of America*, vol. 84, pp. 115-123.

[3] Buder, E.H., Kent, R.D., Kent, J.F., Milenkovic, P., and Workinger, M. (in press). FORMOFFA: an automated formant-moment-fundamental frequency-amplitude analysis of normal and disordered speech. *Clinical Linguistics and Phonetics*.

[4] Steinberg, Dan and Colla, P. (1991). *LOGIT: A supplementary module for SYSTAT*. Evanston, IL: SYSTAT, Inc.

ASSIMILATION OF VOICE IN SPEECH DEVELOPMENT

Cecile T.L. Kuijpers
Max-Planck-Institut für Psycholinguistik, Nijmegen, The Netherlands

ABSTRACT

In Dutch, assimilation of voice is a common phenomenon. In adult speech many factors influence the assimilation process. In the present study we report on assimilation of voice in children. The data show that speech development is a factor which strongly influences the assimilation process. The data are interpreted within a phonetic and a phonological framework.

INTROCUCTION

Assimilation of voice in Dutch has been investigated in several phonetic and phonological studies. With respect to phonology, the assimilation process can be described by an ordered set of rules (see [1]). Summarizing, it may be said that, in a two-obstruent cluster, the first phonemically voiceless obstruent becomes voiced if followed by a voiced stop (e.g. 'klapdeur' /pd/->/bd/, and 'stofdoek' /fd/->/vd/). This process is called *regressive* assimilation. When the second phonemically voiced fricative becomes voiceless if preceded by a voiceless obstruent (e.g. 'potvis' /tv/->/tf/) it is called *progressive* assimilation. Zonneveld [2] argues that progressive assimilation is fed by the rule of final devoicing (e.g.'rondvaart /tv/ -> /tf/) and that the voicelessness of the cluster could be generalized lexically in Dutch.

From a phonetic point of view the assimilation process is characterized by an overlap of articulatory gestures; an inherent feature in a sound segment is altered under the influence of a neighbouring segment, and, since this accomodation occurs systematically, it is incorporated in the phonology of the language. In a phonetic empirical study by Slis [3] assimilation of voice is described in terms of laryngeal adaptation in which several articulatory parameters are involved. Slis assumes that the assimilation process is a form of coarticulation based on the mechanical and aerodynamic properties of the vocal cords.

Many factors influence the assimilation process [4], such as 'sex of the speaker', 'speaking rate', 'phonological composition of the cluster', 'word stress', and 'linguistic boundary'. In a recent study by Menert [5] it is argued that voice assimilation is a gradual and optional process; a discrete classification on the basis of articulatory features is arbitrary. Thefore, she assumes that, primarily, assimilation of voice is part of the grammar of the language user and originates in the internalized phonological rules.

So far, all findings on assimilation of voice in speech production were deduced from adult speech utterances. Actually, assimilation of voice in child speech has never been examined, although it will be most instructive with respect to both the phonetic and phonological explanation of voice assimilation. The research questions in this production study are: Do young children assimilate to the same extent as adult speakers? How are progressive and regressive assimilation distributed across age groups? Do children also assimilate more frequently within compound words than across a word boundary (influence of a Inguistic boundary)?

METHOD

Participants

Three age groups participated in the experiment; 6-year-olds (mean age 6;4), 12-year-olds, (mean age 12;2), and adults (between 25-30). In each group there were three male and three female speakers. On the basis of a pilot experiment we excluded a group of four-year-old children because of problems with both the word material and the recordings.

Material

In total 21 items were selected; 8 two-syllabic compound words, and 13 two-word items (all adjective-noun combinations). In each item the first syllable was stressed. The items contained heteroganic stop-stop, fricative-stop, stop-fricative, and fricative-fricative (C1C2) clusters (see Table 1). All words were known by 6-year-olds, and they were represented by pictures.

Table 1. Illustration of the words with word-medial two-obstruent clusters (for instance, (stropdas= 'tie', knapdier= 'clever animal')

cluster	compound	two-word item
stop-stop	stropdas	knap dier
fric-stop	leesboek	zes ballen
stop-fric	broekzak	leuk vest
fric-fric	grasveld	zes vogels

Task

In all groups the words were elicited by picture cards in a sentence completion task, e.g. 'A book for reading is a ... reading book' (leesboek). In this way the segmental characteristics of the item would not be disturbed. In the two-word items the adjective was always stressed by asking the opposite qualification, e.g. 'These are not five balls but ... six balls' (zes ballen).

Recordings

Recordings were made of both the microphone signal and the electrolaryngograph signal on separate channels of an audiotape (recorder Revox A77). Exact timing of the glottal activity was registered simultaneously with the output of the microphone. All subjects were recorded twice and both recordings were used for further analysis. Besides, no problems occurred during the recordings.

Measurements

Both signals were stored on a microVAX II computer. For actual measurements both visual and auditory information were available. For the adult speakers we maintained the standard criteria of Slis (see [3], [6]). Since the laryngeal configuration of children differs from adults we normalized the adult criteria for the children's realizations. The normalization was done on the basis of voice continuation (voice tail) in intervocalic voiceless stops. In child speech we considered the first obtruent of the cluster to be voiced if the voice tail exceeds 31.5 ms. For the adults the standard 50 ms criterion was used. All other parameters were identical in the children and the adults.

RESULTS

Phonological composition

With respect to the clusters with final fricative (C2) both children and adults display 100% progressive assimilation. This corresponds to the phonological rules [1]. With respect to the clusters with final stop, voice assimilation is either absent, progressive, or regressive. The latter group of items are analysed further for the different age groups.

Age groups

The assimilation categories are distributed differently across age groups (chi^2= 36.54; $p<0.001$; data pooled across words). Both groups of children pronounced the C1C2 clusters with predominantly *progressive* assimilation (see also Figure 1). The adults show mostly *regressive* assimilation. Assimilation does not differ significantly between younger and older children, but it does differ between both groups of children and adults (chi^2= 29.45; $p<0.001$) . The number of words without assimilation do not

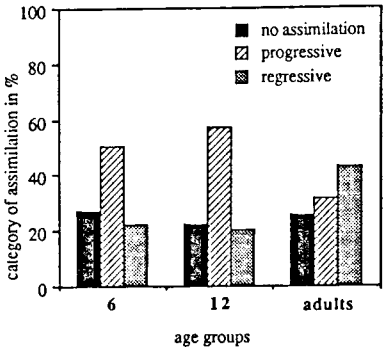


Figure 1. Frequency of occurrence of assimilation (in %) in the age groups.

deviate largely between groups. In none of the age groups a significant difference was found between male and female speakers.

Stop-stop and fricative-stop

In a log-linear analysis possible interactions were tested between the variables 'age', 'manner of articulation of C1' (i.e. stop/fricative), and 'assimilation category'. A significant difference was found for the interactions 'age' by 'category' and 'age' by 'C1' (chi^2=54.52; p<0.001 for fric-stop, and chi^2=10.81; p<0.01 for stop-stop). In the fricative-stop clusters regressive assimilation increases systematically with age, while in the stop-stop clusters a less regular pattern is present.

Linguistic boundary

So far, we did not take into account the linguistic boundary. In one part of the corpus the two-obstruent clusters were situated within a compound word, in the other part they were situated across a word boundary. In the (adult) literature it was found that regressive assimilation occurs more frequently in compound words than across a word boundary. In the present study the adults maniested the same tendency; a significant difference was found for the interaction 'assimilation category' by 'linguistic boundary' (chi^2=28.89; p<0.001). No significant difference was found for the interaction 'age' by 'linguistic boundary'. However, a significant difference was found between children (age 6 plus 12) and adults for the clusters within words (chi^2=18.10; p<0.001) and across a word boundary (chi^2=18.70; p<0.001). The data are indicated in Figure 2. Briefly, children use predominantly progressive assimilation irrespective of the linguistic boundary.

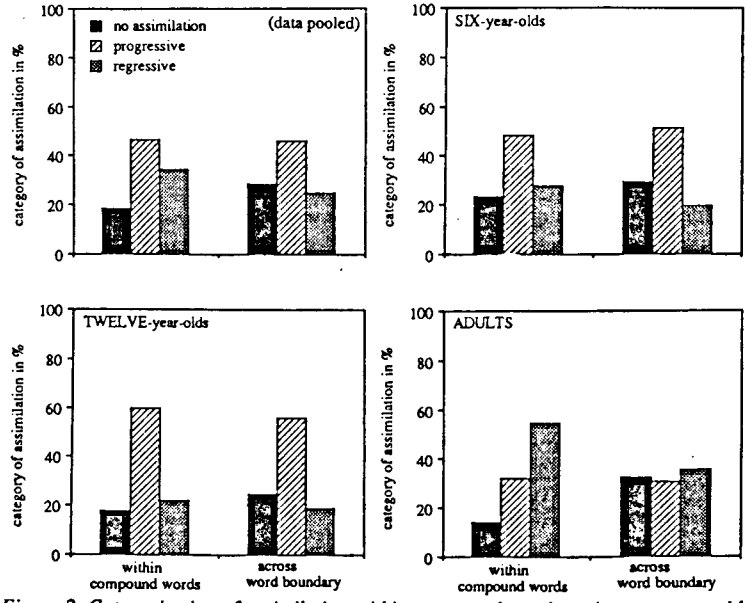


Figure 2. Categorization of assimilation within compound words and across a word boundary for data pooled across age groups, and for the four age groups.

DISCUSSION

We have seen that two-obstruent clusters with a rightmost phonemically voiced fricative, are always assimilated progressively in all age groups. The voiced character of the fricative (closing gesture of the glottis) is overruled by the voiceless character of the preceding obstruent (opening gesture of the glottis). Devoicing of the obstruent-fricative clusters seems to be internalized in the phonological system of the speaker. The data of the young children affirm this claim.

With respect to the obstruent-stop clusters the results lead to the following answers to the research questions. The 6- and 12-year-old children do not assimilate less than adults, but they assimilate differently. While adults show more regressive than progressive assimilation (43% *vs.* 31%), 6- and 12-year-old children display more progressive than regressive assimilation (50% (6) and 57% (12) *vs.* 23% (6) and 20% (12). Furthermore, voice assimilation in children is not influenced by the linguistic context. In adult speakers regressive assimilation occurs more often within words than across a word boundary, whereas in children progressive assimilation dominates irrespective of the strength of the linguistic boundary.

From a phonetic point of view, the predominance of progressive assimilation in children can be explained by the fact that children have more difficulty in maintaining voicing during an obstruction than adults. The articulation-based persistence of voicelessness is easier, and will also be stronger than the influence of the linguistic boundary.

From a phonological point of view, the children's data can be interpreted within a phonological framework. First, they have a general rule deleting voice, fed by the final devoicing. The learning process consists of recognizing that rightmost fricatives cause a voiceless cluster but stops don't. Later on, the interaction between voice assimilation and strenght of the linguistic boundary will be incorporated into their phonological system.

ACKNOWLEDGEMENTS

This reseach was part of my Ph.D study financially supported by the University of Amsterdam and carried out at the Institute of Phonetic Sciences Amsterdam. I would like to thank L. Pols and F. Koopmans-van Beinum for their many helpful suggestions and discussions on the dissertation.

REFERENCES

[1] Booij, G.E. (1981). *Generatieve fonologie van het Nederlands.* Utrecht-Antwerpen: Spectrum.

[2] Zonneveld, W. (1983). "Lexical and phonological properties of Dutch voicing assimilation". In: M. van den Broecke, V. van Heuven & W. Zonneveld (Eds.). *Sound Structures: Studies for Antonie Cohen.* Dordrecht: Foris Publications, pp. 297-312.

[3] Slis, I.H. (1981). "The effect of speaking rate on assimilation of voice". *Proceedings of the Institute of Phonetics of Nijmegen* 5, pp. 160-176.

[4] Slis, I.H. (1986). "Assimilation of voice in Dutch as a function of stress, word boundaries, and sex of the speaker and listener". *Journal of Phonetics* **14**, 311-326.

[5] Menert, L. (1994) *Experiments on voice assimilation in Dutch: Prosodic structures and tempo.* Utrecht: OTS dissertation series.

[6] Kuijpers, C.T.L. (1993). *Temporal coordination in speech development: A study on voicing contrast and assimilation of voice.* Unpublished Ph.D. Thesis, University of Amsterdam, The Netherlands.

YOUNG INFANT'S PERCEPTION OF SEGMENTAL AND SUPRASEGMENTAL INFORMATION: PRELIMINARY RESULTS

Francisco Lacerda, Ulla Sundberg, Christin Andersson and Åsa Rex
Inst. of Linguistics, Stockholm University, S - 106 91 Stockholm, Sweden

ABSTRACT

This paper is a partial progress report from an experimental study addressing the question of whether "prosodic markers" help infants to discriminate between naturally presented word contrasts. The speech materials were carrier sentences in which target words are presented for discrimination. The prosodic characteristics of the carrier sentence were varied to enable comparisons between contrasts when the target word receives the main sentence stress and contrasts when the main stress is shifted to a non-changing word. The infants were randomly assigned to one of two groups — the infant-directed speech group or adult-directed speech group. At this point, 10 infants, with an average age of 8.5 months, were tested with the head-turn technique. The current results indicate that, contrary to what might be expected on the basis of infants' preference for motherese, adult-directed speech leads to better discrimination between the target words than infant-directed speech.

INTRODUCTION

Adults talking to young infants tend to use a type of speech generally referred to as motherese or infant-directed speech. This type of speech can be generally characterised as containing exaggerated features in relation to an adult-to-adult reference speech — higher F_0, larger F_0 excursions and lower speech tempo than adult-directed speech [5]. Because infants attend preferentially to infant-directed speech [1, 3, 12], it is possible that the exaggerated prosodic features of the motherese may also assist the infant in structuring the linguistic information of their ambient language. In addition it has been reported that young infants are sensitive to the correct placement of prosodic juncture markers [8]. Thus, if infants pay preferential attention to motherese that convey rather explicit prosodic information [4], if they can correctly use prosodic information to detect word boundaries, if they are capable of detecting virtually all phonetic contrasts that they have been tested with [6] and if they also are sensitive to the phonotatic patterns of their native language by 9 months of age [9], then infant-directed speech may assist the infant in extracting relevant linguistic information from its ambient language [2, 7].

To test this hypothesis we assessed the infants' capacity to discriminate target words embedded in carrier sentences. Our hypothesis was that the target words were presented in focal position in sentences produced as infant-directed speech should be easier to discriminate than when the same target words occurred in non-focal position or were presented in adult-directed speech sentences.

METHOD

Stimuli

The stimuli were two sets of natural sentences produced by an adult female native speaker of Swedish — one set produced as infant-directed speech and the other as adult-directed speech. The sentences can be described as a presentation sentence (a carrier sentence, "Det är små _____ där", "There are small _____ there") in which the target word and/or the word in focus is changed. Table 1 shows how the sentences contrasted within one of the set of infant-directed speech.

The stimuli were produced by editing the target-words in the appropriate carrier sentences and calibrated with adult listeners.[1] The sentences were organised in one set of infant-directed speech and another of adult-directed speech sentences.

Table 1. Reference and contrast sentences used in the experiments. The word in focus is underlined.

Reference sentence	Word contrasts	Prosodic and word contrasts
Det är små myror där	minor, manar	Det är... minor, manar, myror

Subjects

The present results were obtained from 10 infants living in monolingual Swedish language environments. The infants had an average age of 8.5 months, with a standard deviation of 0.6 months. The subjects were randomly assigned to the "infant-directed speech" group or the "adult-directed speech" group.

Procedure

The infants' ability to discriminate between the reference sentence and its variants was tested with the head-turn paradigm [10]. The test procedure consisted of three phases — conditioning, criterion and test phase. In the conditioning phase the infants were trained to produce head-turns in response to a large contrast between sentences involving differences in direction, focus and maximal change in the target word ("myror"/"manar"). All trials in this phase were change trials. After three consecutive correct responses the infants proceeded to the criterion phase, including both change and no-change (control) trials. In this phase the infants were also requested to generalise from the single contrast to two less clear contrasts. The infants had to produce 7 correct responses within 8 consecutive trials to proceed to the test phase in which only contrasts within a speech direction were used.

Discrimination measures

The discrimination metric used here is the unbiased d' measure. Because d' helds infinite values if the percentage of hits or false-alarms is either 0 or 100, these bottom and ceiling values were adjusted to 0.1 and to 99.9 before performing the d' computations. The d' scores were submitted to a one-way analysis of variance in which the d' obtained for each of the contrasts involving word change, with or without change in the sentence focus were treated as repeated measures. The factor was the intended direction of the speech — adult-directed vs. infant-directed.

RESULTS

The average scores obtained for the adult-directed and for the infant-directed speech are displayed in figure 1. Discrimination scores were poorer for infant-directed speech than for adult-directed speech in four of the five word contrasts. Only the *myror/minor* contrast in non-focal position had higher average scores for infant-directed speech than for adult-directed speech.

[1] The details of this procedure and results from the adult perception tests will be published elswhere [11].

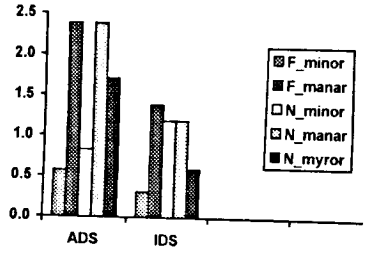


Figure 1. Average discrimination scores (d') for adult-directed speech and for infant-directed speech.

An analysis of variance in which the five discrimination scores from the word contrasts were modelled as repeated measures with direction of the speech as a factor, revealed no significant differences between the two types of speech ($F(1,8)=1.620$, $p < 0.239$). The within subjects' measures indicated a significant difference among the responses to the five word-contrasts ($F(4,32)=3.127$, $p < 0.028$). Since the "myror/minor" non-focal contrast did not match the pattern of the other four contrasts, an additional analysis of variance was performed in which this contrast was excluded. Also this analysis failed to reveal a significant difference between the adult-directed speech *vs.* infant-directed speech ($F(1,8)=2.938$, $p < 0.125$).

To assess specific differences in discrimination performance for each of the word contrasts involved, a series of Kruskal-Wallis analysis of variance was made. The results (two-tailed) reveal a strong tendency for a difference between adult- and infant-directed speech in the case of the single focus contrast, i.e. no change in the target word ("myror", focal *vs.* non-focal, $p < 0.054$). For the focal *vs.* non-focal contrast involving change of the target word, "myror" *vs.* "manar" the two-tailed probability was only $p < 0.199$.

The consequences of the change in the sentence focus were analysed by the Wilcoxon's test. The results indicate only a tendency ($p < 0.08$) towards sensitivity to focus changes in the case of the "myror"/"minor" contrasts, produced with adult-directed speech.

DISCUSSION

In this paper we considered only our current data on infant sentence discrimination. The present results are based on a very small sample. Thus, given the variance in the responses of the subjects, it is wise to view these results only as a preliminary indication of possible response patterns.

Subjects' sensitivity to segmental and prosodic cues

The within-subjects' results indicate that the infants' performance varied significantly depending on the word contrasts that they were tested with. Within each group, discrimination of a target word without associated changes in focus seems to be dependent to the relative prominence of the segmental changes involved. The discrimination scores for "myror"/"minor" contrasts in focal position are lower than those for "myror"/"manar", as it might be expected on the basis of the phonetic differences involved. The contrasts involving displacement of the sentence focus produced large differences in the d' scores but there seems to be a complex interaction between focal displacement and magnitude of the word contrast. It could be expected, for instance, that the within-subjects' discrimination scores for a contrast involving only changes in the target word would be systematically lower that those involving both a lexical and a focus change. In fact this was only true for the contrasts between "myror" and "minor", in adult-directed speech. Thus, it seems that the contrast "myror"/"manar" was so salient by itself that the additional change in focus did not contribute significantly to further improvement in the discrimination scores.

Influence of the type of speech on the discrimination performance

The pattern of variation in the d' scores suggests a difference in the performance of the adult-directed speech group and the infant-directed speech group. However, the variance within each of the groups is too large to enable a statistically significant difference. Since previous research indicates differences in the infants' attention to infant-directed and adult-directed speech, it would not be surprising to find significant differences in this case too. However, because the present results were obtained from a very small sample there is high probability of type II error.

One important aspect emerging from the present results is that the possible significant difference between the adult-directed and the infant-directed types of speech does not occur in the expected direction. The discrimination scores are actually worse in the case of infant-directed speech than for adult-directed speech. If this pattern holds, it suggests that the infants' attention may be overloaded by a focus on paralinguistic aspects. Infants prefer to listen to the prosodically richer speech involved in infant-directed speech [4] but they seem to drop their attention to the segmental information it conveys.

ACKNOWLEDGEMENT

This research is supported by The Bank of Sweden Tercentenary Foundation, Grant 94-0435.

REFERENCES

[1]Fernald, A. (1985), "Four-month-old infants prefer to listen to motherese", *Infant Behavior and Development 8*, 181-195.

[2]Fernald, A. (1989), "Intonation and communicative intent in mothers' speech to infants: Is the melody the message?", *Child Development 60*, 1497-1510.

[3]Fernald, A. (1992), "Human maternal vocalizations to infants as biologically relevant signals: an evolutionary perspective", in Barkow J., Cosmides L. and Tooby J. (eds.) *The adapted mind: evolutionary psychology and the generation of culture*, Oxford Univ. Press, New York, 391-428.

[4]Fernald, A. and Kuhl. P. (1987), "Acoustic Determinants of Infant Preference for Motherese Speech" *Infant Behavior and Development 10*, 279-293.

[5]Fernald, A. and Mazzie, C. (1991), "Prosody and Focus in Speech to Infants and Adults", *Developmental Psychology 27*, 209-221.

[6]Jusczyk, P. (1985), "On characterizing the development of speech perception", in J. Mehler and R. Fox (Eds.), *Neonate cognition: Beyond the blooming, buzzing confusion*, Hillsdale, N.J.: Laurence Erlbaum Assoc., 199-229.

[7]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*, Timonium, MD: York Press, 17-64.

[8]Jusczyk, P., Kelmer Nelson, D., Hirsh-Pasek, K., Kennedy, L., Woodward, A. and Piwoz, J. (1992), "Perception of acoustic correlates of major phrasal units by young infants", *Cognitive Psychology 24*, 252-293.

[9]Jusczyk, P., Luce, P. and Charles Luce, J. (1994), "Infants' sensitivity to phonotatic patterns in the native language", *Journal of Memory and Language 34*,630-645.

[10]Kuhl, P. (1985), "Methods in the study of infant speech perception", in G. Gottlieb and N. Krasnegor (Eds.), *Measurement of Audition and Vision in the First Year of Postnatal Life: A Methodological Overview*, Norwood, N. J., Ablex, 223-251.

[11]Lacerda, F., Sundberg, U. and Andersson, C. (1995), in preparation.

[12]Werker, J. and McLeod, P. (1989), "Infant preference for both male and female infant-directed talk: a developmetnal study of attentional and affective responsivness", *Canadian Journal of Psychology 43*, 230-246.

EMOTIONAL INFORMATION IN YOUNG INFANTS' VOCALIZATIONS

Yoko Shimura and Satoshi Imaizumi***
** Saitama University, **University of Tokyo, Tokyo, Japan*

ABSTRACT

Developmental aspects of four infants' ability to express emotions through vocalizations were studied based on perceptual rating experiments against 9 reference words for 200 voice samples recorded at 2 months of age. Infants even at 2 months of age can produce vocal elements necessary to express emotional contrasts which are identifiable for adult listeners.

INTRODUCTION

There is a hypothesis proposed that "infants begin to communicate through nonlinguistic aspects of voice rather than linguistic aspects at very early stage of their life"[1]. One way to test this hypothesis is to observe interaction scenes between the infant and the parent and collate the contents of such interaction with the infant's vocalization or expression. Several attempts to explore this field have been made so far[2]. Another way to test the hypothesis is to determine experimentally whether conditions necessary for communicating through nonlinguistic aspects of voice exist at an early stage of the infant's development. We adopted the latter method, defined "emotion" as "information communicated through nonlinguistic aspects of voice" and have conducted several experiments. The object of our study was to confirm some of the following conditions for communication through nonlinguistic aspects; (a) infants can produce vocalizations necessary for communication, (b) the parents as well as surrounding adults can interpret meanings contained in the infant's vocalizations with a certain degree of regularity, and (c) Infants can also interpret meanings contained in the vocalizations of other infants at some stage of their life.

Our previous reports suggested the following: (i) adults listeners can perceive a rich variety of emotional contrasts such as "pleasant vs. discomfort" even in voices made by infants older than 6 months of age[3]; and (ii) there were significant consistency in adults attributions of infants' vocalizations, although significant differences due to the child-rearing experiences were also observed[4].

The purpose of the present study is to test (1) if 2 months age young infants can express emotions through vocalizations, and (2) if so, what acoustic aspects of voice convey such emotional information.

METHOD

Recording

Totally 200 voice samples were recorded from four Japanese infants (three male, one female), during playing with their mothers at home, on the day when they were just 2 months of age. They were raised in households where standard Japanese was spoken. Through a questionnaire, it was confirmed that the infants showed normal development and behavior. The voice samples were presented in a random order repeating each sample five times.

Perceptual Rating

The listeners participated were 15 university students with normal hearing whose mother tongue was Japanese. They rated each voice sample using nine 9-point dipole scales related to emotions, listed in Table 1. Nine relatively independent items (happy, sad, laughing, pleased, frightened, demanding, rejecting, seeking affection, angry) were selected as the basic rating scales through preliminary experiments, and "calm", "surprising", "friendly" and "awful" were included to form dipole scales. Five terms representing the manner of vocalization, such as "speaking", "singing", "crying", "shout" and "secret talk" were added to study the relationships between voicing modes and emotional contents.

The experiments were conducted in a quiet room where voice samples were presented via a loudspeaker at listeners' most comfortable level.

Obtained rating scores were analyzed by a principal factor analysis and analysis of variance.

Acoustic Analysis

Using an acoustic analysis system, ten acoustical parameters were extracted from the voice samples. For 60 voice samples which had a large positive or negative factor score on a factor extracted by the principal factor analysis, an analysis of variance was carried out to extract significant relationships between the acoustic parameters and the perceptual factor scores.

Table I. Nine 9- point-rating dipole scales used in Experiment.

Happy	vs.	Sad
Laughing	vs.	Crying
Pleased	vs.	Frightened
Demanding	vs.	Rejecting
Seeking affection	vs.	Angry
Singing	vs.	Speaking
Secret talk	vs.	Shout
Calm	vs.	Surprising
Friendly	vs.	Awful

Happy +--+--+--+--+--+--+--+--+Sad
-4 -3 -2 -1 0 1 2 3 4

Fig. 1. A dipole scale used in perceptual rating.

Table II. Acoustic parameters measured.

1:	Total length
2:	No. of segments
3:	Segment type
4:	Segment length
5:	Type of Fo pattern
6:	Initial value of Fo
7:	Final value of Fo
8:	Maximum Fo
9:	Minimum Fo
10:	Fo range

RESULTS & DISCUSSION

Perceptual Rating

A principal component analysis was carried out to extract a few essential components from the rating scores on the nine dipole scales. Three principal components were extracted. Their accounting for rates in percentage were 56%, 11% and 8%, adding up to 75% in total.

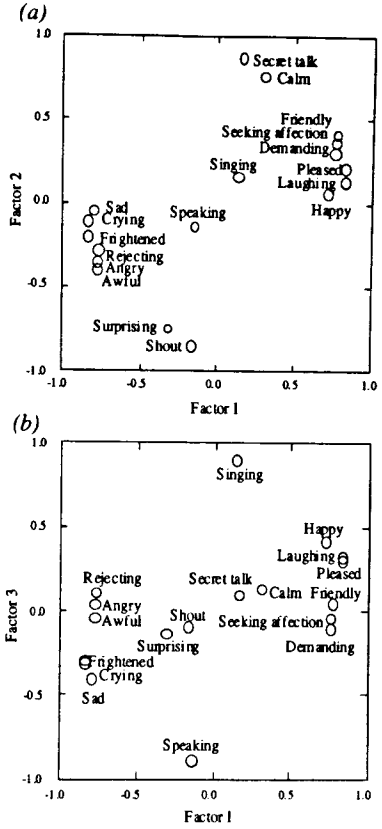


Fig. 2. Factors extracted from the rating scores given by the listeners. (a) Factor 1 vs. 2. (b) Factor 1 vs.3.

Factor loading of the rating scales after Varimax orthogonal rotation is shown in Fig. 2. It can be seen that Factor 1 represents emotions relevant to "laughing vs. crying"(0.833), "pleased vs. frightened"(0.833), "happy vs. sad"(0.791), "friendly vs. awful"(0.775), "seeking affection vs. angry"(0.767) and "demanding vs. rejecting"(0.762) and that Factor 2 represents emotions relevant to "secret talk vs. shout"(0.861), "calm vs. surprising" (0.753). Factor 3 represents emotions relevant to "singing vs. speaking"(0.869).

The above results suggest the following; (1) Factor 1 can be

considered as information pertaining to "laughing / pleased / happy / friendly vs. crying / frightened / sad / awful"; "pleasant vs. discomfort" (2) Factor 2 can be considered as information pertaining to "secret talk / calm / vs. shout / surprising";"calm vs. surprising" (3) Factor 3 can be considered as information pertaining to "singing vs. speaking."

The factor scores were analyzed by an ANOVA to test the significance of individual differences between infants. Differences between infants were significant with respect to all three factor scores ($p<0.0001$).

Figure 3 (a) shows the 90% confidence area of the factor scores on Factor 1 (F1) vs. Factor 2 (F2) for the four infants of 2 months of age, whereas Fig. 3 (b) the 90% confidence area of the factor scores for the six infants of 2 months of age. The later results were obtained from our previous report.

Although individual differences are observed as predicted by the analysis of variance, the 90% confidence areas of 2 months of age infants are narrower than those of the 6 months of age infants. The 90% confidence areas of the former group are restricted in the area representing emotional contrasts between "pleasant vs. discomfort," while those of the later group expand toward the area representing not only emotional contrasts between "pleasant vs. discomfort" but also "speaking" vs. "singing".

These results suggest that although even 2 months of age infants can produce vocalizations inducing significant infant-dependent differences in emotional contrasts, their ability to express emotional contrasts seems to develop with their age.

Acoustic Characteristics

Table III shows the relationships between the acoustic parameters and the perceptual factor scores.

For Factor 1 representing the emotional contrast between "pleasant vs. discomfort," only three acoustic parameters, length, No. of segments, and minimum Fo, had significant differences at 1% level. Voice samples perceived as "discomfort" had a longer length of

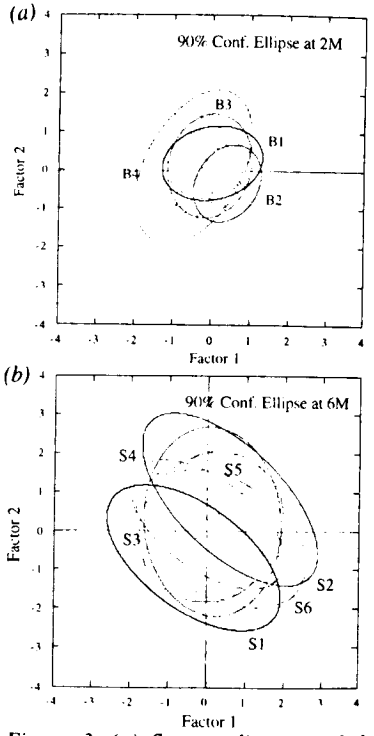


Figure 3. (a) Scatter diagram of the factor scores on F1 and F2 extracted from four 2 months of age infants. (b) Scatter diagram extracted from six 6 months of age infants.

vocalization with a less number of segments, and a lower minimum Fo than the ones perceived as "pleasant".

For Factor 2 representing "calm vs. surprising", two parameters were significant at 1% level and other two at 5% level. The voice samples perceived as "surprising" had a less number of segments with a higher and wider range of Fo than the others. Fo patterns of the "surprising" voice samples were more complex than the others.

Many acoustic parameters correlated with Factor 3 representing "speaking vs. singing." The voice samples perceived as "speaking" had a longer duration with more number of segments, and lower Fo with a narrow Fo range than the others. The voice samples perceived as "speaking" had rising-falling or falling-rising Fo

Table III. Relationships between the acoustic parameters and perceptual factors.

Factor	F1: Pleasant vs. Discomfort		F2: Calm vs. Surprise		F3: Singing vs. Speaking	
Length	short	long**	short	long	short	long**
No. Seg.	many	less*	many	less**	less	many**
Init Fo	high	low	low	high	high	low**
Final Fo	high	low	low	high	high	low**
Max. Fo	high	low	low	high*	high	low**
Min. Fo	high	low**	high	low	high	low**
Fo Range	narrow	wide	narrow	wide*	wide	narrow
Fo Type	∩, −, ／, ＼,complex ,∩∩	∪	∪, −, ／, ＼	∩,∩∩, complex**	−, ／, ＼, ∩∩,complex	∩, ∪

patterns, while the ones perceived as "singing" had flat, rising, falling or complex Fo patters.

These results suggest the following. (1) Infants even at 2 months of age can produce vocal/acoustical elements necessary to express emotional contrasts which are identifiable for adult listeners. These acoustical elements were estimated as Fo range, Fo pattern, minimum and maximum Fo values, vocalization length, and the number of segments. (2) Even 2 months of age infants can produce various voices which induce consistent interpretations or responses in adult listeners about the infants' emotional states. Even if the induced responses or interpretation of the emotional contents might not be the same as the infants' actual emotional state, surrounding people tend to interpret infants' vocalizations in a lawful way. This is important because lawful responses by the surrounding people may induce infant's notice on the social meanings of his/her own vocalizations.

CONCLUSION

Developmental aspects of four infants' ability to express emotions through vocalizations were studied based on perceptual rating experiments against 9 reference words for 200 voice samples recorded at 2 months of age. By a factor analysis for perceptual rating scores, three factors representing emotional contrast of "pleased vs. discomfort", "calm vs. surprising," and "speaking vs. singing" were extracted. Acoustical analyses showed that these factors significantly correlate with the acoustical parameters such as F0 range, F0 pattern, minimum and maximum F0 values, vocalization length, and the number of segments. These results suggest that infants even at 2 months of age can produce vocal elements necessary to express emotional contrasts which are identifiable for adult listeners.

Acknowledgment

This research was partially supported by Grant-in-Aid for Scientific Research for Priority Areas on "Cognitive and Linguistic Development," No. 06205102, Ministry of Education, Science and Culture, Japan.

References

[1] Shimura, Y., Imaizumi, S., et. al. (1990), "Infant' Vocalization Observed in Verbal Communication: -Acoustic Analysis-", *Proceedings of ICSLP90*, pp.1329-1332.

[2] Toda, S., Bornstein, M., et. al. (1991) "The Relationship Between Mother's Behavior and Child's Play and Language Development," *Proceedings of Japan Society of Development Psychology, 1214.*

[3] Shimura, Y. and Imaizumi, S. (1992) "Emotional Expression by Infant through Vocalization - Perceptual Evaluation -", *The Japan Journal of Logopedics and Phoniatrics*, vol. 33, No. 4, pp. 325-332.

[4] Shimura, Y. and Imaizumi, S. (1993) "Listener and Context Effects in Perception of Emotional Aspects of Infants' Vacuolization ", *The Japan Journal of Logopedics and Phoniatrics*, vol. 34, No. 4, pp. 417-424.

VOICE ONSET TIME IN SPEECH DIRECTED TO INFANTS AND ADULTS

Ulla Sundberg
Department of Linguistics, Stockholm University

ABSTRACT

Voice Onset Time (VOT) was measured in three Swedish mothers' infant-directed speech (IDS) and compared to VOT in their adult-directed speech (ADS). In this preliminary study VOT was significantly shorter in IDS than in ADS. The impact of stress was very clear in both IDS and ADS showing significantly longer VOT in stressed positions as compared to the unstressed. The shorter VOT in IDS and the finding that the absolute range of VOT values was smaller in IDS than in ADS, may suggest vocal accommodation [1] from the mothers' part.

INTRODUCTION

Infant-directed speech (IDS) differs in several aspects from adult-directed speech (ADS). The most striking differences are the well documented prosodic modifications with exaggerated intonation contours, shorter utterances and longer pauses, for example [1]. Acoustic investigations of IDS and ADS at the segmental level are rather scarce although some studies have been performed regarding vowel formant structure [2] and vowel modification [3] and pre-boundary vowel lengthening [4] e.g. IDS is often referred to as being more clear and informative than ADS [3, 5, 6]. One factor that might influence clearness in speech is Voice Onset Time (VOT), i.e. the duration of the gap between the burst marking the release of the articulatory closure and the onset of voicing in stops. In most languages VOT may serve as one factor separating voiced from voiceless segments [7, 8] into the two phonemic categories. In speech directed to children during their early language acquisition period, it was found that, in contrast to ADS where VOT varied drastically in duration, the mothers prolonged VOT in voiceless stops, thus producing extreme ones, and in this way efficiently separated voiceless stops from their voiced cognates, (Bernstein-Ratner, 1984, [4] referring to an unpublished investigation by Moslin, 1979.)

One other aspect of IDS that could be illustrated by analysis of VOT is phonetic accommodation, i.e. the interactors' mutual affective engagement resulting in accommodation of speech style to one another [1]. One expression of intimacy in the interaction between the mother and her baby might be reflected in the mother's way of mimicking young children's early speech with e.g. unaspirated stops [9] [1].

The objective of the present study is to make a preliminary assessment of how Swedish mothers' use VOT in IDS and ADS.

PROCEDURE

Three Swedish mothers interacted with their three-month old infants, two girls and a boy, in an isolated booth for 10 to 15 minutes. The mothers were instructed to play with some toys in a way they would normally do at home. The IDS sample was collected from this session. The ADS sample was collected immediately after this session. The investigator entered the room and talked informally with the mother about different topics concerning the infant. The sessions lasted 20–30 min. The first 4-5 minutes of the IDS and ADS samples were selected for analysis. Before analysing the data the investigator made an auditory judgement of the dialogues marking the words that were perceived as being the most prominent one(s) in each utterance. VOT in the syllables carrying lexical stress in these prominent words are in the following called stressed. The VOT was defined as the time gap between the onset of the burst and the onset of voicing, both marked on the speech wave signal.

[1] Older children at the age of 3–4 years may produce longer VOT than adults [10].

RESULTS

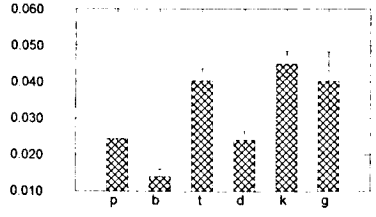


Figure 1a. Mean VOT in msec, in IDS. The bars indicate the standard error.

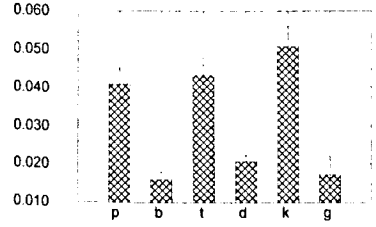


Figure 1b. Mean VOT in msec, in ADS.

VOT values in stops in word initial and medial positions were subjected to a 3-way analysis of variance. Addressee, consonant type and stress were the three factors. All three mothers showed great overlap of VOT within the categories voiced and voiceless in both speech directions. VOT was significantly longer in ADS as compared to IDS, ($F(1,222)=3.955$, $p < 0.05$). A very significant difference was found regarding consonant type and VOT duration, ($F(2,222)=6.812$, $p<0.001$), showing that the labials had shorter VOT than the dentals, and the dentals had shorter than the velars, see figure 1a and b. The voiced stops were excluded from further analysis due to uneven numbers or missing values in some cells, which in turn reflects the finding made in earlier investigations e.g. [7] that Swedish voiced stops lack aspiration phase.

The results revealed significantly longer VOT in words with stress than in words without stress ($F(1,222)=9.656$, $p < 0.002$). No interaction was found between the two factors of stress and consonant type ($F(2,222)=0.326$, $p > 0.7$). A strong interaction, on the other hand, was found between stress and addressee ($F(1,222)=3.128$, $p < 0.08$).

DISCUSSION

Adults use a speech style qualitatively different in their interaction with infants as the style that is used amongst adults. The mothers in this investigation produce shorter VOT in IDS, regardless of stress, than in their ADS. These results may indicate that mothers adjust their speech to a style closer to children's own early speech. (Phonological reductions in IDS may serve a similar purpose as suggested by Shockey & Bond (1980), [11], by "setting a tone of intimacy in a dialogue"). Vocal accommodation have mainly been suggested in the context of children's vocal behaviour [1]. There

are no obvious reasons though, to believe that adults wouldn't adjust their speech to closer conformity with children's speech when interacting with infants.

The influence of stress on VOT was shown very clearly in this study by the significantly longer values in stressed positions as compared to unstressed positions. These results are in accordance with earlier investigations of both so called lab speech [7, 12] citation form, [13, 14] and spontaneous speech [14]. Krull (1991) who made the first VOT measurements of VOT in spontaneous Swedish found 30–100% longer VOT in stressed CV-sequences than in corresponding unstressed syllables. A noteworthy aspect in the present study is that the range of the VOT values in IDS was smaller than in ADS, suggesting a less pronounced separation of voiced and voiceless stops in IDS by means of VOT. This is contrary to the expectations, but it must be kept in mind though that VOT is only one out of several acoustic features affecting the voiced/voiceless distinction. Löfqvist (1976), [12] e.g. suggests a complex differentiation of voiced and unvoiced segments in terms of closure duration, which is inversely related to VOT, comprising acoustical duration of the consonant, acoustical duration of the vowel, duration and spectral extensiveness of the vowel formant transitions, to mention some.

The narrower range of VOT values in absolute terms in IDS as compared to ADS lends support to the suggestion mentioned earlier that VOT in IDS may be one phonetic feature signalling the mothers' vocal accommodation [1]. At the age of 11 months children's production of stop-vowel syllables is often characterised by simultaneous, or almost simultaneous, release of the stop and onset of voicing [9]. By using a speech style mimicking children's way of speaking, virtual 'Baby Talk', the mother adjusts her speech into closer conformity with her interlocutor, a phenomenon that according to Giles (1984) may unconsciously be increasing the perceived attractiveness to the listener and the level of involvement.

The discrepancy between the results from the present investigation showing shorter VOT in IDS than in ADS and those from Moslin (1979), [4], who found the inverse relationship are striking. A possible explanation may be a developmental change of the characteristics in IDS related to the age and language development of the addressee [15]. In Moslin's investigation the children were "in the first stages of language learning", i.e. in the age range 12 to 20 months. The function of IDS to children of these ages are often related to object play and language acquisition [15] that may affect a phonetic feature such as VOT to have a more definitive clarification function. The age of the addressees in the present study is only three months, and the social context and the kind of interaction is here very different from that mentioned in the previous study. In the interaction with young infants the goal is to engage them in a face-to-face play, the tone is intimate and in the present study the mothers often encouraged the infants to vocalise. The differences in context and in goal of interaction in the two studies may thus account for the discrepancy of VOT in the observed IDS and ADS patterns.

ACKNOWLEDGEMENT

I would like to express my gratitude to Francisco Lacerda for assistance with the statistical analysis and helpful discussions.

REFERENCES

[1] Locke, J.L. (1993), *The child's path to spoken language*, Harvard University Press, Cambridge, Mass.

[2] Davis, B. & Lindblom, B. (1992), "Prototypical vowel information in Baby Talk", *PERILUS,* No XV, pp 119-124.

[3] Bernstein Ratner, N. (1984), "Patterns of vowel modification in mother-child speech", *Journal of Child Language*, 11, pp 557-578.

Moslin, B. (1979), "*The role of phonetic input in the child's acquisition of the voiced-voiceless contrasts in English stops: A VOT analysis",* Unpublished doctoral dissertation, Brown University.

[4] Bernstein Ratner, N. (1986), "Durational cues which mark clause boundaries in mother-child speech", *Journal of Phonetics*, 14, pp 303-309.

[5] Fernald, A. (1983), "The perceptual and affective salience of mothers' speech to infants", in *The Origins and Growth of Communication,* (Eds. L. Feagans, C. Garvey, R. Golinkoff) Ablex Publishing Corporation, Norwood, New Jersey, pp 5-29.

[6] Fernald, A. & Mazzie, C. (1991), "Prosody and focus in speech to infants and adults", Developmental Psychology, Vol. 27, No. 2, pp 209-221.

[7] Fant, G. (1973), *Speech sounds and features*, MIT Press, Cambridge, Mass. p 110 ff.

[8] Lisker, L. & Abramson, A., (1964), "A cross-language study of voicing in initial stops: Acoustic measurements", *Word*, 20, pp 384-422.

[9] Fletcher, P. & Garman, M. (1986), *Language Acquisition, Studies in first language acquisition.* Cambridge University Press, p. 161.

[10] Menyuk, P. & Klatt, M. (1974), "Voice onset time in consonant cluster production by children and adults", *Journal of Child Language*, 2, pp 223-231.

[11] Shockey, L. & Bond, Z.S. (1980), Phonological processes in speech addressed to children. *Phonetica* 37: 267-274.

[12] Löfqvist, A. (1976), "Closure duration and aspiration for Swedish stops", *Working Papers 13, Phonetics Laboratory, Dept of General Linguistics*, Lund University, pp 1-39.

[13] Engstrand, O. (1983), Articulatory co-ordination in selected VCV utterances: A means - end view. (Ph.D. diss.) *Reports from Uppsala University Department of Linguistics* (RUUL) 10.

[14] Krull, D. (1991), "VOT in spontaneous speech and in citation form words", *PERILUS* XII, Stockholm University, pp 101-107.

[15] Stern, D.N., Spieker, S, Barnett, R.K., & MacKain, K. (1983), "The prosody of maternal speech: Infant age and context related changes", *Journal of Child Language*, 10, pp 1-15.

INTRINSIC F0 IN THE BABBLING OF MANDARIN-LEARNING INFANTS

D. H. Whalen, Andrea G. Levitt[1], Pai-Ling Hsiao[2], Julia Irwin[2] and Winifred McGowan
Haskins Laboratories ([1]also Wellesley College) ([2]also University of Connecticut)

ABSTRACT

Intrinsic F0 (IF0) is the tendency for high vowels to have a higher F0 than low vowels. We previously found IF0 in babbling of French- and English-learning infants, suggesting an automatic effect. Here, we find IF0 in four 12-month-old Mandarin babblers, even though they are learning a tone language. Thus it seems that IF0 is to be explained not as an enhancement of the vowel quality difference but rather as an automatic consequence of vowel formation.

INTRODUCTION

Languages differ on many dimensions, but other features are consistent across languages. One phonetic feature that has been found to accompany vowels is "intrinsic F0" or "intrinsic pitch" (IF0, from here on). This is the tendency for the high vowels, such as [i] and [u], to have a higher fundamental frequency than the low vowels, such as [a] and [œ]. IF0 was first noticed for German [1] and has since been found in every language that has been examined for it. In a previous survey, we found evidence of IF0 in 31 languages from 11 of the 29 major language families of the world, and no instances of a lack of IF0 [2].

There has been considerable debate about the mechanism responsible for this effect . The present work does not directly support any particular explanation, so the reader is directed to the various surveys of the explanations to be found elsewhere [3-7] All of the prior explanations have assumed that IF0 is an automatic consequence of articulation, probably due to the pull of the tongue on the laryngeal system, or to an acoustical interaction between F1 and F0. (Steele [8] argues that there must be a contribution of subglottal pressure.) Some authors suggest that only a combination of explanations can account for all of the facts.

However, more recently there have been proposals that IF0 is a deliberate manipulation of F0 that is introduced in the signal to enhance the differences between vowel categories [9-11] On this account, speakers try to accommodate their listeners by making F0 closer to F1 for high vowels and farther from F1 for low vowels. There is some evidence that listeners perceive vowel height not in terms of F1 by itself but by a difference between F1 and F0 [12]. The enhancement account, then, asserts that speakers have control over this aspect of the F0, and IF0 is simply a particularly useful enhancement.

There are two predictions that come from the enhancement account. First, there should be some population that chooses not to enhance its vowels in this way. Second, there should be some developmental change in the use of IF0 if it is an enhancement that needs to be learned.

The first prediction, of a language which chooses not to use IF0, has not been borne out. The survey of Whalen and Levitt [2] found no instances of languages which lacked IF0. The survey included languages with quite different vowel inventories, and still found no difference across languages, even though an enhancement would seem to be more useful in a crowded space than in a sparse one. It is never possible to prove that there is no language that exhibits a certain trait, since an example could be waiting to be discovered. For IF0, however, the sample of languages was broad and also included cases where one would expect to find no difference. A cogent example of such a language is one which uses F0 for lexically distinctive tones and, at the same time, has a small vowel inventory. (Enhancement by IF0 would seem to be most useful in crowded vowel spaces.) One such language is Mandarin, and Mandarin has in fact been shown to have IF0 [13].

The second prediction, that the IF0 effect should change during language development, has also not been supported. Whalen, Levitt, Hsiao and Smorodinsky [14] examined infants in two language environments (French and English), at the ages of 6, 9 and 12 months. Despite the fact that these same infants showed significantly different use of F0 for intonation [15], they showed a typical IF0 effect that did not differ across language environment or across age. That paper also surveyed six studies with older children (6-11 years), and found no developmental trend at older ages either. If the IF0 effect is present from the beginning of linguistic production (and there is very little linguistic phonation before 6 months), it seems very unlikely that it is a learned enhancement.

An enhancement account might assume that infants are imitating IF0. It is true that every language the infant hears will show IF0 (since it is universal), and it is thus logically possible that the IF0 in babbling is imitative. However, it is not clear how the infant would know to extract this property of the signal, since the infant lacks vowel categories in the babbling stage. Furthermore, the speech directed to infants ("motherese") contains very large changes in F0 [16], which would make the extraction of the relation between vowel height and F0 that much more difficult. Finally, children learning a tone language would also hear each vowel at very different F0s, depending on the tone used with it. All of these factors make the task of detecting the IF0 extremely difficult for the child.

Nonetheless, if any population were to benefit from avoiding IF0, it would seem to be learners of a tone language. Tone is crucial for lexical distinctions, and it depends largely on F0, which is a phonetic dimension that seems to be under the infant's control earlier than segmental ones. Indeed, tonal categories seem to be mastered sooner than segmental categories [17]. Even if the IF0 contribution were a deliberate enhancement in tone languages, it would seem that the learner of a tone language would be most likely to use F0 just for emerging tone distinctions instead. In order to test this directly, we recorded the babbling of four Mandarin-learning infants and measured the F0s of the vowels to see whether the IF0 effect found for French- and English-learning infants also appeared for these infants learning a tone language.

THE EXPERIMENT

We measured the F0 of all non-central vowels in the babbling of four Mandarin-learning infants.

Subjects

The infants were being raised as monolingual speakers of Mandarin. Infants were selected for the study only if both parents were native speakers of Mandarin Chinese. Most were from the Beijing area. The children were living in Storrs, Connecticut, while one or both parents attended the University of Connecticut. Most of these students planned to return to Mainland China after graduation and were therefore raising their children as monolingual speakers of Mandarin. One of the four had a monolingual Mandarin-speaking grandparent taking care of him for the duration of the recording sessions. Another had a grandparent visiting during part of the recording period.

Recordings

The infants were recorded in the home every other week for a session lasting 30-45 minutes. Recordings started at six or seven months of age and ended at 11-16 months. A Panasonic SV-3700 DAT tape recorder was used in conjunction with a Realistic wireless microphone. The microphone itself was sewn into a vest (concealed as the center of a flower) which the infant wore during the session. In this way, a relatively constant distance between the infant's mouth and the microphone could be maintained without restraining the child.

Analysis

The recordings were transferred to a VAX computer for analysis. The utterances were selected as being speech-like and separated from other sounds by 750 ms or more. All utterances were then transcribed by a native speaker of Mandarin. The symbols of the IPA were used, with the understanding that some of the utterances would be ambiguous at this level of detail.

For the present analysis, only the 12 month recordings were used. One subject (BX) returned to China in his

eleventh month, so his eleventh month recordings were used instead. We further restricted the analysis to non-central vowels; we also excluded /ɛ/ (a very common vowel in these transcriptions) as a practical way of reducing the number of tokens to be analyzed without sacrificing the points of interest.

F0 was measured by delimiting ten pitch periods by hand in the acoustic waveform. This was performed with the program HADES, written at Haskins Laboratories [18]. We tried to take a measurement at a point 40% of the way into the syllable. In the best case, there were five periods on either side of that point, which would give us a single, average value for that stretch of speech. If that portion turned out to be unmeasurable, a stretch of ten periods as close to the 40% point and still within the syllable was found. These selection criteria resulted in 3155 vowel tokens that were measured. F0 values greater than 850 Hz were excluded to reduce the influence of occasional outlier, resulting in a final analysis of 3054 tokens. Of these, 2752 were transcribed as /e/.

RESULTS

As Table 1 shows, IF0 is present in the babbling of these four Mandarin-learning 12-month-olds. (Analyzing the results according to front/back as well as height was not possible because all subjects had gaps in their results that way.) The one negative difference (TZ) can be presumed artifactual, because of the small number of low vowels for this subject. Similarly, the large difference for BX also depends on a small number of low (and high) vowel tokens. The other two subjects show a difference of just the size we would expect based on our previous work with babbling. The overall difference is smaller than expected only because a large proportion of the low vowels happen to come from a speaker (EW) with a high overall F0.

DISCUSSION

Intrinsic F0 (IF0), which has been found in every language measured so far and in our previous study of babbling, has been found here in the babbling of Mandarin-learning infants as well. The size of the effect is of the same magnitude as in the earlier study of French- and English-learning infants.

Table 1. Average F0 for vowels of three heights (in Hz) for the four subjects and in a weighted average across subjects. Number of tokens is given below the F0 value.

Vowel Height	High	Med	Low	H-L
EW	404 (41)	369 (781)	348 (82)	56
TZ	328 (25)	293 (718)	336 (17)	-8
YL	332 (70)	311 (996)	279 (25)	53
BX	395 (10)	346 (274)	253 (15)	142
Mean	356	326	324	32

Thus even in a language that uses a sparse vowel space and lexical tones, infants exhibit IF0 in their own productions.

These results are incompatible with the notion that IF0 is a deliberate enhancement of the speech signal. That position assumes that the increase in F0 for high vowels helps to shift the effective F1 and thus enhance the vowel category differences. Infants learning Mandarin need to learn to produce the tone contours if they are to become successful speakers. Therefore, they have every reason to attend to the tonal aspects of F0 and to ignore, if possible, confounding factors such as IF0. If there was a population that would seem to benefit from ignoring this (potential) enhancement, it would appear to be the Mandarin learners. The fact that they do not is further evidence that IF0 is not an enhancement. Rather IF0 appears to be an automatic consequence of vowel production (from whatever source or combination of sources), even in infants.

ACKNOWLEDGMENT

This work was supported by NIH grant DC-00403 to Catherine T. Best and Haskins Laboratories.

REFERENCES

1. Meyer, E.A. (1896-7), "Zur Tonbewegung des Vokals im gesprochenen und gesungenen einzelwort." *Phonetische Studien (Beiblatt zu der Zeitschrift Die Neuren Sprachen)*, vol. 10, pp. 1-21.

2. Whalen, D.H. and A.G. Levitt. (in press), "The universality of intrinsic F0 of vowels." *Journal of Phonetics*, in press.

3. DiCristo, A., D.J. Hirst, and Y. Nishinuma. (1979), "L'estimation de la F0 intrinsèque des voyelles: etude comparative." *Travaux de L'Institut de Phonétique D'Aix-en-Provence*, vol. 6, pp. 149-176.

4. Shadle, C.H. (1985), "Intrinsic fundamental frequency of vowels in sentence context." *Journal of the Acoustical Society of America*, vol. 78, pp. 1562-1567.

5. Fischer-Jørgensen, E. (1990), "Intrinsic F0 in tense and lax vowels with special reference to German." *Phonetica*, vol. 47, pp. 99-140.

6. Sapir, S. (1989), "The intrinsic pitch of vowels: Theoretical, physiological and clinical considerations." *Journal of Voice*, vol. 3, pp. 44-51.

7. Silverman, K.E.A. (1987), *The structure and processing of fundamental frequency contours*. Unpublished Ph.D. thesis, University of Cambridge.

8. Steele, S.A. (1986), "Interaction of vowel F0 and prosody." *Phonetica*, vol. 43, pp. 92-105.

9. Diehl, R.L. and K.R. Kluender. (1989), "On the objects of speech perception." *Ecological Psychology*, vol. 1, pp. 121-144.

10. Diehl, R.L. (1991), "The role of phonetics within the study of language." *Phonetica*, vol. 48, pp. 120-134.

11. Kingston, J. (1993), "The phonetics and phonology of perceptually motivated articulatory covariation." *Language and Speech*, vol. 35, pp. 99-113.

12. Traunmüller, H. (1981), "Perceptual dimension of openness in vowels." *Journal of the Acoustical Society of America*, vol. 69, pp. 1465-1475.

13. Shi, B. and J. Zhang. (1987). "Vowel intrinsic pitch in standard Chinese", in *Proceedings XIth International Congress of Phonetic Science*, pp. 142-145. Academy of Sciences of the Estonian SSR: Tallinn, Estonia.

14. Whalen, D.H., A.G. Levitt, P.-L. Hsiao, and I. Smorodinsky. (1995), "Intrinsic F0 of vowels in the babbling of 6-, 9- and 12-month-old French- and English-learning infants." *Journal of the Acoustical Society of America*, vol. 97, pp. 2533-2539.

15. Whalen, D.H., A.G. Levitt, and Q. Wang. (1991), "Intonational differences between the reduplicative babbling of French- and English-learning infants." *Journal of Child Language*, vol. 18, pp. 501-516.

16. Fernald, A., *et al.* (1989), "A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants." *Journal of Child Language*, vol. 16, pp. 477-501.

17. Clumeck, H. (1980). "The acquisition of tone", in *Child phonology, volume 1: Production*, G. Yeni-Komshian, J.F. Kavanagh, andC.A. Ferguson, Editor, pp. 257-275. Academic Press: New York.

18. Rubin, P.E. (1995). "HADES: a case study of the development of a signal analysis system", in *Applied Speech Technology*, A. Syrdal, R. Bennett, andS. Greenspan, Editor, pp. 501-520. CRC Press: Boca Raton, FL.

PALATOMETRIC SPECIFICATION OF HINDI /S/ AND /S/*

R. Prakash Dixit
Louisiana State University, Baton Rouge, La, USA

ABSTRACT

The area of tongue-palate contact, the length of front cavity, the place of groove, the width of groove and the length of groove during /s/ and /S/ production were obtained from a native speaker of Hindi using the technique of dynamic palatometry. Nonsense words of the form /bi-íb/, /ba-áb/ and /bu-úb/, containing the sibilants in a symmetrical vocalic context and embedded in a carrier sentence, were used for data collection. On average, the area of tongue-palate contact was greater, while the length of the front cavity and the side-to-side width of the groove were lesser for /s/ than for /S/. Front-to-back length of the groove was similar for /s/ and /S/. Both /s/ and /S/ were found to be produced in the prealveolar zone (anterior part of the alveolar ridge) of the roof of the mouth. However, the location of the groove center for /S/ was about 4 mm behind than that for /s/.

INTRODUCTION

Hindi /s/ and /S/ are, traditionally, said to be produced by placing the tongue tip against the back of the upper front teeth and the palate, respectively, and by leaving an opening along the central line of the tongue tip. (Sharma [1]). There is , virtually, no information on the type of opening (groove/slit), the width of opening (broad/narrow) and the length of opening (long/short). Moreover, there is a complete lack of quantitative data in Hindi on the above production parameters of /s/ and /S/; quantitative data on the place of groove, the length of the front cavity and the area of tongue-palate contract are also not available on Hindi sibilants. Importance of some of these parameters in distinguishing /s/ from /S/ and in constructing electrical or mechanical models of the vocal tract, which may be capable of predicting acoustical consequences of /s/ and /S/ production cannot be over stated. Thus, the purpose of the present study was to generate quantitative data on the /s/ and /S/ production parameters indicated above.

METHOD

An adult male native speaker of Hindi who had no apparent articulatory abnormalities, served as subject.

A custom-made palatometer (electropalatograph) containing 96 electrodes arranged from front-to-back in 11 rows with a 2x2 mm regular grid pattern was used. The first row was located in the dental zone and the eleventh row in the prepalatal zone 5 mm and 25 mm above and behind the edges of the central maxillary incisor teeth, respectively. The front-to-back location of the electrode rows in relation to the maxillary teeth of the subject is shown in figure 1. This figure also shows the typical pattern of tongue-palate contact for /s/ and /S/. Such plots as those shown in Figure 1 were used to take measures of various production parameters of /s/ and /S/. Threshold for generating the plots was 80%.

During recording session, the subject was seated in an anechoic room and was allowed 15 minutes to adapt to the electropalatograph after it was positioned in his mouth. The subject practiced the test sentences during this time. For data collection, the subject produced in a random order 15 repetitions of each of the nonsense words /bisíb/, /basáb/, /busúb/, /biSíb/, /baSáb/ and /buSúb/ in the carrier sentence /didi-lizije/ "Elder sister - (please) take". Recording system, described in detail elsewhere (Fletcher et al [2]), was calibrated before data collection began.

RESULTS

Quantitative data on various production parameters of /s/ and /S/ are presented in Table I. As shown in this table, the area of tongue-palate contact, in terms of contacted electrodes, was consistently greater for /s/ than for /S/. It was 43.20 (range 40.33-45.33) electrodes for /s/ and 36.82 (range 36.53-37.00) electrodes for /S/.

The front-to-back length of the groove was slightly greater for /s/ than for /S/. It was 3.69 mm (range 2.93-4.13) for /s/ and 3.16 mm (range 2.27-4.27 mm) for /S/. However, the difference in the individual measures of the groove length between /s/ and /S/ were small and unsystematic. Thus, the groove length for /s/ and /S/ can be deemed as similar.

The side-to-side width of the groove was consistently smaller for /s/ than for /S/. During /s/, the width of the groove was 5.07 mm (range 4.67-5.60 mm), while during /S/, it was 8.44 mm (range 7.20-9.33 mm).

Both /s/ and /S/ were found to be produced in the prealveolar zone of the roof of the mouth. However, the center of the groove for /s/ was consistently located about 4 mm anterior to that for /S/. During /s/, the groove center occurred 1.42 mm (range 1.06-1.66 mm) behind the lateral incisor (gingival incisor) line which forms the boundary between the dental zone and the prealveolar zone, whereas during /S/, the groove center occurred 5.33 mm (range 4.46-5.93 mm) behind the lateral incisor line.

The length of the front cavity was consistently smaller during /s/ than during /S/. It was 8.42 mm (range 8.07-8.67 mm) during /s/ and 12.33 mm (range 11.47-12.93) during /S/. The length of the front cavity was determined by adding 7 mm - the distance between the edges of the central maxillary incisors and the lateral incisor line - to the measures of the location of the groove center.

DISCUSSION

The area of tongue-palate contact was found to be consistently and substantially greater for /s/ as compared to that for /S/. Fletcher and Newman [3] reported similar differences in the area of contact between /s/ and /S/ of English. This is not an unexpected result since the area of contact and the contour of airflow channel largely depend on the location and the width of the groove: the more anterior and narrower the groove, the larger the area of contact.

In standard phonetic texts, it is generally assumed that the groove during /S/ from front-to-back is longer and from side-to-side is wider than that during /s/ (See, for example, Pike [4]). Contrarily, the front-to-back length of the groove was found to be similar during /s/ and /S/ in this study. Probably, the groove length does not play any role in separating /s/ from /S/.

On the other hand, the assumption that side-to-side width of the groove was broader during /S/ than during /s/ was strongly supported by the results of the present study. Further support for the above assumption comes from the studies by Fletcher [5] and Fletcher and Newman [3]. Like their studies, the groove for /S/ as compared to that for /s/ was found to be broader by about 3 mm in the present study. However, one of their subjects showed a difference of about 6 mm between the groove widths of these sibilants.

Supporting another assumption of the standard phonetic texts, the place of the groove was found to be more posterior for /S/ than for /s/. The difference in the place of the groove between /S/ and /s/ was about 4 mm. Similarly, a difference

of 3 - 4 mm between /S/ and /s/ groove places was reported in Fletcher [5]. In Fletcher and Newman [3] the difference in the groove places of /S/ and /s/ was, however, about 7 mm. Contrary to the phonetic description of the place of production for /S/ and /s / given in Sharma [1], both /S/ and /s/ were found to be produced in the prealveolar area about 2 mm anterior to the canine line and about 2 mm posterior to the lateral incisor line, respectively.

The front cavity length was found to be about 8 mm during /s/ and 12 mm during /S/. Similarly, Fletcher and Newman [3] reported the front cavity length of about 7 mm during /s/ and 14 mm during /S/ for one of their two subjects. These measures of front cavity length are close to those used in modeling studies by Heinz and Stevens [6], and Shadle [7]. A 10 mm front cavity length was found to be appropriate by Heinz and Stevens for the production of /s/ resonances using an electrical circuit model; and a 15 mm long front cavity was considered adequate by Shadle to produce /S/-like sibilant noise using a mechanical model of the vocal tract.

CONCLUSION

Consistent and substantial differences observed in the measures of the groove width, the place and the front cavity length suggest that these parameters singly or in a certain combination play an important role in distinguishing /s/ and /S/.

It is expected that the measures of the above parameters, reported in this study, will be useful in constructing electrical circuit models or mechanical models of the vocal tract, which may be capable of predicting acoustical consequences of /s/ and /S/ productions.

Vocalic context did not influence systematically any of the production parameters of /s/ and /S/ suggesting their resistance to coarticulatory effects.

REFERENCES

[1] Sharma, A. (1958). *A basic grammar of the Hindi language*, Agra: Agra University Press.

[2] Fletcher, S.G., McCutcheon, M.J. & Wolf, M.B. (1975), "Dynamic palatometry", J. Speech hear. Res., vol. 18, pp. 812-819.

[3] Fletcher, S.G. & Newman, D.G. (1991), "[s] and [S] as a function of linguapalatal contact place and sibilant groove width", J. Acoust. Soc. Am., vol. 89, pp. 850-858.

[4] Pike, K.L. (1958), *Phonetics*, Ann Arbar: The University of Michigan Press.

[5] Fletcher, S.G. (1989), "Palatometric specification of stop, affricate and sibilant sounds", J. Speech Hear. Res., vol. 32, pp. 736-748.

[6] Heinz, J.M. & Stevens, K.N. (1961), "On the properties of voiceless fricative consonants", J. Acoust. Soc. Am., vol. 33, pp. 589-596.

[7] Shadle, C.H. (1985), "The acoustics of fricative consonants", Tech. Rep. 506, MIT Res. Lab., Cambridge, MA.

*/S/ = /ʃ/

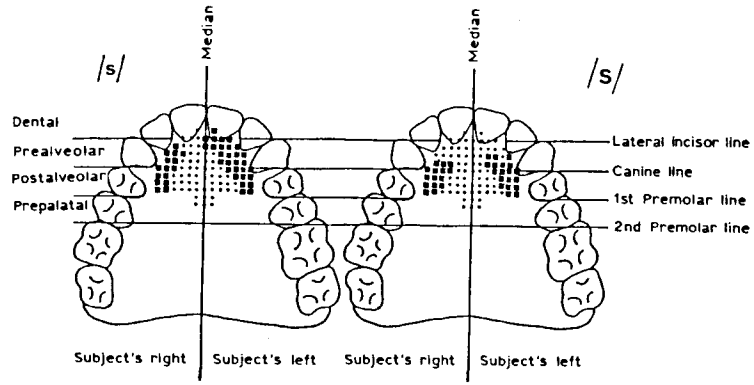


Figure 1. Front-to-back location of 11 rows of 96 electrodes on the oral surface of the palatometer in relation to the maxillary teeth of the subject. Also shown are the typical patterns of tongue-palate contact for /s/ and /S/. The electrodes contacted in 80% or more of the tokens are shown by filled squares. Notice that the airflow channel is skewed to the right of the subject's mouth.

Table I. Means ($\bar{X}$) and standard deviations (SD) for the area of tongue-palate contact, the front-to-back length of the groove, the side-to-side width of the groove, the place of the groove from the lateral incisor line, and the front cavity length from the edges of the central maxillary incisors to the location of the groove center.

	Area Contacted		Groove Length (mm)		Groove Width (mm)		Groove Place (mm)		Front Cavity Length (mm)	
Context	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD
/s/										
/i-i/	40.33	2.19	2.93	1.28	5.60	1.12	1.53	0.63	8.53	0.64
/a-a/	45.33	3.09	4.13	1.19	4.67	1.45	1.06	0.59	8.07	0.59
/u-u/	43.93	2.96	4.00	1.51	4.93	1.03	1.66	0.81	8.67	0.82
Group	43.20	3.45	3.69	1.41	5.07	1.25	1.42	0.72	8.42	0.72
/S/										
/i-i/	36.53	2.00	2.93	1.03	9.33	0.98	4.46	0.51	11.47	0.52
/a-a/	37.00	3.30	2.27	0.70	7.20	1.26	5.60	1.18	12.60	1.18
/u-u/	36.93	3.43	4.27	1.98	8.80	1.01	5.93	0.59	12.93	0.59
Group	36.82	2.92	3.16	1.57	8.44	1.41	5.33	1.02	12.33	1.02

GLOTTAL OPENING IN GERMAN OBSTRUENTS

Michael Jessen
Institute of Natural Language Processing, University of Stuttgart, Germany and Department of Modern Languages and Linguistics, Cornell University, Ithaca, New York

ABSTRACT

A transillumination study was carried out on the production of stops and fricatives by a native speaker of German. Different parameters involving the glottal opening gesture were measured and evaluated with phonation type and place of articulation as the independent variables. Maximum degree of glottal opening was found to be the most reliable correlate of phonation type. Parameters of oral-laryngeal coordination turned out to be prominent as correlates of place.

1. INTRODUCTION

Compared to our knowledge of glottal opening in the obstruent production of most other Germanic languages, relatively little evidence exists for German [4]. Using the transillumination technique two tasks are addressed in this study. One is to investigate the realization of the opposition between the tense obstruents /p,t,k,f,s/ and the lax obstruents /b,d,g,v,z/ in terms of laryngeal behavior and oral-laryngeal coordination and to compare the results to the realization of related two-way phonation type oppositions in other languages (cf. [3]). A second task is to address how differences in place of articulation are expressed articulatorily. It has been shown that aspiration differences due to place of articulation are associated with the coordination between the location of maximum glottal opening and stop release in English and German [2,4].

2. METHOD

The set of obstruents that can occur in a phonation type opposition in German were produced by a male speaker of German. The obstruents occur in two different contexts, one being intervocalic position preceded by [i] and followed by schwa (e.g. [íp^{h}ə, íbə, ít^{h}ə] etc., which are nonsense words), the other being word-initial position preceded and followed by [i] (e.g. *nie dir* [ni: dɪ̯ɐ] 'never you', *nie Tier* [ni: t^{h}ɪ̯ɐ] 'never animal' etc., which are existing words in German). Recordings were made in two different experimental sessions three weeks apart. Data across sessions were not pooled in the analysis since values for degree of glottal opening are partially specific to factors of the session such as the exact location of the fiberscope. Recordings were made of both the acoustic signal and the transillumination (TI) signal. Both the TI signal and the calculated velocity curve were smoothed to facilitate extraction of the relevant parameters of the glottal opening gesture (Figure 1). Further details of data recording and processing follow the methodology reported in [7].

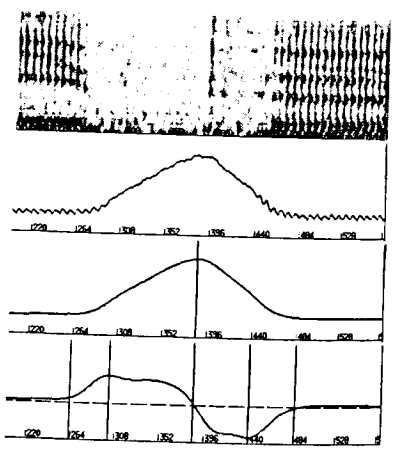


Figure 1. Spectrogram, unsmoothed TI curve, smoothed TI curve, and smoothed velocity curve of one representative token of [íp^{h}ə]. Outmost cursors indicate OG and EG, innermost cursor indicates P.

The following set of parameters was measured. Aspiration duration (Asp) and closure duration (Clos) in stops, and total duration (Total) in fricatives were determined in a spectrogram on the basis of the end (Clos, Total) and beginning (Asp, Total) of F_2 in the adjacent vowel. Maximum degree of glottal opening (Gmax) was measured as the intensity value in the smoothed TI signal at the point in time corresponding to a zero crossing in the velocity curve (P). The remaining parameters are temporal intervals involving the following oral and laryngeal events: onset of consonant (OC), onset of vowel (OV), and stop release (R), as employed in the measurement of Asp, Clos, and Total, as well as the onset of the glottal opening gesture (OG) and its end (EG), and finally the moment of peak glottal opening (P) that is employed for the Gmax measurement. OG and EG were defined as the point in time corresponding to 10% of maximum and minimum velocity, respectively. The interval between OG and EG was taken as the duration of the glottal opening gesture (Gdur).

3. RESULTS

The results are presented in Table 1. The values in each cell of Table 1 are mean values from the measurements of approximately 20 tokens. For many lax fricatives in word-initial position glottal opening was too small for measurement application or absent. This context was therefore not considered for fricatives.

Table 1. Mean values and standard deviations (separated by a slash) of the different parameters (first column) in two contexts (second column), and two experimental sessions (third column) for the stops /p,t,k,b,d,g/ (above) and the fricatives /f,s,v,z/ (below). "P-R", for example, stands for "P minus R". The values are in milliseconds, except for Gmax which is expressed in arbitrary units.

Param.	Ctx.	S.	p	t	k	b	d	g
Asp	í_ə	1	61/9	83/10	73/10	18/2	21/3	32/7
Asp	í_ə	2	49/6	89/8	66/16	16/3	25/4	34/7
Asp	i#_í	1	84/14	93/7	111/14	25/3	34/3	57/9
Asp	i#_í	2	81/13	90/10	104/8	23/5	28/4	48/10
Clos	í_ə	1	106/11	69/7	90/8	87/6	54/5	73/13
Clos	í_ə	2	97/7	67/7	92/9	82/7	54/6	77/11
Clos	i#_í	1	126/13	112/10	112/11	138/10	126/7	122/18
Clos	i#_í	2	117/8	106/6	114/14	114/8	108/10	111/13
Gdur	í_ə	1	202/8	183/11	208/8	100/7	80/8	140/11
Gdur	í_ə	2	188/14	188/8	190/10	90/13	69/8	132/15
Gdur	i#_í	1	237/15	223/11	236/15	161/17	151/10	188/13
Gdur	i#_í	2	213/15	203/14	220/16	117/18	126/19	165/15
Gmax	í_ə	1	-51/ 121	-156/ 104	307/105	-585/44	-572/52	-437/29
Gmax	í_ə	2	-862/142	-907/85	-796/139	-1305/25	-1314/23	-1280/40
Gmax	i#_í	1	-8/189	-122/247	214/303	-628/122	-677/40	-513/38
Gmax	i#_í	2	-842/183	-928/116	-659/189	-1335/24	-1316/30	-1296/16
P-R	í_ə	1	-1/5	18/9	9/7	-1/2	7/5	-7/7
P-R	í_ə	2	-5/4	28/10	12/7	0/4	10/4	-6/9
P-R	i#_í	1	0/5	11/7	17/10	-32/29	-26/19	-33/15
P-R	i#_í	2	-5/6	8/6	12/8	-17/14	-10/12	0/20
OG-OC	í_ə	1	0/5	2/5	-7/5	30/3	21/7	-1/5
OG-OC	í_ə	2	3/5	6/5	7/5	34/7	29/7	7/10
OG-OC	i#_í	1	5/7	5/3	6/11	19/7	23/6	6/9
OG-OC	i#_í	2	7/5	16/10	17/19	31/10	22/8	6/2
P-OC	í_ə	1	104/9	88/5	100/6	85/4	62/7	65/9
P-OC	í_ə	2	91/4	96/7	100/20	83/7	65/7	70/6
P-OC	i#_í	1	125/10	124/7	130/11	105/26	100/18	88/20
P-OC	i#_í	2	111/7	115/7	128/14	97/14	97/14	112/25
P-OV	í_ə	1	-63/6	-64/4	-63/7	-20/3	-13/5	-40/6
P-OV	í_ə	2	-55/4	-61/7	-55/7	-16/5	-14/4	-41/5
P-OV	i#_í	1	-84/12	-81/6	-93/8	-57/27	-60/17	-91/19
P-OV	i#_í	2	-87/13	-82/8	-90/8	-41/16	-40/11	-48/16
EG-OV	í_ə	1	33/7	32/7	36/8	24/3	24/6	31/5

EG-OV	í_ə	2	44/10	37/11	36/9	25/6	18/3	27/7
EG-OV	i#_í	1	30/7	22/3	17/4	17/6	14/3	14/7
EG-OV	i#_í	2	21/7	22/6	17/3	10/4	11/5	10/4

Param.	Ctx.	S.	f	s	v	z
Total	í_ə	1	163/9	183/11	112/13	122/12
Total	í_ə	2	160/8	170/9	106/9	116/13
Gdur	í_ə	1	200/10	218/12	152/73	236/25
Gdur	í_ə	2	195/9	200/11	152/45	222/48
Gmax	í_ə	1	265/185	261/195	-625/35	-555/68
Gmax	í_ə	2	-616/268	-872/98	-1313/41	-1297/51
OG-OC	í_ə	1	-16/7	-12/6	-4/51	-60/20
OG-OC	í_ə	2	-15/6	-7/5	-18/36	-33/29
P-OC	í_ə	1	78/9	87/7	76/28	67/15
P-OC	í_ə	2	74/7	84/7	73/13	67/15
P-OV	í_ə	1	-84/7	-95/9	-36/21	-55/11
P-OV	í_ə	2	-85/7	-85/6	-33/17	-48/13
EG-OV	í_ə	1	20/4	21/4	35/40	53/19
EG-OV	í_ə	2	18/4	21/3	26/32	73/42

One-way ANOVAs were calculated, separately for stops and fricatives and for each context and session, with each of the parameters as the dependent variable. In one set of ANOVAs the independent variable was phonation type (tense, lax), in another set, in which only the tense obstruents were included, the independent variable was place of articulation (labial, alveolar, velar). Not all details of the statistical results can be reported here. Instead the results are presented on a more general level, focusing on the question of the reliability of the parameters across different factors. A given parameter will be considered a maximally reliable correlate of phonation type if tense and lax obstruents differ significantly in this parameter across the factors of context, session, and place of articulation, and if in addition the difference has the same directionality across these factors (e.g. Asp in tense stops always longer than in lax stops). Similarly, reliability for place of articulation is determined with respect to statistical significance and unidirectionality across contexts and sessions. In addition, reliability for stops, that involve three places of articulation, has been further determined on the basis of post-hoc tests (Fisher PLSD) as the occurrence of significant differences in all three possible pairwise comparisons of place.

The parameters Asp, Gdur, Gmax, and P-OC emerge from the statistical analysis as maximally reliable correlates of phonation in stops, since differences between tense and lax stops are significant and unidirectional across place, session, and context. P-OV is lower in reliability, since in one combination of factors (word-initial velar stops of the first session) tense/lax differences are nonsignificant. Likewise, one nonsignificant case is found for the parameter EG-OV, which is the same as for P-OV. P-R shows nonsignificant or heterodirectional instances in two combinations of factors (first and second session of intervocalic labial stops). Next in reliability is OG-OC with two nonsignificant and one heterodirectional case all involving velar stops. Lowest in reliability is Clos that basically shows reliable differences only for intervocalic position. Turning to fricatives, the parameters Total, Gmax, and P-OV are maximally reliable phonation correlates. The remaining parameters come out as basically nonreliable, since in two out of four possible combinations of place and session tense/lax differences are nonsignificant or counter to the dominant directionality. Evaluating the complete set of obstruents, the single parameter that is maximally reliable across stops and fricatives is Gmax. Thus tense obstruents are reliably produced with a larger maximum glottal opening than lax obstruents in German according to the present results.

Evaluating place of articulation for stops, none of the parameters is a maximally reliable correlate in the sense defined above. Most reliability is achieved by Asp and P-R on the one hand, and Clos and Gmax on the other. For Asp and P-R results are significant in all conditions, but differ in directionality across contexts. Clos and Gmax are nonsignificant for certain place comparisons, but are unidirectional throughout. The other parameters are less reliable, since they involve several nonsignificant cases and reversals of directionality. Among fricatives, Total and P-OC is maximally reliable. The rest of the parameters are of lower reliability, because each of them is significant in only one session, although all parameters are unidirectional across sessions.

4. DISCUSSION

The results show that the tense obstruents /p,t,k,f,s/ of German are produced with a significantly larger maximum degree of glottal opening than the lax obstruents /b,d,g,v,z/. German, like many other Germanic languages (but not Dutch) distinguishes tense from lax stops in terms of aspiration. Assuming that aspiration is to a large extent caused by a widely opened glottis [5], it is expected correctly that tense stops in German show a much wider opened glottis than lax ones, which is also reported for other aspiration-based languages (cf. [3] for references). Other than the degree of glottal opening, aspirated and unaspirated stops differ also in terms of oral-laryngeal coordination, most strikingly in terms of P-OC. The relation between this parameter and aspiration has also been shown for Swedish [6], independent of the fact that aspiration investigated in [6] was allophonic (depending on stress), whereas it is phonemic (expressing the tense/lax opposition) here.

The question arises whether the small glottal opening found for most of the lax obstruents is produced actively, as claimed by some authors for Danish and Icelandic (cf. [3]), or whether it results passively from an increase in oral air pressure [1]. The passive account is supported by the fact that lax fricatives, which are expected to have lower oral air pressure than lax stops, have been produced more often completely lacking glottal opening than lax stops in this study.

This study can confirm earlier results for German [4] and English [2] that in the articulatory expression of place of articulation factors of oral-laryngeal coordination, most specifically P-R, are of particular importance. A comparison of the parameters Asp and P-R in stops is revealing in this respect. For both parameters the directionality is vel. > alv. > lab. word-initially and alv. > vel.> lab. intervocalically, suggesting a close relationship between P-R and Asp in the expression of place of articulation.

REFERENCES

[1] Bickley, C.A. & Stevens, K.N. 1987. Effects of a vocal tract constriction on the glottal source: data from voiced consonants. In Baer, T., Sasaki, C. & Harris, K.S. (eds.) *Laryngeal function in phonation and respiration*. Boston etc.: College-Hill Press. 239-253.

[2] Cooper, A.M. 1991. Laryngeal and oral gestures in English /p,t,k/. *PICPS* 12,2:50-53.

[3] Dixit, R.P. 1989. Glottal gestures in Hindi plosives. *Journal of Phonetics* 17: 213-237.

[4] Hoole, P., Pompino-Marschall, B. & Dames, M. 1984. Glottal timing in German voiceless obstruents. *PICPS* 10,2b: 399-403.

[5] Kim, C.-W. 1970. A theory of aspiration. *Phonetica* 21: 107-116.

[6] Löfqvist, A. 1980. Interarticulator programming in stop production. *Journal of Phonetics* 8: 475-490.

[7] Munhall, K.G., Löfqvist, A., & Kelso, S. 1994. Lip-larynx coordination in speech: effects of mechanical perturbations to the lower lip. *JASA* 95, 6: 3605-3616.

The study was supported by Grant DC-00865 from the National Institute on Deafness and Other Communication Disorders to Haskins Laboratories. Thanks to Anders Löfqvist for assistance with the transillumination experiment.

EFFECTS OF PROSODIC POSITION ON /t,d/ TONGUE/PALATE CONTACT

Patricia A. Keating
Phonetics Lab, Linguistics Department,
University of California, Los Angeles, USA

ABSTRACT

Two experiments tested the effects of lexical stress, phrasal stress, and position in utterance on the epg contact of onset and coda /t/ and /d/. Onsets have more contact than codas; utterance-initial onsets have by far the most contact. Effects of stress are variable and effects of accent are not significant.

INTRODUCTION

Several previous studies have shown that English non-continuant consonants in syllable- or word-initial position (onsets) generally have larger oral gestures and more oral contact than they do in syllable- or word-final position (codas) (e.g. [1], [2], [3], [4], [5]). Descriptively, this difference can be called *coda weakening*. Coda weakening is not due to some simple left-to-right weakening in an utterance: even codas which precede onsets have less contact [3]. It is not the same thing as American English flapping: coda weakening of alveolars is less extreme than flapping and occurs in non-flapping contexts [5]. It affects stops but not the fricative /s/ [3]. Aside from these observations, however, not much is known about the generality of coda weakening, in particular, whether coda weakening is limited to, or is enhanced by, particular prosodic positions. Prosodic here refers both to stress/ prominence and to phrasal groupings of various sizes. The two experiments reported here consider the effects of lexical stress, phrasal stress (sentence stress with a pitch accent), and phrasal position on word-initial and word-final /t/ and /d/. Additional data will be reported at the Congress.

The reason to consider the effect of lexical and phrasal stress on coda weakening is that they all appear to involve effects on degree of stricture. DeJong [6] shows that English stress results in a hyperarticulation, or strenthening, of segmental contrasts in the stressed syllable. It seems to be assumed that this hyperarticulation applies to all segments in the syllable, codas as well as onsets. It is possible, though, that because codas are generally weakened they would not be subject to the contradictory effect of prosodic strengthening. More subtle relations between the two are also possible. Expt. I was designed to test the effects of lexical and phrasal stress on onsets and codas.

The reason to consider the effect of phrasal position on coda weakening is that we know that glottal articulations are sensitive to phrasal position. Glottal opening associated with /h/ and with aspiration increases in magnitude word-initially and phrase-initially ([2],[7],[8]), just as it increases in magnitude with stress ([7]). We could expect initial oral articulations to pattern similarly. Since word-final aspiration is relatively rare [9], and glottal opening is reduced [2] word-finally, we could expect coda oral articulations to be reduced, rather than strenthened, word-finally. At the same time, we do not know much about how coda consonants are affected by prosodic groupings above the word. Expt. II was designed to test the effect of position in utterance on onsets vs. codas. Taken together, the two experiments explore how these three potentially different effects on degree of stricture (coda weakening, stress, phrasal position) interact.

METHOD

Equipment

The data in these experiments comes from electropalatography (EPG). EPG contact shows the net effect of jaw and tongue position on degree of stricture. The Kay Elemetrics Palatometer uses custom pseudopalates embedded with 96 contact electrodes to measure contact patterns over the hard palate and the inner surface of the molars. The EPG sampling interval is 10 ms, with the Palatometer taking 1.7 ms to complete a single sweep of the 96 electrodes.

Subjects

Subjects included in the present report were two middle-aged women, one a native speaker of Californian English and the other a native speaker of a mid-Atlantic dialect (the author). Because only two subjects have been tested, the results reported here may be viewed as preliminary.

Experimental materials

The test sentences were composed of real English words. In Expt. I, the variables under study were position of the consonant in the word (onset vs. coda), whether the consonant's syllable had lexical stress (stressed vs. unstressed), and whether the test word had sentence stress (nuclear pitch accent). The consonant /t/ occurred at the beginning or end of the test words, which were: *timid* (t is initial and stressed), *timidity* (t is initial and unstressed), *limit* (t is final and unstressed), *emit* (t is final and stressed). Each word appeared in the sentence "I wonder if (word) means anything". Phrasal stress was varied so that in half the sentences the test word had the nuclear accent of the sentence: "I wonder if EMIT means anything"; in the other half the nuclear accent was on "anything": "I wonder if emit means ANYTHING". No instructions were given about sentence accents other than the nuclear accent.

In Expt. II, /d/ occurred at the beginning or end of the test words, which were all content words of one or two syllables. Since the monosyllabic words had a lexical stress, the consonants in those words were in the lexically-stressed syllable. The disyllabic words had lexical stress on the syllable which did not contain the test consonant. Thus the test words were *deaf* (d is initial and stressed), *fed* (d is final and stressed), *demand* (d is initial and unstressed), *aphid* (d is final and unstressed). Thus the stressed /d/s come from monosyllables and the unstressed /d/s come from disyllables, but they will be referred to simply as stressed vs. unstressed.

These 4 words were put into sentences in three different positions: initial, medial, and final. I will refer to initial, medial, and final onsets and codas even though when a word with a coda /d/ is utterance-initial, the /d/ itself is of course not initial in the utterance, and so on. The segmental contexts were kept very similar across utterance conditions. Absolute initial and absolute final positions might have special articulations due to a neutral or rest position. Therefore the test sentences were preceded or followed by extra words, though the subjects were instructed to produce the test sentences as separate utterances. The test sentences for *deaf* were: *(up.) DEAF bugs go out.*
Pick up DEAF bugs now.
Pick them up DEAF. (Bugs.)
In this experiment, the test word always had a pitch accent, which was the nuclear, and only, accent of the sentence.

Procedure

Subjects read the test sentences from a printed sheet, 8 times each for Expt. I and 9 times each for Expt. II, in a different order each time.

Analysis

The percent of all electrodes that were contacted was measured for each token at the point of greatest contact for the consonant. These measures were then analyzed by ANOVA. Because the results for the two subjects were somewhat different, only individual-subject analyses will be discussed here.

RESULTS

Experiment I

ANOVA showed that the two subjects shared only one significant effect, the main effect for onset vs. coda /t/. Both subjects showed some effects of both lexical stress and accent, but in different ways.

Both subjects had some kind of significant effect of lexical stress on EPG contact. For subject P there was a main effect of stress: stressed consonants had significantly *less* contact than stressless consonants (Fig. 1). For subject B there was an interaction of stress with syllable position: onsets had *more* contact when stressed; coda contact showed no effect of stress. Furthermore, for this subject, onsets differed from codas only in stressed syllables (Fig. 2).

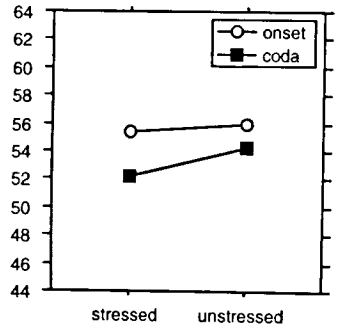


Figure 1. Percent contact for /t/, subject P: position in syllable x stress

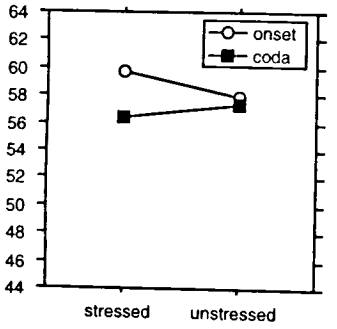


Figure 2. Percent contact for /t/, subject B: position in syllable x stress

Both subjects also had some non-significant effect of accent on EPG contact. For subject B, accent had a tendency ($p<.08$) to result in more contact for both onsets and codas. For subject P, there was a non-significant tendency ($p=.06$) for accent and syllable position to interact. The onset/coda difference was much stronger in accented than in unaccented syllables, and the effect of accent was seen only in codas, where accented codas had less contact than unaccented.

In sum, the only consistent effect for the two subjects was that onsets have more contact than codas in stressed syllables. Lexical stress clearly affects EPG contact too, but in different directions for the two subjects. Only for one subject did it affect onsets differently from codas. The effect of accent, in contrast, was only a trend for each subject.

Experiment II

For this experiment, statistically-reliable results will be presented only for subject P. However, the results appear similar for the other subject, with one exception to be noted. In Expt. II, accent was not varied, but lexical stress was again varied, in addition to the new variable of interest, position in utterance, all for the consonant /d/. All factors gave significant main effects for subject P. First, as before, onsets have more contact than codas. Second, with respect to stress, the result was different from before: stressed consonants, whether onset or coda, have more contact. Finally, with position in utterance, for subject P all three positions are significantly different from one another: initials have the most contact, finals the next, and medials the least.

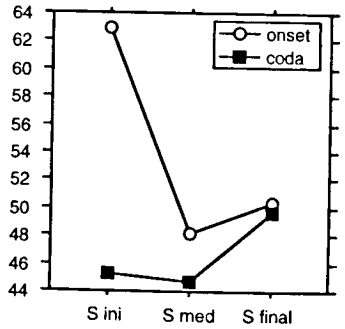


Figure 3. Percent contact for /d/, subject P: position in syllable x position in utterance

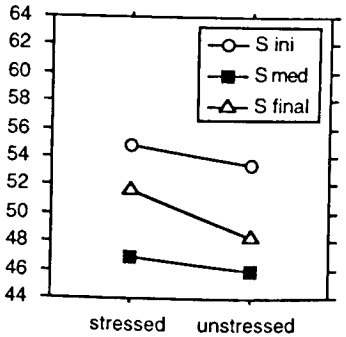


Figure 4. Percent contact for /d/, subject P: stress x position in sentence

This effect is fairly specific: there is a large increase in contact for utterance-initial onsets and a smaller boost of utterance-final codas. (Subject B appears to have no such coda boost.) As a result, the onset/coda asymmetry is much stronger utterance-initially than utterance-medially (Fig. 3). Furthermore, stress and position interact: the effect of lexical stress (which is to increase contact) is strongest in utterance-final position (Fig. 4).

CONCLUSIONS

These results suggest that "coda weakening" is a fairly robust phenomenon; indeed, it is the only consistent result of the two experiments. However, it does not occur all the time: for one subject it is not observed utterance-finally, and for the other it is not observed in unstressed syllables.

In these data, we see first that lexical stress and phrasal accent are not the same; lexical stress has significant effects on consonant contact while phrasal accent has only weak effects. Lexical stress for one subject operated in a way that is difficult to understand, because Expts. I and II showed opposite effects. Possibly the results of Expt. I were contaminated by an asymmetry in the carrier sentence, which had "if" (with [I]) before onsets but "means" (with [i]) after codas. Under an intricate scenario of how words overlap, it might be possible to explain the stress effect in codas as an artifact of the experiment. Expt. II used a more symmetrical carrier frame, and in this experiment the results were in accord with one of the predicted outcomes: stress resulted in more consonant contact in both onsets and codas, even against the fact that the stressed vowels in these test words were more open. Here, stress and coda weakening are independent effects, with the hyperarticulation of stress affecting both consonants about the same.

On the other hand, lexical stress for the other subject increased contact only in onsets, and only in stressed syllables was the onset/coda asymmetry observed. Possibly this result is also due to the asymmetry of the frame. However, for this subject, it is also entirely possible that the onset/coda difference is due to an asymmetry in the realization of stress: that stress causes hyperarticulation in CV but not VC within the stressed syllable. Statistically-reliable data from Expt. II will be needed to distinguish these interpretations for this subject.

With respect to phrasal position, we see in Expt. II that lingual contact is greatest adjacent to an utterance boundary. Most notably, onsets are strongest utterance-initially, with more contact than any other consonants. For one subject, codas are strongest utterance-finally, though the strengthening of onsets utterance-initially is much greater in magnitude than the strengthening of codas utterance-finally. As a result of this strengthening, codas make up somewhat for their overall relative weakness, so that these strongest codas have about the same contact as a weak onset for this speaker. For the other subject, no final coda strengthening is seen. Clearly, to the extent coda strengthening occurs, it is a small effect.

REFERENCES

[1] Krakow, R. A (1989), *The articulatory organization of syllables: a kinematic analysis of labial and velic gestures*, PhD dissertation, Yale U.

[2] Browman, C. & L Goldstein (1992), "Articulatory phonology: an overview", *Phonetica* 49: 155-180.

[3] Byrd, D. (1994), *Articulatory timing in English consonant sequences. UCLA Working Papers in Phonetics 86.*

[4] Keating, P. A. and R. Wright (1994), "Effects of position-in-syllable on consonant articulation and acoustics" (abstract), *JASA* 95(5.2):2819.

[5] Wright, R, (1994), "Coda lenition in American English consonants: an EPG study" *JASA* 95(5.2),:2819.

[6] DeJong, K. (1995), "The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation", *JASA* 97(1):491-504.

[7] Pierrehumbert, J. & D. Talkin (1992), "Lenition of /h/ and glottal stop", *Papers in Laboratory Phonology II* (eds. Docherty & Ladd):90-116.

[8] Jun, S. (1993), *The phonetics and phonology of Korean Prosody*, PhD dissertation, Ohio State U.

[9] Keating, P., M. Huffman, and W. Linker (1983), Patterns in allophone distribution for voiced and voiceless stops, *J. Phonetics 11*:277-290.

CROSS-LINGUISTIC ASPECTS OF COARTICULATION: AN ACOUSTIC AND ELECTROPALATOGRAPHIC STUDY OF DENTAL AND RETROFLEX CONSONANTS

Krull D[1], Lindblom B[1], Shia B-E[2] and Fruchter D[2]
[1]Department of Linguistics, Stockholm University, Stockholm S-10691, Sweden
[2]Department of Linguistics, University of Texas at Austin, 78712 Austin, Texas, USA

ABSTRACT

In Hindi, Swedish and Tamil retroflexes show a more posterior articulation at the beginning of the closure than at the release. Also their exact place of closure is vowel-dependent, whereas that of dentals is constant. Locus equation parameters provide a clear basis for separating dentals from retroflexes at closure onset, but fail to support the idea that *degree of vowel-consonant coarticulation* varies with place, and/or is language-dependent.

MEASUREMENTS

The present data come from two speakers of Hindi, two speakers of Swedish and two speakers of Tamil. Electropalatographic and acoustic records were obtained from all of them. The Hindi and Tamil speakers were asked to produce isolated words of either [a'CV] or ['VCa] structure with V = /i/, /e/, /a/, /o/ or /u/. The Swedish utterances, prosodically similar to *gul hatt*, had symmetrical ['V:'CV:] structure. The present findings are based on measurements for V = /i/, /e/, /ɛ/, /a/, /o/ or /u/. The test words were read five to six times. Formant estimates were made from spectrograms and short-term spectra using the MIX software of R Carlson (KTH). For all samples, formant frequencies were measured in the first vowel 80 ms before the VC boundary; at the last glottal pulse of first vowel before closure (=VC boundary); at the burst; at the first glottal pulse of second vowel (=CV boundary); and in the second vowel 80 ms after the CV boundary. The EPG data were collected using the Reading system [1].

RESULTS

Figure 1 compares several aspects of the data. Average values are shown from a Swedish speaker (OE, left two columns) and for a Hindi speaker (RM, right two columns). For dentals the place of articulation is at the first row of EPG electrodes. There is no variation with vowel context. Nor is there a change from the VC to the CV condition. However, the retroflex data differs in that the exact place does indeed depend on the vowel front vowels having more anterior variants of retroflection. Also there are marked differences between the VC and the CV samples: During the closure the place of contact slides forward so that the contrast between dental and retroflex is larger at the VC than at the CV boundary. The retroflex data for the other speakers exhibit similar patterns of vowel dependence and closure displacement. These findings confirm previous findings on Hindi [2]. The formant measurements are in the form of "locus" plots with the F2 and F3 onsets-offsets plotted against the F2 in the adjacent vowel (first vowel for VC, second for the CV). There is a marked difference between the lines fitted to the F3 data. For dentals, horizontal patterns prevail. For retroflexes, lines are steeper roughly parallel to those of F2.

A comparison of slopes and intercepts for locus equations fitted to the F2 data for each speaker individually reveals no

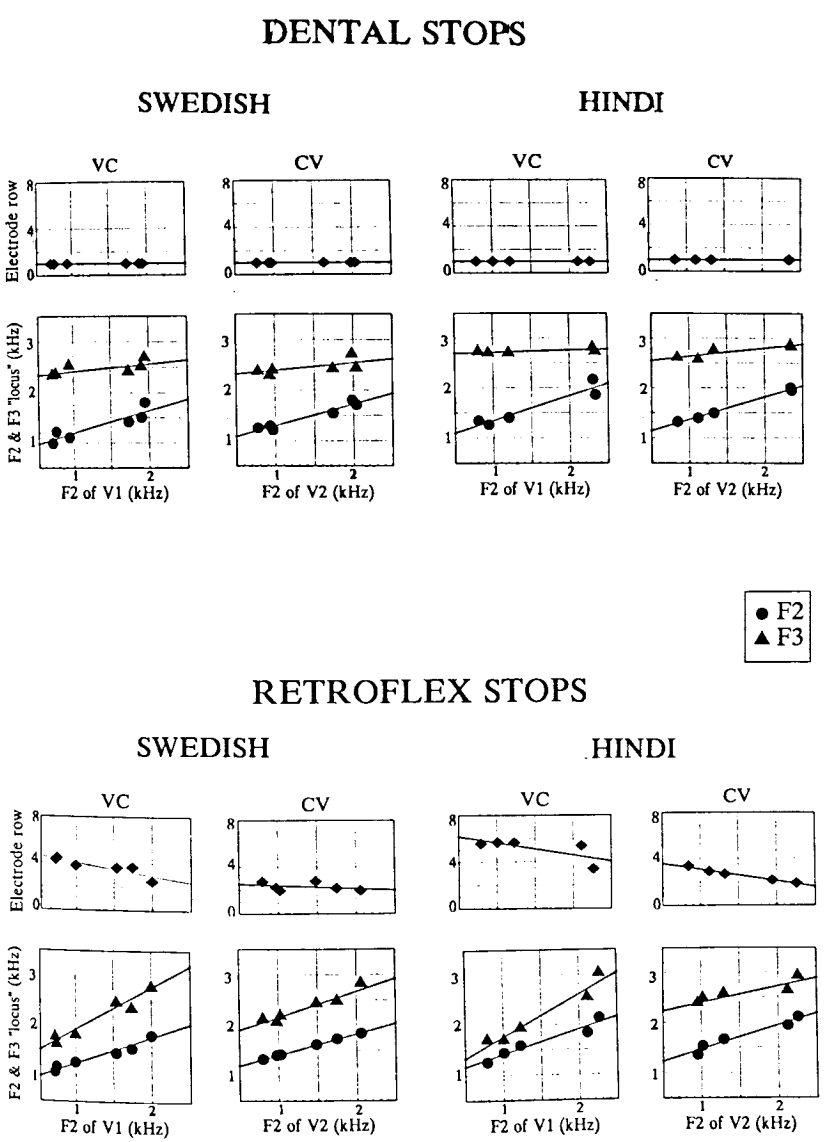


Figure 1. The top half of the figure pertains to dental, the lower half to retroflex observations. The column to the left shows measurements made at the VC boundary, to the right CV observations. The smaller panels (EPG) indicate place of stop closure (specified by electrode number) and are paired with a plot of F2 and F3 onset (VC or CV) versus F2 in vowel (first or second). The calibration of the x-axis is F2 of first or second vowel.

major differences between dentals and retroflexes. This is in agreement with the results reported by Sussman et al [3]. Since, theoretically, F2 ought to be associated mainly with the cavity behind the closure, this finding implies that dentals and retroflexes invoke similar coarticulation patterns with respect to underlying tongue body configurations. Consequently, a more posterior retroflection does not necessarily presuppose a tongue body which is also more posterior. Sublaminal articulations are allegedly typical of Tamil retroflexes [4]. They involve the tongue underside and might therefore be assumed to constrain the mobility of the tongue body even more severely than laminal retroflexes and dentals. Krull [5] has suggested that, for a given place of articulation, variations in the slope andintercept of locus equations could be seen as variations in *degree of coarticulation.*

However the present investigation provides no basis for identifying significant differences in slopes and intercept values in the F2 of Swedish and Hindi laminal retroflexes and Tamil sublaminal retroflexes.

A partial summary of the locus equation results is presented in Figure 2: The diagrams pertain to F3 at the VC and at the CV boundary.

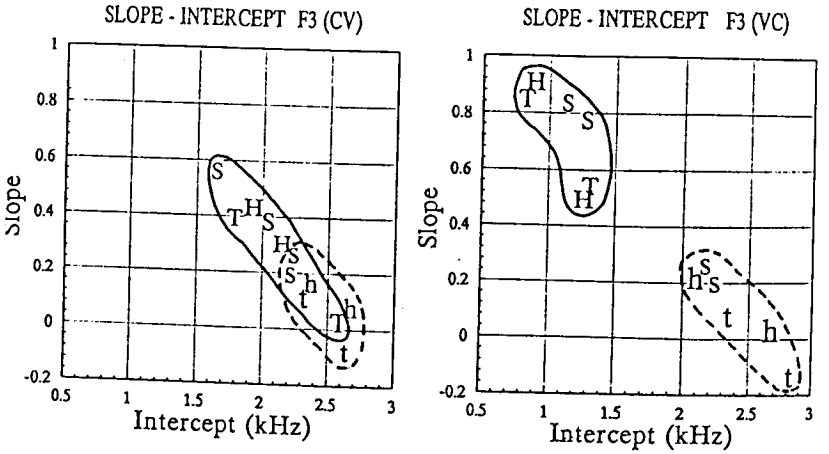


Figure 2. Locus equation slopes and intercepts for F3 at the VC boundary (right) and CV boundary (left). Data for all speakers indicated by language name initial. Uppercase stands for retroflexes, lowercase for dental articulations.

Here slopes are plotted against intercepts for all speakers. Dentals and retroflexes form clearly separated clusters in the case of VC, but they are less distinct for CV. There is no clear-cut or systematic grouping of the speaker pairs which leads us to conclude that coarticulation patterns are similar in the six speakers investigated.

CONCLUSIONS

1. With respect to the time course of retroflex production, it was found that, in all three languages, retroflexes showed a more posterior articulation at the beginning of the closure than at the release. Also their exact place of closure is vowel-dependent, whereas that of dentals is constant (Figure 1).

2. Do locus equation parameters provide invariant place correlates separating dentals and retroflexes in a vowel-independent manner [6]? The answer based on the present data is no with respect to F2, but yes in terms of F3 (Figures 1 and 2).

3. Is there evidence of a more restricted tongue body variation (less coarticulation) in retroflexes than in dentals? On the basis of considerations of articulatory synergy [7], it might be assumed that a more posterior retroflection would create a preference for a tongue body which is also more posterior. Applying the reasoning of Krull [5], we conclude that, for the present analyses, *degree of coarticulation* does not seem to be less in retroflexes. This conclusion is based on the fact that the slopes and intercepts of F2 locus equations were found to be remarkably similar in comparisons of each speaker's dentals and retroflexes.

4. We were unable to identify reliable articulatory or acoustic evidence for the laminal variant of retroflection supposedly characteristic of Hindi and Swedish as opposed to the sublaminal articulation of Tamil [4]. That may be due to the need for more fine-grained analyses than the ones undertaken so far.

REFERENCES

[1] Engstrand O (1989): "Toward an electropalatographic specification of consonant articulation in Swedish", 115-156 in *Perilus X,* Department of Linguistics, University of Stockholm.

[2] Dixit R P (1990): "Linguotectal contact patterns in the dental and retroflex stops of Hindi", *J of Phonetics 18*:189-201.

[3] Sussman H M, Hoemeke K A and Farhan S A (1993): "A cross-linguistic investigation of locus equations as a phonetic descriptor for place of articulation", *J Acoust Soc Am* 94(3):1256-1268.

[4] Ladefoged P and Bhaskararao P (1983): "Non-quantal aspects of consonant production", *J of Phonetics 11*:291-302.

[5] Krull D (1988): *Acoustic properties as predictors of perceptual responses: A study of Swedish voiced stops,* doctoral dissertation, *Perilus VII,* Stockholm University.

[6] Sussman H M, McCaffrey H A and Matthews S A (1991): "An investigation of locus equations as a source of relational invariance for stop place categorization", *J Acoust Soc Am* 90:1309-1325.

[7] Lindblom B, Pauli S and Sundberg J (1975): "Modeling coarticulation in apical stops", in Fant G (ed): *Proceedings of SCS-74,* Stockholm, Almkvist&Wiksell.

ACKNOWLEDGEMENTS

This research was supported by HSFR of Sweden (project APEX).

The authors are grateful to Ivar Wäneland for the odontological work needed to produce the EPG palates, and to Agneta Lundgren for her help with the data processing.

A CROSSLINGUISTIC VOT ANALYSIS OF EARLY STOP PRODUCTION: THE DEVELOPMENT OF THE FEATURE VOICING

P. Larrañaga, C. Lleó and M. Prinz
University of Hamburg, Germany

ABSTRACT

This paper focuses on the development of the feature voicing and its VOT implementation. The VOT values of early stops produced by German and Spanish monolingual children were analyzed and compared. A language-specific difference of the mean VOT values was already found at the babbling stage. Moreover, in early German word production significant voicing contrasts for coronals and labials were found. The Spanish data provided no statistically significant VOT differences.

INTRODUCTION

It has been proposed that at the phonological level voicing is a binary feature, phonetically implemented by means of two out of three discrete categories of the VOT continuum: voiced (=lead), voiceless unaspirated (=short lag) and voiceless aspirated (=long lag), depending on the particular language; and that, at the phonetic level, short lag is the unmarked option, whereas lead and long lag are marked [1]. This entails that voiceless unaspirated will always implement one of the terms of the phonological voicing contrast. In this respect, research on early child stop production provides relevant data to test the particular hypotheses, although the available literature - [2], [3], [4], [5], [6] and [7] among others - does not present a unitary picture and leaves many questions open.

This paper is concerned with the acquisition of the voicing contrast and with its implementation by means of VOT. It deals with some of those questions on the acquisition of VOT which current research has left partly open: When do the VOT values begin to constitute a contrast, and when is a voicing contrast of the target language acquired? Do stop consonants at the babbling stage already manifest a tendency to the VOT values of the target language? Do babbling data and early words exclusively contain voiceless unaspirated stops?

In order to find an answer to these questions, data on early stop production by four German and four Spanish monolingual children were selected for VOT analysis. The target languages German and Spanish were chosen because of their opposing implementation of the feature voicing: German, like English, implements voiceless by means of long lag and it implements voiced by means of short lag, whereas in Spanish voiceless corresponds to short lag and voiced to lead.

METHODS

The data reported here belong to a larger longitudinal investigation of five children acquiring German in Hamburg and four children acquiring Spanish in Madrid. Both groups were audio-recorded at their homes in unstructured play sessions, using a high-fidelity Sony TCD-D10 PRO cassette recorder and a portable Beyerdynamic microphone. Due to the nature of the recordings as unstructured play sessions, the collected data are heterogeneous, and the number of tokens for analyzable stops varies considerably within sessions and children. Many relevant data could not be taken into consideration, because of disturbing noise; this was especially true for the Spanish data, which had thus to be extensibly reduced.

Utterances beginning with stops (voiced and voiceless) in initial stressed CV sylables were selected for analysis, corresponding to three developmental points: babbling, 25 word point and 100 word point. The material for the word stage was classified according to the intended target consonant and not according to the produced sound; a stop was further considered for analysis only if it had the same place of articulation as the intended target consonant. In the babbling stage, no target model being available, the classification was made according to the produced sound. Stops selected for analysis belonged to the three places of articulation: labial, coronal and velar. The relevant speech signals were digitalized at a 22KHz rate from a Revox-B215 tape recorder. We used both wide band (300 Hz) and narrow band (59 Hz) FFT, with 6 dB pre-emphasis and 0.5 frame advance for the acoustic analysis. The acoustic analysis was made with a computer-implemented program (Sound Scope 16 for Macintosh). Two windows were used: the time signal in the lower window and the power spectrum in the upper window. Time as well as frequency signals were taken into account to calculate the values. If the time signal was not very clear we additionally measured the power spectrum (both narrow and wideband) and compared the results with the time signal values. Only if the results were similar, we took them into account.

To test statistical significance, variance measures (two-tailed student-t tests) were conducted on the VOT scores for the various places of articulation within each language group and across languages, at the three developmental points.

RESULTS AND DISCUSSION

Our results confirm some of the findings of other researchers, but bring some new points to light. At the babbling stage, German children made a statistically significant difference between two coronal categories, which were perceived as voiced and voiceless, respectively ($p = .001$). No other relevant differences could be ascertained neither in the German nor in the Spanish data.

A crosslinguistic comparison was only possible at the coronal articulation place. VOT values for coronals in both language groups cluster around the short lag and (short) lead, but a significant difference between the mean VOT values of the two language groups was found ($p =$

.0369), German children being more at the short lag side and Spanish children at the (short) lead side. Even though lead voicing predominates only in the Spanish data, it has to be emphasized that it is manifested in both language groups. As expected, in German lead voicing decreases in the course of time, i.e., at the early word stage, only two children manifest a preference for lead voicing in the production of target voiced stops. This result is only comparable to [5], who found more voicing lead than other researchers, i.e., [4].

At the 25 word point German children produce significant voicing contrasts for coronals ($p = .0118$), whereas there is no contrast for labials ($p = .4991$) nor for dorsals ($p = .5$). At the 100 word point, there is a significant contrast for coronals ($p = .0325$) as well as for labials ($p = .00592$). The values for target voiced and target voiceless velars are not significantly different at any of the two word points. At both word points, the categories implementing the voicing contrasts are different from those of the target language. Thus the mean values for voiced stops are slightly under null at the three articulation places and the values for the voiceless stops are within the short lag domain, showing the expected progression from a shorter lag for labials, slightly longer for coronals and definitely longer for dorsals, especially at the 100 word point.

The Spanish data provide no statistically significant differences of VOT, neither at the 25 word point nor at the 100 word point. This agrees with the findings of [4]: at the early word stage, VOT does not play a significant role in Spanish. In our data, this finding is also due to the scarcity of the data: very few target voiced stops fulfilled the criteria for analysis. As regards the target voiceless, their values lie in the short lag range, also showing the expected progression from a (short) lead for labials to a short lag for coronals and a slightly longer lag for velars.

A crosslinguistic comparison of the word data gives the following results. At the 25 word point stops corresponding to the target voiceless have a significantly higher VOT in German than in Spanish at the labial place ($p = .0253$). And at the 100 word point a slight tendency to significancy is manifested at the coronal point of articulation ($p = .115$).

These results agree with the expectations only insofar as the first phonetically implemented stops are in the domain of short lag. Unexpectedly, the first contrast produced in German appears between short lag and (short) lead, regardless of the target language distinction between short lag and long lag. In fact, according to at least part of the literature [6, 7], the first expected opposition should be established between short lag and long lag.

CONCLUSION

The data presented in this paper have shown that the short lag option is by no means the only category present at the babbling stage nor at the early word stage. Short lag and short lead were both present at this early stage in both languages, although only the target language Spanish contains lead voicing. Obviously, the VOT values do not match those of the target languages. This is especially true for the German data, the target language not implementing lead voice. Assuming that at the early stages of language acquisition unmarked options are chosen, the present results call into question the exclusive unmarked status of short lag. Lead voicing, although at the short range, seems to constitute an unmarked category as well, already present at the babbling stage. Interestingly, the first voicing opposition at the 25 word point is made between short lag and (short) lead, in spite of the fact that the target language opposes short lag to long lag.

As to the question when children start making a contrast between a voiced and a voiceless series of stops, no significant opposition could be found in Spanish. In the German data, a significant contrast was already found for coronals at the 25 word point and for labials at the 100 word point. Furthermore, at the babbling stage, German children made a statistically significant difference between two coronal categories, perceived as voiced and voiceless, respectively.

ACKNOWLEDGEMENT

The project on which this research is based has been supported by a grant of the Deutsche Forschungsgemeinschaft. We want to express our gratitude to this institution, to the children of the project and to their mothers.

REFERENCES

[1] Keating, P.A. (1984), Phonetic and phonological representation of stop consonant voicing. *Language*, vol. 60, pp. 286-319.

[2] Jakobson, R. (1941), *Kindersprache, Aphasie und allgemeine Lautgesetze*. Uppsala: Almqvist and Wiksell.

[3] Kewley-Port, D. & M.S. Preston (1974), Early apical stop production: a voice onset time analysis. *Journal of Phonetics*, vol. 2, pp. 195-210.

[4] Macken, M.A. & D. Barton (1980), The acquisition of the voicing contrast in Spanish: a phonetic and phonological study of word-initial stop consonants. *Journal of Child Language*, vol. 7, pp. 433-458.

[5] Eilers, R.E., D.K. Oller & C.R. Benito-García (1984), The acquisition of voicing contrasts in Spanish and English learning infants and children: a longitudinal study. *Journal of Child Language*, vol. 11, pp. 313-336.

[6] Deuchar, M. & A. Clark (1992), Bilingual acquisition of the voicing contrast in word-initial stop consonants in English and Spanish. *Cognitive Science Research Paper, Serial No. CSRP 213*. University of Sussex.

[7] Macken, M.A. & D. 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.

STOP CONSONANT PRODUCTION: AN ARTICULATION AND ACOUSTIC STUDY

Kelly L. Poort
Massachusetts Institute of Technology, Cambridge, MA USA

ABSTRACT

Articulation and acoustic data for stop consonant production were examined with the aims of (1) describing the movements and coordination of the articulatory structures and (2) developing procedures for interpreting acoustic data in terms of articulatory movements. Findings were placed in the context of existing acoustic and aerodynamic models.

INTRODUCTION

Articulatory information for the stop consonants, derived from recordings of the physical movements of the articulators, was combined with simultaneous acoustic recordings. The voiceless stop consonants were chosen for detailed study because there is less acoustic information present immediately following the release than for voiced stops. In particular, the labial and alveolar voiceless stops were chosen for in-depth investigation. This paper focuses on the production of the labial voiceless stop consonant /p/. The production of the alveolar stop /t/ will be discussed in the presentation in August. The coupling of articulatory and acoustic information provides an improved understanding of stop-consonant production, leading to knowledge of the sequencing and timing of articulator movements as well as the resultant acoustics. The primary objective of the study is to refine existing acoustic and aerodynamic models to reflect the new level of understanding. In addition, examination of the articulatory information leads to improved interpretation of the acoustic signal, such that in the future the acoustic waveform will be the only information required to determine many of the important aspects of the vocal-tract movements for stops. The results of the investigation are applicable to the areas of speech recognition, speech synthesis, and the study and remediation of disordered speech production.

PROCEDURE

Movements in the midsagittal plane of points on the lower jaw, lips, tongue blade and tongue body were measured using an electromagnetic midsagittal articulometer [1]. Acoustic data were recorded simultaneously. Three normal speaking, normal hearing male subjects spoke single-syllable words /CVt/, composed of one of the voiceless stop consonants /p, t/ followed by one of the vowels /a, i/ and the consonant /t/, imbedded in a carrier phrase. Half the tokens were preceded by the fricative consonant /s/. One example is, "Say spot again." A minimum of three repetitions of each utterance were recorded per speaker.

Events in time, such as the end of the vowel in say, the end of /s/ in the token (if present), the stop release, and the onset of the vowel following the stop, were identified in the acoustic waveform based upon researcher's judgment. The corresponding times were located in the time-aligned articulatory displacement waveforms. In order to preserve the times and magnitudes of each of these event times in the articulation data, as well as the general shape of the displacement waveform between event times, the standard technique of linear time warping was adapted and applied. The repetitions of each utterance for a given speaker were averaged together using the adapted technique. For example, the eight repetitions of "Say spot again." recorded by one of the speakers were averaged together using the modified linear time warping technique to become one, representative displacement waveform for that speaker.

An estimate of the constriction cross-sectional area change with time following the release of the stop consonant was calculated with the aid of the articulation data. The upper and lower lip transducers are located on the vermilion borders of the lips. The movements of these two transducers, averaged as described above, were used to determine rate of vertical lip separation following release. The difference between the upper and lower lip trajectories was used as an estimate of the rate of lip separation, with the difference at the time of release zeroed. The rate of horizontal lip separation was taken from a study by Fujimura [2]. An estimate of constriction cross-sectional area change with time following stop release was obtained by approximating the lip opening cross-sectional area as a rectangle whose height and width change with time according to the vertical and horizontal lip separations, respectively. For example, the cross-sectional area of the lips in the labial /p/, derived in this manner for one subject speaking the token spot, increases by approximately 35 cm^2/sec within the first 5 - 10 msec following the stop release.

ACOUSTIC ANALYSIS

During the first few milliseconds following the stop consonant release, the constriction cross-sectional area change with time can be related to the corresponding acoustics in two important ways: (1) by modeling the vocal tract as a series of tubes of varying cross-sectional area (in the case of a labial stop, a Helmholtz resonator) and calculating the transition of the first formant frequency F1; and (2) by calculating the time course of the burst when the constriction cross-sectional area is used as a parameter in a circuit model of the vocal tract [3].

Fujimura, in a stroboscopic motion picture study [2], observed a three-stage transition in the first formant frequency F1 following labial stop release into a vowel. The first stage occurs during the 5 - 10 msec immediately after the lips begin to open. It consists of an abrupt increase in lip opening cross-sectional area corresponding to a rapid upward shift in F1. Within just the first 5 msec, F1 will shift from 0 Hz (assuming the vocal tract walls are rigid) to a value of 200 - 400 Hz, depending upon the following vowel. The rapid rise of F1 is modeled as a Helmholtz resonator. The short front tube represents the lips during the production of the bilabial stop consonant. The second stage consists of a slower rise in F1, corresponding to a continued increase in lip opening cross-sectional area in conjunction with a downward movement of the jaw. The third stage is the movement of F1 during the initial portion of the following vowel.

The present study has also investigated the F1 transition following the stop release. The constriction cross-sectional area change with time following release, as calculated from the articulation data, is utilized as the rate at which the cross-sectional area of the short front tube of the Helmholtz resonator increases with time. A correction term of 180 Hz has been incorporated into the model to reflect the impedance of the vocal tract walls. From the solution of the Helmholtz resonator, F1 transitions from 180 to 400 Hz during the first 5 - 6 msec following labial stop release in the utterance spot. This study finds that the rate of constriction cross-sectional area change with time immediately after lip opening is less than the 100 cm^2/sec determined by Fujimura. Consequently, F1 does not initially transition upward as rapidly as predicted by Fujimura [2].

An aerodynamic circuit model of the pressures and flows in the vocal tract [3] was employed to calculate the airflow through the lips for the first few milliseconds following the stop release. The derived constriction cross-sectional area was utilized in the determination of one of the parameters in the circuit model, specifically the rate of decrease in constriction resistance following the stop release. From the constriction cross-sectional area of the lip opening and the airflow through the lips immediately following release, the amplitude and duration of the frication noise source can be calculated [4, 5, 6, 7 and others]. The burst contains a peak in amplitude approximately 5 - 6 msec after the release. The duration of the noise burst up to a time where the amplitude is 10 dB down from the burst peak in spot is approximately 8 - 10 msec. Both the shape and location of the peak in the noise burst as well as the duration of the burst derived from the model in this fashion agree well with the shape and duration of the noise burst in the corresponding acoustic waveform for spot.

ARTICULATION ANALYSIS

In addition to the constriction cross-sectional area calculation discussed earlier, the articulation data were examined to determine effects of phonetic context on production. Findings include a constraint on lower jaw position during /s/ production, maximum downward velocity for the lower jaw occurring approximately at the time of the consonant release, and a correlation between increasing distance articulators must travel and faster rates of movement. In addition, the articulators not involved in forming the constriction were found to anticipate the positions required for the upcoming vowel much more so than the constriction-forming articulator(s). For example, the lips are constrained to form the constriction for /p/ in spot; however, the jaw and tongue can and do move to some extent into position for the following vowel during the production of /p/. In a similar fashion, the restriction on jaw movement during the production of /s/ forces the jaw to remain in a high position throughout the duration of /s/. As a result, there is a more rapid downward velocity of the jaw at the time of the /p/ release in spot than in pot. This finding is thought to be a compensating mechanism. In order to reach the low jaw position required for the production of the vowel /ɑ/ in spot in approximately the same amount of time as it takes to reach the /ɑ/ in pot, the jaw increases its downward velocity. Relating the finding to an earlier observation, the jaw must move downward faster for /p/ in spot because it travels farther.

CONCLUSION

One of the primary results of the study is the ability to obtain an estimate from the articulation data of the constriction cross-sectional area change with time for the first few milliseconds following the stop consonant release. For example, this rate for /p/ in spot is 35 cm^2/sec. An acoustic model representing the vocal tract as a Helmholtz resonator yields a rapid transition for F1 following the stop release of 35 - 45 Hz/msec for the same utterance. An aerodynamic model coupled with calculations of the frication noise burst reveal an agreement between calculated (utilizing the articulation data) and acoustic waveform noise burst shapes and durations. The shape is found to contain a peak approximately 5 - 6 msec after release and the duration is approximately 8 - 10 msec for /p/ in spot. This agreement suggests that the original estimate of the constriction cross-sectional area change with time of 35 cm^2/sec for /p/ in spot during the first few milliseconds following the stop consonant release is reasonable.

Inferences, such as those described above for /p/ in spot, made from detailed examination of the articulation and acoustic data will aid in developing a more comprehensive model of stop consonant production. From comparison of model outputs and acoustic data, refinements can be made to the existing aerodynamic and acoustic models to more accurately represent the acoustic signal. The quantitative variations in production resulting from various phonetic contexts can be incorporated into the models in order to broaden their applicability to essentially all stop consonant production. The findings of the study contribute to the goal of a single comprehensive model which incorporates all the acoustic, aerodynamic, and articulatory observations in order to explain the resultant acoustic output.

ACKNOWLEDGEMENT

My deepest appreciation to Professor Ken Stevens for his guidance and support. I also thank Joe Perkell, Melanie Matthies, and Mario Svirsky for the use of the electromagnetic midsagittal articulometer system and their technical assistance with my project. This research has been supported in part by grant DC00075 from the National Institutes of Health.

REFERENCES

[1] Perkell, J., et al. (1992), "Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements," *J. Acoust. Soc. Am.* 92(6), pp. 3078-3096.

[2] Fujimura (1961), "Bilabial Stop and Nasal Consonants: A Motion Picture Study and its Acoustical Implications," *J. Speech and Hear. Res.* 4(3), pp. 233-247.

[3] Stevens, K. (1993) "Models for the production and acoustics of stop consonants," *Speech Comm.* 13, pp. 367-375.

[4] Fant, G. (1960), *Acoustic theory of speech production*, The Hague: Mouton & Co.

[5] Stevens, K. (1971) "Airflow and turbulence noise for fricative and stop consonants," *J. Acoust. Soc. Am.* 50, pp. 1180-1192.

[6] Shadle, C. (1985) *The acoustics of fricative consonants*, RLE Technical Report 506, Massachusetts Institute of Technology, Cambridge, MA.

[7] Pastel, L. (1987) *Turbulent noise sources in vocal tract models*, SM Thesis, Massachusetts Institute of Technology, Cambridge, MA.

ASPIRATED STOPS IN SCOTS GAELIC

Henry Rogers
University of Toronto

ABSTRACT

Aspiration is the primary difference between the lenis and fortis stops in Scots Gaelic: postaspiration initially, and preaspiration medially and finally. With faster speech, the postaspirated stops show general shortening, and the preaspirated stops are shortened in the voiceless duration preserving the perceptual salience of the aspiration. The details of the aspiration and the shortening are viewed as controlled, language-specific behaviour.

INTRODUCTION

Recent research [1, 2] has emphasised the role of non-automatic, allophonic phonetic activity. This paper presents data on pre- and postaspirated stops in Scots Gaelic at different rates of speech and argues this aspiration is an example of such controlled, subphonemic activity.

In Scots Gaelic [3] the fortis stops /p t k/ have postaspiration [p^h t^h k^h] in initial position, and preaspiration [$^h p$ $^h t$ $^x k$] medially and finally. In the dialects analysed here, preaspiration before /k/ is realised as a velar fricative. The term 'fortis' is used for the phonemes /p t k/ and 'lenis' for the phonemes /b d g/; 'voiceless' and 'voiced' refer to activities of the vocal folds. The lenis stops are typically voiceless in all environments.

Two speakers read the material, consisting of 120 one- and two-syllable words in a frame of *Can X a nis* /kɑn X ə niʃ/ 'Say X now', four times at a normal speed, and then twice at a fast speed. Speaker RM is from Harris and FS from Lewis; both women have lived in Toronto for several years.

Preaspiration

Preaspirated stops have aspiration preceding the closure as opposed to postaspirated stops with aspiration following the release of the stop. This is a rather rare phenomenon in the world, reported primarily in Northern Europe (Icelandic, Sami, Scots Gaelic) and in North America (Fox, Hopi, and Malecite/Passamaquoddy) [4-6]. Most of the research on preaspiration has been on Icelandic [7-12] with less on Sami [13-16]. Relatively little work has been done on preaspiration in Scots Gaelic [17-19].

Measurements

Preaspiration (Preasp), Closure Duration (CD), Voiceless Duration (VlessD), and Voice Onset Time (VOT) were measured [20]. VlessD is the entire period of voicelessness including VOT. Figure 1 *tobhta* /totə/ [$t^h o^h t ə$] 'walls of a house' shows both postaspiration and preaspiration. The waveform is shown with the individual portions labelled. The breathy voice which has been mentioned in some research [17, 19] was only sporadically present and where found has been considered part of the aspiration.

Figure 1. Waveform of tobhta /totə/ [$t^h o^h t ə$] 'walls of a house'.

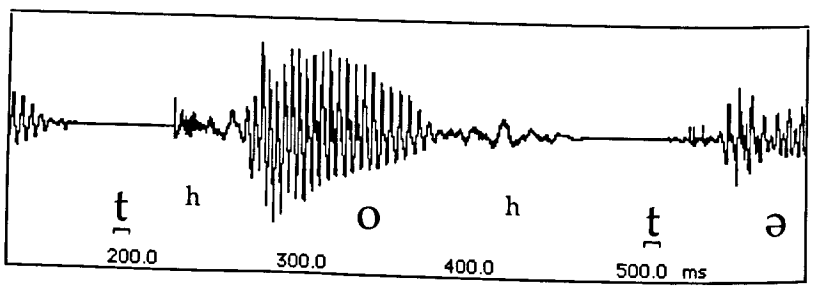


Table 1. Means of lenis and fortis stops in milliseconds. The means of all fortis–lenis pairs are significantly ($p < .05$) different except those in bold face; italics indicate pairs with a significant difference, but in the unexpected direction. In slow speech, FS paused at the end of the elicited word so that measuring the end of voicelessness was not possible .

	RM					FS				
Slow	N	Preasp	CD	VlessD	VOT	N	Preasp	CD	VlessD	VOT
Initial										
lenis	35		202	172	27	54		**161**	186	20
fortis	36		142	191	77	65		**145**	243	97
Medial										
lenis	38		114	117	*47*	41		105	119	22
fortis	35	171	66	267	*31*	39	148	76	255	33
Final										
lenis	75		140	170	**61**	83		133		
fortis	25	181	81	349	**73**	36	220	90		
Fast										
Initial										
lenis	45		87	88	22	42		*77*	95	17
fortis	39		73	110	53	49		*89*	162	75
Medial										
lenis	17		**78**	95	*36*	20		86	97	**22**
fortis	11	94	**66**	189	*28*	20	102	65	187	**26**
Final										
lenis	39		86	96	*33*	38		108	113	**22**
fortis	45	107	68	200	*22*	18	151	79	241	**22**

Gestures

In the postaspirated stops, the oral gesture begins before the laryngeal gesture begins and ends before the laryngeal gesture ends. With the preaspirated stops, the laryngeal gesture completely overlaps the oral gesture extending beyond it at both ends. From the acoustic data, the measurements are consistent with the hypothesis that, with aspirated stops, the peak of the glottal gesture is coordinated with the end of the oral gesture [7]. The acoustic activity of the preaspirated stops suggests, however, that the peaks of the glottal and oral gestures cooccur, but that the glottal gesture is larger.

RESULTS

Lenis v. fortis

Table 1 compares the lenis and fortis stops. The lenis stops have a voiceless closure followed by a short period of aspiration. The fortis stops have a voiceless closure with longer aspiration, postaspiration initially and preaspiration elsewhere. As expected, the fortis stops often have a longer closure duration than the lenis stops; in two cases, however, the difference is not significant, and in one, the fortis closure is longer. The lenis stops always have a significantly shorter voiceless duration than the fortis stops. The VOT is longer for the fortis stops in initial position, as we would expect; otherwise, it is erratic.

If slow and fast rates of speech are compared (Table 2), the lenis stops show a general shortening in all portions of the consonant, except for FS medial lenis VOT. The fortis stops show a similar general shortening in initial position; in medial and final position, however, the closure duration and VOT are not always significantly different, especially with RM.

DISCUSSION

Lenis Fortis

Aspiration is the feature which always serves to distinguish fortis and lenis stops: postaspiration initially, and preaspiration medially and finally. Closure duration is not a reliable cue in distinguishing the stops. The total amount of voicelessness

Table 2. Means of low and fast rates of speech in ms. The means of all slow-fast pairs are significant ($p < .05$) except for those in bold face. In slow speech, FS paused at the end of the elicited word so that measuring the end of voicelessness was not possible.

	RM					FS				
Initial	N	Preasp	CD	VlessD	VOT	N	Preasp	CD	VlessD	VOT
lenis										
slow	35		202	172	27	54		161	186	20
fast	45		87	88	22	42		77	95	17
fortis										
slow	36		142	191	77	65		145	243	97
fast	39		73	110	53	49		89	162	75
Medial										
lenis										
slow	38		114	117	47	41		105	119	**22**
fast	17		78	95	36	20		86	97	**22**
fortis										
slow	75	171	**66**	267	**31**	39	148	**74**	255	33
fast	39	94	**66**	189	**28**	20	102	**65**	187	26
Final										
lenis										
slow	75		140	170	61	83		133		
fast	39		86	96	33	38		108	113	22
fortis										
slow	25	181	**81**	349	73	36	220	88		
fast	45	107	**68**	200	22	18	151	79	241	22

is distinctive; however, the aspiration, produced with an open vocal tract, is the most audible and perceptually the most salient part of this voiceless period. The unsystematic variation of VOT in noninitial position is not important since preaspiration serves to distinguish lenis and fortis stops in those positions.

Kingston & Diehl [2] have argued that postaspiration in English is a controlled allophonic aspect of production. Their arguments would apply equally well to Scots Gaelic. Further, the argument that preas-piration is also nonautomatic can be made even more strongly, given its rarity in the world.

Rate of speech

In faster speech, in contexts where there is no preaspiration, a general shortening occurs. With the preaspirated stops, all shortening tends to be in the VlessD. The relative stability of the CD at different rates of speech implies that the major adjustments for rate of speech are made during the adjacent pre- and postaspirated periods.

To speak faster, something has to be shortened. This shortening is not necessarily done evenly in all parts of the utterance [18, 21]. In previous work, I have shown that languages use a variety of language-specific strategies to shorten elements in order to talk faster. In Mongolian [22-24], the VlessD for fortis stops remains steady at different rates of speech; in French [23], the VOT remains unchanged, but the voiceless portion of the closure is shortened; in Turkish (Rogers, 1994) the fortis stops show general shortening. Now, in Scots Gaelic, the stops without preaspiration show general shortening, but the preaspirated stops show a shorter voiceless duration.

These findings are consistent with the position of Docherty [1] and Kingston and Diehl [2] that considerable allophonic variation must be accounted for in the grammar of the language, and not by recourse to automatic processes.

CONCLUSION

Aspiration has been shown to be the primary difference between the lenis and fortis stops: postaspiration in initial position, and preaspiration elsewhere. Both types of aspiration are produced by controlled activity at an allophonic level. With an increased rate of speech, a language-specific observation was made that stops with postaspiration show general shortening, and those with preaspiration are shortened in the voiceless duration.

REFERENCES

[1] Docherty, Gerard, (1992), *The timing of voicing in British English obstruents*, Berlin, New York: Foris Publications.

[2] Kingston, John and Randy Diehl, (1994), "Phonetic knowledge", vol. *Language*, pp. 419–54.

[3] Rogers, Henry, (1972), "The initial mutations in modern Scots Gaelic", *Studia Celtica*, vol. 7, pp. 63–85.

[4] Hurch, B., (1988), *Über Aspiration: Ein Kapitel aus der natürlichen Phonologie*, Tübingen: G. Narr.

[5] Laver, John, (1994), *Principles of Phonetics*, Cambridge: Cambridge University Press.

[6] Rogers, Henry, (1991), *Theoretical and Practical Phonetics*, Toronto: Copp Clark Pitman.

[7] Löfqvist, Anders and Hirohide Yoshioka, (1981), "Laryngeal activity in Icelandic obstruent production", *Nordic Journal of Linguistics*, vol. 4, pp. 1–18.

[8] Pétursson, Magnús, (1972), "La préaspiration en islandais moderne: Examen de sa réalisation phonétique chez deux sujets", *Studia Linguistica*, vol. 26, pp. 61–80.

[9] Pétursson, Magnús, (1976), "Aspiration et activité glottale", *Phonetica*, vol. 33, pp. 169–98.

[10] Pind Jürgen, (1994), "Constancy and normalization in the perception of voice offset time as a cue for preaspiration", *Acta Psychologica*, to appear.

[11] Pind, Jürgen, (1994), "Perception of aspiration and preaspiration in Icelandic", draft.

[12] Thráinsson, Höskuldur, (1978), "On the phonology of Icelandic preaspiration", *Nordic Journal of Linguistics*, vol. 1, pp. 33–54.

[13] Engstrand, Olle, (1987), "Preaspiration and the voicing contrast in Lule Sami", *Phonetica*, vol. 44, pp. 103–116.

[14] Kylstra, A.D., (1972), "Die Präaspiration im Westskandinavischen und im Lappischen", *Orbis*, vol. 21, pp. 367–82.

[15] McRobbie, Zita, (1991), "Preaspiration in Skolt Sámi", *SFU Working Papers*, vol. 1, pp. 77-87.

[16] McRobbie, Zita (1993), "The role of pre-aspiration duration in the voicing contrast in Skolt Sámi", *International Congress on Spoken Language*, Banff.

[17] Chasaide, Ailbhe Ní. and Cathair Ó. Dochertaigh, "Some durational aspects of pre-aspiration," in *Topics in Linguistic Phonetics: In Honour of E.T. Uldall*, J.-A.W. Higgs and R. Thelwall, Editor. 1984, New University of Ulster: pp. 141–57.

[18] Rogers, H., (1994), "Preaspiration in Scots Gaelic", *Proceedings of the 1994 Annual Conference of the Canadian Linguistic Association*, pp. 465–76.

[19] Shuken, Cynthia, "[ʔ], [h], and parametric phonetics," in *Topics in Linguistic Phonetics: In Honour of E.T. Uldall*, J.-A.W. Higgs and R. Thelwall, Eds, 1984, New University of Ulster: pp. 111–39.

[20] Brown, W.S., Jr., R.J. Morris, and R. Weiss, (1993), "Comparative methods for measurement of VOT", *Journal of Phonetics*, vol. 21, pp. 329–336.

[21] Löfqvist, Anders, (1991), "Proportional timing in speech motor control", *Journal of Phonetics*, vol. 19, pp. 343–50.

[22] Rogers, Henry, (1992), "Laryngeal timing in Mongolian", *Proceedings of the 1992 Canadian Linguistic Association Conference*, pp. 241–8.

[23] Rogers, Henry, (1993), "A revised theory of articulatory binding", *Proceedings of the 1993 Canadian Linguistic Association Conference*, pp. 573–84.

[24] Rogers, Henry, (1995), "The effect of rate of speech on laryngeal timing in medial stops in Mongolian", to appear.

THE CHARACTER OF /r/-SOUNDS: ARTICULATORY EVIDENCE FOR DIFFERENT REDUCTION PROCESSES WITH SPECIAL REFERENCE TO GERMAN

N. O. Schiller and C. Mooshammer
Max-Planck-Institute for Psycholinguistics, Nijmegen, The Netherlands
and Forschungsschwerpunkt Allgemeine Sprachwissenschaft, Berlin, Germany

ABSTRACT

The class of /r/-sounds is quite heterogeneous from the acoustic and the articulatory point of view. Many /r/-allophones are the result of phonetic reduction processes. In contemporary standard German /r/ can be realized as an apical or uvular trill, a fricative, an approximant or a vowel. In the experimental study presented here, two German subjects were investigated by means of EMA and spectrography. The results provide articulatory evidence for the gestural affinity between the different /r/-allophones.

1. INTRODUCTION

Laterals and /r/-sounds (or so-called rhotics) have many phonological and phonetic similarities (cf. [1], [2]) (e.g. distribution in the syllable, behaviour in sound change, articulatory, acoustic and perceptual features) and therefore belong to the class of liquids. The liquids constitute a small sound class; but nevertheless, they are represented in the majority of languages (cf. [3], [4]). For instance, /r/-sounds can be found in 76% of the UPSID languages (cf. [3]: 73). But the subclass of rhotics is phonetically quite heterogeneous (cf. [5]: 114) and a definition of /r/-sound is not available. Beside a set of core elements (e.g. trills, taps and flaps), there are some fricatives, approximants and vocalized allophones of /r/. The reason why there are so many allophones of /r/ is to be found in the phonetic nature of the rhotics: some /r/-sounds, e.g. trills, are quite complex in their articulation. They are learned relatively late by children, and their articulation causes serious problems for many adult speakers (cf. [6]). That is why trills are not very stable sounds; they are prone to undergo sound change resulting in different allophones.

In German, e.g., historically there was (presumably) only an apical-alveolar trill [r] (cf. [7], [8]); a uvular-postdorsal trill [R] developed later (cf. [9]) and was considered to be a sub-standard allophone of /r/ for a long time (cf. [10]: 51). In contemporary Standard German there are different fricatives, approximants and vocalized forms called /r/-sounds (cf. [11], [12]). Whereas acoustic analyses of different /r/-sounds have already been published (cf. [2], [13]), to our knowledge articulatory studies are not yet available. One of the few experimental studies on the production of /r/-sounds was provided by Recasens ([14]). He found in an EPG study that "[...] an alveolar tap and an alveolar trill show contrasting degrees of resistance to coarticulation from the adjacent vowels" ([14]: 279). Recasens concluded that taps and trills are produced by means of two different gestures.

The aim of this paper is to show that the different allophones of /r/ in German are closely related to each other articulatorily. Our hypothesis was that the constriction location for a certain class of /r/-sounds, e.g. uvular /r/s, is the same, and that the allophones of that particular class differ only in the constriction degree. To give articulatory evidence for the relatedness of superficially different looking /r/-sounds an EMA experiment was carried out.

2. EXPERIMENT

An experimental investigation was carried out to examine the articulatory relatedness of different forms of /r/-reduction in German.

2.1. Subjects

Two German speakers served as subjects. Subject 1 was a male speaker of standard German with an uvular variant of /r/. Subject 2 (the second author) was a female speaker of standard German but spoke an apical allophone of /r/ because of her Bavarian dialectal background.

2.2. Speech material

Test utterances were designed to test several aspects of /r/ articulation in German: coarticulatory effects, position dependent reductions and articulatory similarities between sounds with a similar place of articulation. Accordingly, the first part of the corpus consisted of items including /r/ in varying vowel contexts, e.g. /bara/, /bari/, /bira/ etc. The effect of syllable position was tested by means of pairs such as /rup/ vs /bur/, /rip/ vs /bir/ and /rap/ vs /bar/. To compare /r/ with neighbouring sounds, items like /tap/, /sap/, /pat/, /pas/ were tested for [r]. [R] was compared with /kap/, /pak/, /xap/, /pax/. Both subjects produced the test items in the carrier utterance "Ich habe ____ erwähnt" ("I mentioned ____ "). 50 test utterances were designed in total. Five repetitions were run with the test utterances randomised in all repetitions.

2.3. Method

Electromagnetic articulography (EMA, AG100, Carstens Medizinelektronik, Göttingen, Germany) was used to monitor tongue movements (cf. [15]). For this method three transmitter coils (mounted on a helmet) are used to generate an alternating magnetic field at three different frequencies. Five sensor coils, attached to the subjects' articulatory organs by means of physiological glue, detect the magnetic field strength which is roughly inversely proportional to the cube of the distance between sensor and transmitter (cf. [16] for details). The raw distance signals are then converted to the x-y coordinates in the mid-saggital plane. Three coils were mounted on the mid-saggital line of the tongue from about one to five cm from the tongue tip (coil 1 was placed one cm behind the apex, coil 2 was attached to the tongue blade and coil 3 was positioned approximately five cm from the tongue tip on the tongue dorsum). To compensate for head movements two reference coils were attached to the upper incisors (coil 4) and to the bridge of the nose (coil 5).

Simultaneously to the EMA recordings, acoustic recordings were made. Subjects read the corpus of 50 test utterances five times to get 250 recordings from each of the two subjects.

3. RESULTS

In this chapter, due to space limitations, we will mainly relate to the observed reductions of uvular /r/.

3.1. Acoustic analyses

The acoustic analyses were run with the XHADES program on a VAX computer at the NICI[1].

On the whole, it can be said that the uvular trill was rarely realized by speaker 1. [R] was only produced in initial position or in initial C-clusters. Most often, /r/ is reduced to a fricative (see figure 1). In final position after a long stressed vowel, /r/ is realized as a vowel [ɐ] (see figure 2).

The apical /r/ is initially realised either as a trill or as an approximant. Apical /r/ is not pronounced as a fricative because [z] already has phonological status in German. In syllable final position, apical /r/ is most often vocalized or articulated as an approximant.

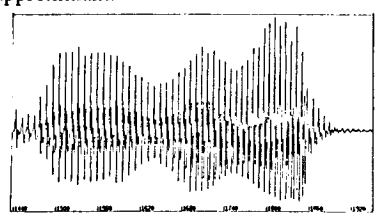


Figure 1. Wave form of the test stimulus /rip/ [ʁip].

[1] Nijmegen Institute for Cognition and Information, Nijmegen, The Netherlands.

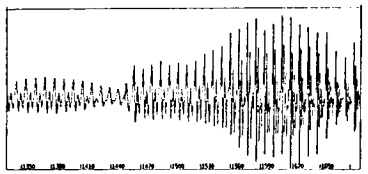


Figure 2. Wave form of the test stimulus /bir/ [biɐ].

3.2. Articulatory analyses

Articulatory data were rotated in such a way, that the line from the upper incisors to the bridge of the nose was parallel to the y-axis. Due to problems with the hardware, some recordings were lost or expelled from the data analyses.

Articulatory analyses of uvular /r/ were carried out for all repetitions of /bar/, /bir/, /bur/, /rap/, /rip/ and /rup/. Therefore, steady states of the movement signals of coil 3 served as the criterion for the target positions of uvular /r/. As can be seen in figures 3 and 4, the position of the tongue dorsum was higher when /r/ was in initial position in the vicinity of high vowels (/u/ and /i/) and lower when /r/ was produced in final position. No dependency on syllable position could be observed when /r/ was adjacent to the low vowel /a/ (see figure 5).

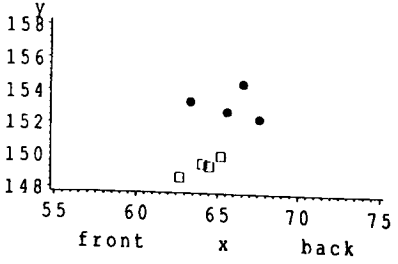


Figure 3. Plot of the tongue body position (coil 3) for initial vs final uvular /r/ (V=[u]). • = syllable initial position, □ = syllable final position.

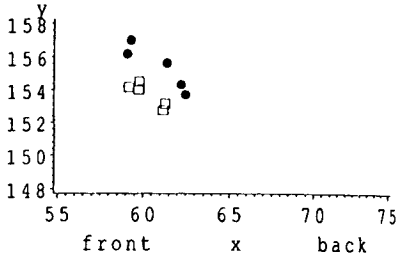


Figure 4. Plot of the tongue body position (coil 3) for initial vs final uvular /r/ (V=[i]). • = syllable initial position, □ = syllable final position.

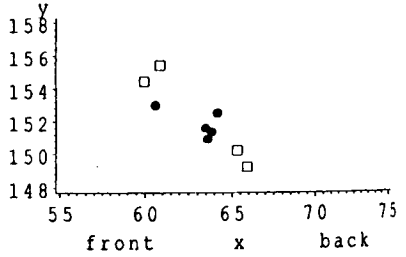


Figure 5. Plot of the tongue body position (coil 3) for initial vs final uvular /r/ (V=[a]). • = syllable initial position, □ = syllable final position.

4. DISCUSSION

The acoustic analyses of the productions of uvular /r/ revealed several forms of articulatory reduction. In syllable initial position, uvular /r/ is often produced as a post-dorsal fricative. In syllable final position after a stressed long vowel, however, uvular /r/ is often vocalized. This is in accord with previous studies on the reduction of /r/ (cf. [11]). The articulatory analyses carried out by means of EMA showed that, in general, the tongue body position is higher when uvular /r/ is syllable initial and lower when it is syllable final. This is true when /i/ or /u/ was the nucleus of the syllable; when /a/ was the nucleus, the situation was rather unclear.

These findings are in accord with the hypothesis that the different forms of articulatory reduction for uvular /r/ result from different degrees of tongue body constriction in the post-dorsal area. In the framework of *Articulatory Phonology* developed by Browman and Goldstein (cf. [17]) this can be described in a very simple way. It seems that the uvular trill and its forms of reduction (fricative, approximant and vowel) all belong to the same gestural family, e.g. a post-dorsal constriction gesture. The only difference between these allophones lies in the degree of tongue body constriction.

Acknowledgements

The authors are indebted to P. H. H. M. van Lieshout (NICI, Nijmegen, The Netherlands) for his great help with running the acoustic analyses. Furthermore, we would like to thank Jörg Dreyer (FAS, Berlin, Germany) for having been subject 2 and for his technical support in the articulatory analyses and Daniel Swingley (Stanford University, Stanford, USA) for proof-reading the paper.

5. REFERENCES

[1] Bhat, D. N. S. (1974), "The phonology of liquid consonants", *Working Papers on Language Universals*, vol. 16, pp. 73-104.

[2] Ladefoged, P., Cochran, A., Disner, S. (1977), "Laterals and trills", *Journal of the International Phonetic Association*, vol. 7, pp. 46-54.

[3] Maddieson, I. (1984), *Patterns of sound*, Cambridge et al.: Cambridge University Press.

[4] Göschel, J. (1971), "Artikulation und Distribution der sogenannten Liquida r in den europäischen Sprachen", *Indogermanische Forschungen*, vol. 76, pp. 84-126.

[5] Lindau, M. (1980), "The story of /r/", *UCLA Working Papers in Phonetics*, vol. 51, pp. 114-119.

[6] Luchsinger, R., Arnold, G. E. (1970), *Handbuch der Stimm- und Sprachheilkunde. Zweiter Band: Die Sprache und ihre Störungen*. Third edition, Wien, New York: Springer.

[7] Lehmann, W. P. (1951), "The distribution of Proto-Indo-European /r/", *Language*, vol. 27, pp. 13-17.

[8] Penzl, H. (1961), "Old High German ⟨r⟩ and ist phonetic identification", *Language*, vol. 37, pp. 488-496.

[9] Moulton, W. G. (1952), "Jacob Böhme's uvular r", *The Journal of English and Germanic Philology*, vol. 51, pp. 83-89.

[10] Siebs, T. (1898), *Deutsche Bühnenaussprache*, Berlin: de Gruyter.

[11] Hildebrandt, B. F. O., Hildebrandt, L. M. (1965), "Das deutsche R. Regelhaftigkeiten in der gegenwärtigen Reduktions-Entwicklung und Anwendung im Fremdsprachenunterricht", *Linguistics*, vol. 11, pp. 5-20.

[12] Hall, T. A. (1993), "The phonology of German /R/", *Phonology*, vol. 10, pp. 83-105.

[13] Meyer-Eppler, W. (1959), "Zur Spektralstruktur der /r/-Allophone des Deutschen", *Acustica*, vol. 9, pp. 247-250.

[14] Recasens, D. (1991), "On the production characteristics of apico-alveolar taps and trills", *Journal of Phonetics*, vol. 19, pp. 267-280.

[15] Schönle, P. W. (1988), *Elektromagnetische Artikulographie. Ein neues Verfahren zur klinischen Untersuchung der Sprechmotorik*, Berlin et al.: Springer.

[16] 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", *Journal of the Acoustical Society of America*, vol. 92, pp. 3078-3096.

[17] Browman, C. P., Goldstein, L. (1989), "Articulatory gestures as phonological units", *Haskins Laboratories Status Report on Speech Research*, vol. SR-99/100, pp. 69-101.

VARIABILITY OF LINGUAL STOPS IN ENGLISH: AN ELECTROPALATOGRAPHIC STUDY

D. Waters, K. Nicolaidis, W.J. Hardcastle and F. Gibbon
Queen Margaret College, Edinburgh, UK

ABSTRACT

This electropalatographic (EPG) study investigates aspects of normal articulatory variability in English lingual stops in VCV sequences. Variability due to the effects of vocalic context, voicing and type of speech material (real versus nonsense words) are examined. The methodology for data collection, segmentation and analysis is described. Results of the analysis will be available at the Stockholm Congress.

INTRODUCTION

Quantification of the extent of normal variability of articulatory movement during speech is of importance both in the development of robust automatic speech recognition systems and in the assessment and remediation of disordered speech. For the latter, it is essential to use baseline normative data against which to compare abnormal articulatory patterns. EPG investigation of tongue contacts with the hard palate during speech offers a means of making systematic investigations of aspects of normal articulatory variability. Previous studies have suggested various factors potentially affecting the variability of EPG patterns [1,2]. This study investigates variability associated with inter-speaker differences, vowel context, effects of voicing and of different types of speech material (real and nonsense words). It contributes towards establishing regions of relative invariance and variability in tongue contact patterns for lingual stops in English.

DATA

The speech data (from the EUR-ACCOR database, [3]) consisted of VCV sequences in English in both real and nonsense words, where V = /i, a/ and C = /t, d/. Five English speaking subjects each produced five repetitions of each VCV combination in nonsense and real words (80 tokens from each subject). Data were acquired with the mulitchannel system developed by Reading University and IBM. Electropalatographic (Reading EPG), laryngographic and acoustic data were recorded simultaneously for each item. [4, 5].

DATA SEGMENTATION PROCEDURE

A multi-level approach was adopted in the segmentation of the data. The placement of annotation points was based on information from the acoustic waveform, the electropalatographic data and the laryngographic signal. Twelve annotation points were marked and labelled for each VCV sequence. They were saved in separate annotation files for each utterance. These files were then used for further analysis of the data. Five of these annotation points are relevant for the current analysis of stops.

Annotation points

sce (stop closure) Taken at the first EPG frame showing complete constriction at the alveolar region. In cases of incomplete closure it was identified as the first frame of maximum constriction in the alveolar region.

sre (stop release) Taken at the last frame of complete constriction at the alveolar region. In cases of incomplete closure it was taken as the last frame of maximum constriction at the alveolar region.

sm (stop midpoint) Taken at the temporal midpoint between 'sce' and 'sre'.

mce (maximum contact from the EPG) Taken at the first frame of maximum constriction in the first four rows of electrodes during the stop closure.

lgp (last glottal pulse) Based on the laryngographic signal this point was taken at the end of periodic pulsing after the onset of closure for the consonant.

DATA ANALYSIS

Analyses were carried out using the Paradox 3 relational database and statistical analyses were performed using a PC running the SPSS statistical package. A number of quantitative and qualitative analysis procedures were undertaken to answer specific questions concerning the nature of articulatory variability of the two alveolar plosives.

Preliminary analysis of the data had indicated a number of cases where the stops were realised as fricatives with evidence of turbulent noise in the acoustic waveform trace. Such items were excluded from the quantitative analysis, however they were annotated separately and included in the qualitative analysis. The following analyses were carried out.

Contact totals

The total number of contacted electrodes in the four front and four back rows of the EPG palate were calculated separately. These measures were made for each /t/ and /d/ token involving symmetrical vowel environments. The measurements were made at the three consonantal points 'sce' (onset of constriction), 'mce' (first frame of maximum constriction in the first four rows) and 'sre' (end of constriction). This analysis aimed to identify any differences in the amount of contact between voiced and voiceless plosives. Because of the aerodynamic requirements for the production of voicing during stops the prediction would be that /d/ would involve less lingual-palatal contact than /t/.

Variability Index

EPG prototypical frames displaying frequency of electrode contact in five repetitions were computed. Based on these a variability index was calculated following Farnetani & Provaglio [6]. This index was calculated for consonants in symmetrical sequences at point 'sm' and used to examine differences in variability between the two consonants, between the two utterance types and among the five subjects.

Coarticulatory index

Contextual variability was quantified using a coarticulatory index (CI) [7]. This index quantifies the amount of tongue-palate contact in different vocalic environments. CIs calculated for symmetrical environments show global effects of the vocalic environment on the consonant. Contact for symmetrical sequences can then be compared with asymmetrical environments to determine possible anticipatory and carryover effects at the beginning (sce) and end (sre) of the consonant respectively.

Voicing

A descriptive analysis was made of the percentage of voicing during the closure phases for the consonants /t/ and /d/. Comparisons were made among subjects and between the two kinds of speech material (real and nonsense words). Results are displayed in the form of histograms.

Qualitative Analysis

Further descriptive analysis was carried out in order to display examples of contact patterns for /t/ and /d/ at the extremes of variation (maximum and minimum contact required for the

production of /t/ and /d/) for each subject.

STATISTICAL ANALYSIS

Various statistical analyses are currently being explored. The factors to be examined are:

Subject.	Five subjects.
Consonant.	/t/, /d/.
Vowel 1.	/ i /, /a/.
Vowel 2.	/ i /, /a/.
Speech Material.	Real word, nonsense word.

RESULTS

At the time of writing the above analyses are in progress. Results will be displayed as a poster at the Stockholm Congress and will be made available in printed form to delegates, together with a discussion of the results, as a supplement to this paper.

REFERENCES

[1] Butcher, A., & Weiher, E. (1976), "An electropalatographic investigation of coarticulation in VCV sequences." *Journal of Phonetics,* vol. 4, No.1, pp. 59-74.

[2] Dagenais, P.A., Lorendo, L.C. & McCutcheon, M.J. (1994), "A study of voicing and context effects upon consonant linguapalatal contact patterns." *Journal of Phonetics,* vol. 22, pp. 225-238.

[3] Hardcastle, W.J. & Marchal, A. (1990), "EUR-ACCOR: a multi-lingual articulatory and acoustic database." *ICSLP 90, Proceedings of the International Conference on Spoken Language Processing, Kobe, Japan, November 18th - 22nd,* vol. 2, pp.1293-1296.

[4] Hardcastle, W., Gibbon, F. and Nicolaidis, K. (1991), "EPG data reduction methods and their implications for studies of lingual coarticulation." *Journal of Phonetics,* vol. 19, No. 3/4, pp. 251-266.

[5] Hardcastle, W.J., Jones, W., Knight, C. Trudgeon, A. and Calder, G. (1989), "New developments in electropalatography: a state-of-the-art report. *Clinical Linguistics & Phonetics,* 3, pp. 1-38.

[6] Farnetani, E., & Provaglio, A. (1991), "Assessing variability in lingual consonants in Italian." *Quaderni del Centro di Studio per le Ricerche di Fonetica del C.N.R.,* vol. X, pp. 117-145.

[7] Farnetani, E., Hardcastle, W. J. & Marchal, A., (1989), " Cross-language investigation of lingual coarticulatory processes using EPG." In Tubach, J.P. & Mariani, J.J. (Eds) *Eurospeech 1989, European Conference on Speech Communication and Technology, Paris,* pp. 429-432.

INTENTIONALITY IN THE SPEECH ACT AND REDUCTION PHENOMENA

Lourdes Aguilar, María Machuca
Departament de Filologia Espanyola, Facultat de Lletres,
Universitat Autònoma de Barcelona, Bellaterra 08193, Barcelona, Spain

ABSTRACT

The present study focus on the influence of intentionality of the speech act in the characterisation of speaking styles. Two procedures to elicitate spontaneous speech in a laboratory environment are presented: the map task and a semidirected interview, both performed with the same speaker. Differences in elicitation procedures are interpreted in basis of the intentionality of the speech act. Related to this, phonetic behaviour of vocalic groups in Spanish for a given speaker is observed. Results show that speech obtained by means of the described procedures is spontaneous, even when intentionality in the speech act appears.

1. INTRODUCTION

It is well known that in all languages speakers have some pronunciation, lexical or grammatical choices which are not a matter of the basic structure of the language but a matter of style; moreover, they have an implicit knowledge of the appropiateness of the speaking style they are using with respect to the situation where they are. Nevertheless, variations are presented in a continuum scale and the factors shaping a given speaking style can have a linguistic, sociolinguistic or pragmatic nature [1, 2].

In this study, we focus on the importance of intentionality of the speech act in the characterisation of speaking styles: Searle [3] uses the structure of the speech acts as a heuristic guide in order to elucidate the structure of intentional states.

Two procedures to elicitate spontaneous speech in a laboratory environment are presented: the map task and a semidirected interview, both performed with a same speaker. Differences in elicitation procedures are interpreted in basis of the nature of the speech act. An overt intentionality in the speech act exists in the map task: both speaker and listener are involved in the task and they want to achieve the objectif with the maximum success. The semidirected interview, on the contrary, lacks of intentionality: there is not an explicit purpose to reach.

The manifestation of vocalic reduction phenomena in Spanish is taken as the index of study to determine if the presence of intentionality affects the speech in a sense of more carefulness and in a loss of naturalness. From an experimental point of view, the main interest consists in determining if the described procedures are suitable to obtain spontaneous speech.

2. EXPERIMENTAL PROCEDURE

2.1. Speech Situations

In order to study the influence of the intentionality of the speech act in the phonetic manifestation of vocalic sequences, two types of corpus have been used: the model of HCRC Map Task Corpus [4, 5] and the model of semidirected interviews.

Elicitation procedures are different, but some variables are controlled: the participants are the same, they maintain a familiarity relation and they have comparable speech rates.

The aim is to compare different speaking styles with the same linguistic content for a given speaker.

2.1.1. Map Task

The map task follows the model in which two actors inhabiting a simple micro-world cooperate to get practical goals: in this case, the completion of a route designed in one actor's map but inexistent in the other actor's map. In the course of the task, conversation arises.

The collection of the corpus by means of the map task allows to consider the degree in which a communicative act characterised by an overt intentionality and requiring cooperation, affects the language use: there is a clear objectif, which must be reached and can be only achieved by means of the verbal interaction of the speakers.

2.1.2. Semidirected interview

In the semi-directed interview, the roles of the interviewer and the interviewee are previously determined, although both participants are familiar.

The interviewer proposes the subjects of conversation which change depending on the interest of the speakers; both interviewer and interviewee participate actively in the conversation so it is far from being a monolog.

2.2. Recordings

Samples of the speech of a male speaker aged 25, with high-level studies, were collected from both situations.

The recordings have been done in a sound-treated room in the Autonomous University of Barcelona. The recordings for each speech situation takes approximately an hour and a half .

2.3. Corpus

In Spanish, the difference between vocalic groups in hiatus -vowel+vowel sequences- and diphthongs -glide+vowel and vowel+glide sequences- is an important idiosincrasy of the language: the fact that a sequence can be realised as a hiatus or must be pronounced as a diphthong is a lexical property of the words, and speakers have strong intuitions concerning the pronunciation in hiatus or in diphthong of the vocalic sequences. In spontaneous speech, a continuum of reduction going from hiatus to diphthong to vowel can be described [6].

In this study, the following vocalic combinations are observed: hiatuses ['ia], ['io]; diphthongs [i̯a], [i̯o]. The vowels [a, o], acoustically distant from [i], have been chosen in order to determine easily the reduction to a diphthong or to a vowel.

2.4. Procedure analysis

The traditional nondynamic procedure of acoustic analysis of diphthongs and hiatus consisting in the segmentation of the sequence in three areas corresponding to an initial segment, a transition and a final segment [7, 8] has been adopted here. If transition can not be observed, the sequence is segmented in the initial segment and the final segment.

As for the analysed parameters, in the temporal domain global duration and duration of the initial segment, the transition (if possible) and the final segment have been considered; in the frequential domain, the first two formant frequencies in the centre of the initial segment, the initial boundary of the transition, the final boundary of the transition (when existing) and the centre of the final segment are observed. Data have been obtained by means of the MacSpeech Lab II speech analysis software.

3. RESULTS

Results are organised around two questions: the manifestation of phonetic reduction processes related to vocalic sequences, and the acoustic cues that differentiate hiatuses, diphthongs and vowels in each of the observed speech situations. We have considered hiatus when two segments appear clearly, diphthongs when a transition existed from one segment to the other, and vowels when only one segment can be observed.

3.1. Phonetic reduction processes

Concerning vocalic groups in Spanish three types of processes can be observed:

a) strenghthening, where a diphthong is realised as a hiatus;

b) maintenance, where the vocalic group doesn't change its phonetic quality;

c) weakening, which can show a three-fold result: a diphthongisation, where a hiatus is pronounced as a diphthong; a vocalic deletion in a hiatus; or a vocalisation of a diphthong, manifested as a fusion in an intermediate element, sharing properties of the original segments of the group, or as a deletion of one of the segments.

Figure 1 shows in percentages the phonetic results of the hiatus and diphthongs in the map task and in the semidirected interviews.

Cases of strenghthening have been found only in the speech excerpted from the semidirected interviews (10.26%), mainly due to the presence of emphasis. On the contrary, the processes of

weakening are present in both types of corpus, referred to hiatuses and to diphthongs: in the semidirected interviews, the 27% of hiatus is pronounced as a diphthong and 12.1% as a vowel, whereas in the map task corpus, any case od diphthongisation is found but a 33.33% of vowel reduction; concerning the acoustic manifestation of diphthongs, 59.09% is reduced to a vowel in the map task corpus and 43.59% in the semidirected interviews.

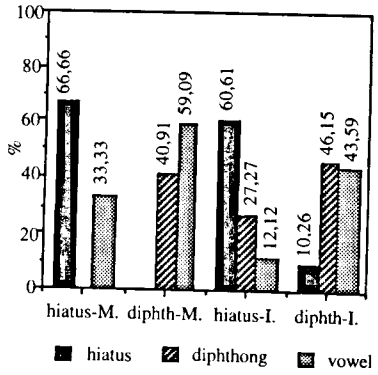


Fig. 1. Percentages of the phonetic results of the hiatuses and diphthongs in the map task (M) and in the semidirected interview (I).

If we pooled data referred to vocalic sequences, we found 44% of maintenance cases and 56% of weakening cases in the corpus of map task; in the semidirected interview, 52.7% of maintenance cases, 41.6% of weakening cases and 5.55% of strenghthening cases are obtained.

3.2. Acoustic Parameters

Hiatuses, diphthongs and vowels, the three acoustic manifestations which are the result of the vocalic sequences analysed, can be differentiated in basis of the total duration, and the duration and the second formant frequency of the first segment.

The number of cases (n), mean values (x) and standard deviation (sd) of these parameters can be observed in table I. A hiatus is always longer than a diphthong in the two type of corpus, and both are longer than the vowel, fact that could favour the bivocalic nature of a diphthong; with respect to the frequency, the first element of a hiatus shows a higher value of F2 than the first element of a diphthong in the semidirected interview corpus, but in the map task the relation is inverse: the initial element of the diphthong presents a higher value than the initial element of the hiatus. The consonantic context, in most of cases of the map task, a vibrant, could explain this fact.

An ANOVA analysis shows that the differences between the global duration of the sequence, the duration and the F2 frequency of the first element are significative at a 5% level of significance.

If we focus on the source of procedence of diphthongs, a difference between hiatus becoming diphthongs (weakening cases) and those diphthongs which come from diphthongs (maintenance cases) can be noted: the F2 frequency of [i] is higher in diphthongs coming from a hiatus than in diphthongs which have not suffered a change: 2067 Hz. and 1754 Hz respectively.

Finally, as far as interstyle differences is concerned, any important difference have been found. An ANOVA analysis applied to the global duration of hiatus, diphthongs and vowel in each type of corpus, and to the duration and F2 of the first element in hiatus and in diphthongs have not shown significative differences at a level of 5% in any case.

Table I. Number of cases (n), mean values (x) and standard deviation (sd) of the global duration of hiatus and diphthongs, the duration and the F2 frequency of the first element of hiatus, and diphthongs, and the duration and the F2 frequency of the vowel in the map task corpus and in the semidirected interview.

		Total dur.			Dur initial segment		F2 initial segment		
		H	D	V	H	D	H	D	V
M	n	2	9	14	2	9	2	9	14
a	x	137	104	68	69	47	1842	1997	1729
p	sd.	31	33	16	21	7	432	90	258
I	n	24	25	21	24	25	24	25	21
n	x	144	88	58	69	48	2134	1993	1813
t.	sd.	38	26	19	19	13	200	139	225

4. CONCLUSIONS

The comparison of the acoustic manifestations of vocalic sequences in speech obtained in two types of corpus, which are differentiated by the presence of intentionality, shows two main trends.

On one hand, phonetic reduction processes appear in both types of speech, pointing out the tendency to hypoarticulation, characteristic of relaxed speech styles. The presence of strenghthening processes in the semidirected interview is related to the presence of emphasis.

On the other hand, acoustic cues between vowels, diphthongs and hiatus are the same in both types of speech: duration and F2 frequency [6, 8,9].

It can be said that the strategies described in this study are suitable to elicitate spontaneous speech: the presence of intentionality, in the map task corpus, can not be directly related to the presence of a higher degree of carefulness -by contrast, these fact depends on emphasis.

On the other hand, naturalness is not affected, as it is shown by the presence of phonetic reduction phenomena: this fact can be explained by the behaviour of the speaker involved in the map task. The implication in the task shifts the speaker's attention over his language, and an unconstrained speech is obtained.

REFERENCES

[1] LLISTERRI, J. (1992) "Speaking Styles in Speech Research", *Proc. ELSNET/ ESCA/ SALT Workshop on Integrating Speech and Natural Language. 15-17th july, Dublin*, Ireland.

[2] ESKÉNAZI, M. (1993) "Trends in Speaking Styles Research" en *Proceedings of Eurospeech'93. 3rd European Conference on Speech Communication and Technology. Berlin. Germany. 21-23 September 1993*, vol.1: 501-509.

[3] SEARLE, J. (1983) *Intentionality. An Essay in the philosophy of mind*, Cambridge: Cambridge University Press.

[4] McALLISTER, J- SOTILLO, C.- BARD, E.G.-ANDERSON, A. (1990) *Using the Map Task to investigate variability in speech*, Ocassional Paper, Department of Linguistics, University of Edinburgh.

[5] ANDERSON, A.H.- BADER, M.- BARD, E.G.- BOYLE, E.-DOHERTY, G.- GARROD, S.- ISARD, S.- KOWTKO, J.-McALLISTER, J.- MILLER, J.- SOTILLO, C.- THOMPSON, H.- WEINERT, R. (1991) "The HCRC Map Task Corpus", *Language and Speech*, 34, 4: 351-66.

[6] AGUILAR, L. (1994) *Los procesos fonológicos y su manifestación fonética en diferentes situaciones comunicativas: la alternancia vocal/ semiconsonante/ consonante*, PhD dissertation, Bellaterra, Universitat Autònoma de Barcelona.

[7] LEHISTE, I.- PETERSON, G. (1961) "Transitions, Glides and Diphtongs", *Journal of the Acoustical Society of America*, 33(3): 268-277; en LEHISTE, I.(ed), *Readings in Acoustic Phonetics*, Cambridge, Mass.:The M.I.T.Press, 1967 pp. 228-237.

[8] BORZONE DE MANRIQUE, A.M. (1979) "Acoustic Analysis of the Spanish Diphtongs", *Phonetica* , 36,3: 194-206.

[9] BORZONE DE MANRIQUE, A.M. (1976) "An acoustic study of i, u in the Spanish diphtongs", *Language and Speech*, 19, pp. 121-128.

ACOUSTIC PHONETICS IN SOCIOLINGUISTICS AND DIALECTOLOGY: THE VARIATION OF ENGLISH VOWELS IN SPONTANEOUS SPEECH

Reijo Aulanko and Terttu Nevalainen
Department of Phonetics and Department of English, University of Helsinki, Finland

ABSTRACT

The paper examines the potential of acoustic phonetics in dialectology. Using Wells's standard lexical sets as our frame of reference, we discuss the methodological issues that arise in the reconstruction of a Somerset speaker's vowel system from spontaneous speech. They include vowel-internal variation, lexical type and token variation, the effects of the immediate phonemic context, and linear as opposed to logarithmic scaling of the results.

ACOUSTIC DIALECTOLOGY?

Over the last decades, acoustic measurements have been gaining ground in urban sociolinguistics. They are typically used in vowel studies to show individual speakers' vowel systems within a framework that readily lends itself to further interindividual comparisons and generalizations [1]. We think that this methodology could also be profitably applied to more traditional dialectology.

The focus of our study [2] is methodological: we explore the various options that acoustic analysis opens up to a sociolinguist and dialectologist, as well as some of the problems connected with the approach. In this paper we show how the results of the vowel analysis may differ depending on a) the exact temporal location of the measurement point, b) the number of measurements per lexeme category, and c) the phonemic context of the sound studied.

DATA: EAST SOMERSET VOWELS

The material of our study consists of the stressed monophthongs of a rural West Country speaker. Our topic and data were first suggested to us by the late Professor Ossi Ihalainen, who made recordings of Somerset folk speech in the 1970s. Our study is based on a 45-minute spontaneous interview with a 69-year-old farmer from Wedmore, East Somerset, in his home in 1976. Although its quality is not ideal, the tape recording is fully intelligible and thus lends itself to acoustic analysis.

We identified the possible vowel contrasts (not necessarily all phonemic) in our data on the basis of Wells's [3] standard lexical sets. They consist of twenty-four matching pairs for vowels in strong syllables in standard British (RP) and American English (GA). More comprehensive, these sets seemed preferable to using either RP, 'Middle English', or GA alone as a baseline for the analysis of dialectal speech. A valuable earlier source of comparison was provided by the Wedmore data from 1956 in the *Survey of English Dialects* [4].

The lexical sets examined in this study are: 1. KIT, 2. DRESS, 3. TRAP, 4. LOT, 5. STRUT, 6. FOOT, 7. BATH, 8. CLOTH, 9. NURSE, 10. FLEECE, 12. PALM, 13. THOUGHT, 15. GOOSE, 21. START, 22. NORTH, and 23. FORCE.

ACOUSTIC ANALYSIS

The data were acoustically analysed with the Intelligent Speech Analyser (ISA) system designed by R. Toivonen [5], implemented in a Macintosh Quadra 700 computer. The ISA is an interactive speech analysis system which enables the analyst to make measurements from several simultaneous displays (FFT and LPC spectra, sound spectrograms etc.) and to monitor the digitized signal auditorily. The system also makes it possible to show the measured values graphically in different ways; e.g. the frequency scales in formant (F1/F2) charts can be linear or psychoacoustic (Bark scales).

In this study each vowel target was first located by means of a wide-band spectrogram and an intensity curve and by auditory monitoring of the signal. An LPC-based automatic formant analysis was used to suggest values for F1 and F2, which were accepted only if they conformed to the visual appearance of the formant structure in the spectrogram and the FFT spectrum. Usually only one pair of F1/F2 values was recorded for each vowel token, i.e. its possible diphthongization was not traced in this analysis (except for Fig. 1 below).

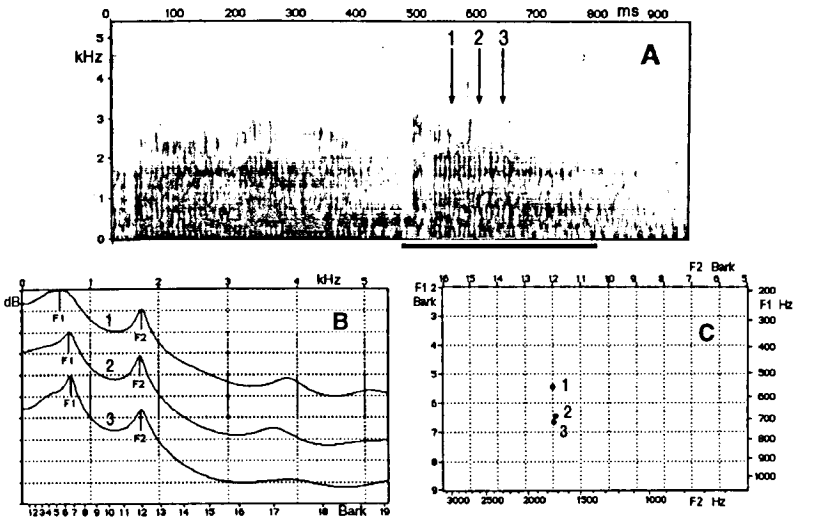


Figure 1. The effect of vowel-internal variation. The spectrogram in **A** *shows the target word in context "... cider did get, ...", the three LPC spectra in* **B** *were calculated from the time points indicated in* **A**, *and the F1/F2 plot in* **C** *shows the changes in the formant values of the target vowel.*

RESULTS

Vowel-internal variation

As only one LPC spectrum is normally determined for each monophthong, we may ask how stable the vowel is acoustically. To study its internal stability, several measurements can be made from one vowel (but care should be taken to exclude formant transitions, usually < 50 ms, due to neighbouring consonants). Fig. 1 shows changes in F1/F2 values caused by making the measurements early or late in the vowel, in addition to the normal mid-point measurement. As can be seen, in this case the effect is more pronounced in F1 than in F2 (160 Hz v. 30 Hz), indicating a slight lowering during the vowel in the word *get*. Cases like this seem to support the standard practice of using the vowel mid-point (2 in Fig. 1) as a value representative of the whole monophthong.

For the rest of the measurements, a total of 511 instances, we located the steady state of the vowel, a point where its quality would stay reasonably invariant for at least 30 ms in the middle region of the target vowel.

Lexeme token variation

Vowels may also differ in different tokens of the same word. Each dot in Fig. 2 shows the F1/F2 values from the middle points of four instances of the target word *get* pronounced in varying sentence stress conditions. The range of variation is 120 Hz for F1 and 70 Hz F2. The variation between tokens is thus approximately as large as the vowel-internal variation discussed above.

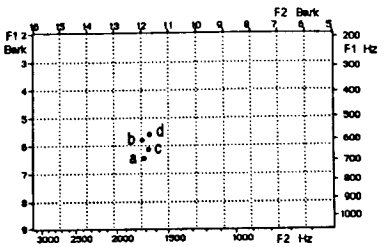


Figure 2. The distribution of four measurements representing different tokens of the word get *from the contexts*
(a) ...the cider did get,...;
(b) ...you could get 'em for...;
(c) ...you could get 'em,...;
(d) ...you could get 'em to what...

Lexeme type variation

Lexeme types may also give rise to variation. Fig. 3 shows the range of 35 F1/F2 values measured from the middle of the target vowel in the following lexemes which all represent the lexical set DRESS: *bed, dead* (2), *elm* (3), *fell* (2), *get* (3), *head* (2), *help, left* (2), *let* (3), *neck* (3), *net* (3), *plenty* (2), *press* (3), *sell* (3), and *went* (2). The number of lexeme tokens is shown in brackets.

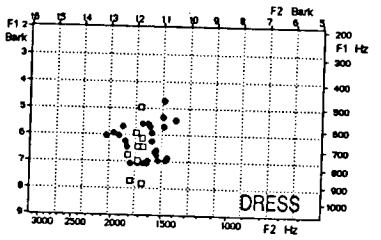


Figure 3. The distribution of 35 individual measurements representing 15 different word types from the DRESS *set. The 9 instances of the context* [C(C)_t#] *are shown by white squares, the others by black circles. (Adapted from* [2].*)*

As might be expected, the phonemic context of the vowel has an effect on its acoustic realization even within one and the same lexical set. In Fig. 3 most of the centralized tokens of the DRESS vowel (lowest F2) occur before /l/ in *elm, fell, help* and *sell*, and following /w/ in *went*. Fronted tokens (highest F2) are found in *head*, where the vowel lengthens before a voiced stop.

One way to minimize the effects of lexeme type variation is to introduce a fixed frame for all vowels. We tested this common practice with our material by selecting a context shared by the majority of our lexical sets, [C(C)_t#]. The white squares in Fig. 3 indicate the dispersion of values in this fixed context, while the black circles show all the other instances. As can be seen, the fixed context clearly reduces F2 variation in this case, whereas F1 variation is not reduced at all.

Another, more laborious approach is to take the vowel of the lexical set to be the mean value of all its lexical tokens. Fig. 4 shows that in the DRESS set there is no great difference between the means of the fixed context and the whole material.

Mean values in fixed contexts

When reconstructing vowel systems, contextual variation is usually minimized by calculating mean values for the data. Fig. 4A shows the average values obtained for eleven of our sixteen sets (i.e. 1, 2, 3, 4, 5, 6, 10, 13, 15, 21, and 22) in the fixed context [C(C)_t#]. For the sake of comparison, Fig. 4B gives the mean values for all the lexeme types representing each of the sixteen sets.

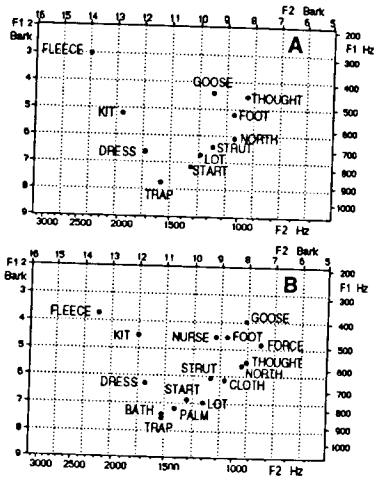


Figure 4. Average formant values measured from [C(C)_t#] *contexts (***A***) compared with the corresponding values from all available phonetic contexts (***B***). (***B *adapted from* [2].*)*

Fig. 4 indicates that the fixed frame yields a more regular distribution for the front vowels in FLEECE, KIT, DRESS and TRAP by increasing the distance between the close front vowel /i/ in FLEECE and the half-close /ɪ/ in KIT. The START, LOT and STRUT vowels are similarly kept separate in both figures, but the fixed frame suggests a more central realization for LOT.

The GOOSE and FOOT vowels are also distinguished in the two figures, but the value for GOOSE is more central in 4A. Finally, the main distributional difference between the two figures can be seen in the values for THOUGHT and NORTH. 4A clearly separates them, while 4B suggests little or no difference. The reason for this is that, as opposed to 26 cases in 4B, our material only contains

two tokens for THOUGHT in the fixed context, *caught* and *thought*. Their mean in 4A also conceals a 100 Hz difference in their F1 values. Hence the distinction between the two vowels suggested by 4A may be more apparent than real.

Visual scaling of the results

How salient are the vowel differences in Fig. 4 perceptually? Various scales have been proposed for presenting vowel formant values [1, 6].

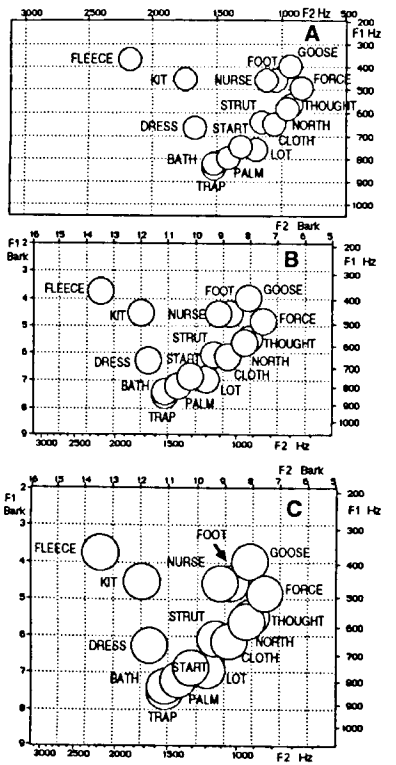


Figure 5. Average F1/F2 values for all lexical sets. **A:** *linear frequency scale (with an arbitrary circle size);* **B:** *Bark scale (with a one-Bark circle size; adapted from* [2]*);* **C:** *Bark scale with the F1 scale and the circle size expanded by 35% (see* [6]*).*

In this study we have used the Bark scale, which modifies the frequency scale according to the critical bands of human hearing. Bark-sized circles marking F1/F2 values visually simulate the psychoacoustic distances that separate any given pair of vowel qualities that are clearly distinguishable for the listener.

Fig. 5 shows the mean F1/F2 values of our lexical sets using three different frequency scales: linear, psychoacoustic (Bark), and a modified version of the latter with the F1 scale expanded by 35% (see [6]). When we set out to explore our speaker's phoneme system, the advantages of the psychoacoustic display are obvious. A comparison of different visual scaling methods in Fig. 5, however, shows that the relative differences between linear and logarithmic displays are in fact not very great.

FINAL REMARKS

Acoustic vowel studies offer dialectologists a neutral basis for comparing sound systems across speakers. The 'free' variation contained in spontaneous speech may also be an invaluable indicator of a sound change in progress.

Compared with the *SED*, our data, for instance, suggest changes in progress in the rural speech of East Somerset. The open front vowel /a/ is losing some of its functional load, while the rounded back vowels are undergoing both qualitative and distributional changes [2].

REFERENCES

[1] Labov, W. (1994), *Principles of Linguistic Change. Volume 1: Internal Factors*, Oxford: Blackwell.

[2] Nevalainen, T. & Aulanko, R. (in press) Stressed vowels in East Somerset: An acoustic-phonetic case study. In J. Klemola, M. Kytö & M. Rissanen, eds., *Studies in Memory of Ossi Ihalainen*, Frankfurt am Main: Peter Lang.

[3] Wells, J.C. (1982) *Accents of English. Vol. 1.* Cambridge: Cambridge University Press.

[4] Orton, H. & Wakelin, M.F. (1967–1968) *Survey of English Dialects, The Basic Material, Vol. IV, The Southern Counties, Parts I–III.* Leeds University.

[5] Toivonen, R. (1988) Intelligent Speech Analyser (ISA) järjestelmän käyttöohje. Tampere, 22.6.1988.

[6] Iivonen, A. (1992) Articulatory vowel gesture presented in a psychoacoustical F1/F2-space. In R. Aulanko & M. Lehtihalmes, eds., *Studies in Logopedics and Phonetics 3,* Publications of the Department of Phonetics, University of Helsinki, Series B, 4, 19–45.

GESTURAL OVERLAP AND GESTURAL WEAKENING IN CONNECTED SPEECH

Martin C. Barry

Department of Linguistics, University of Manchester, UK

ABSTRACT

This paper reports an investigation into the temporal characteristics—duration and relative timing— of coronal and dorsal lingual gestures in [nk], [tk] and similar clusters in English and Russian. Electropalatographic evidence is considered in relation to the question whether the 'assimilations' described in this context for a number of languages may be viewed as instances of gestural overlap, gestural reduction, or of genuine assimilation (change of one entity into another). Electropalatographic evidence suggests that the general case involves, initially, increased overlap between coronal and dorsal gestures, accompanied by a reduction in the magnitude of the coronal gesture once this is masked by the dorsal; there may additionally be some compensatory lengthening of the dorsal gesture. The paper therefore gives qualified support to the account proposed in the theory of Articulatory Phonology, while contributing to well-motivated explanation of the phenomenon.

1. 'ASSIMILATION' IN [nk] CLUSTERS

The phenomenon attested in many of the world's languages, whereby a sequence of consonants consisting of a dental or alveolar consonant followed by a velar may exhibit a process generally (but loosely) termed 'assimilation', yielding what has generally been identified as a geminate or partial geminate sequence [kk] or [ŋk], has been widely reported. A number of studies (cf. [1]) have sought to investigate whether 'assimilation' *stricto sensu* might not be an appropriate label for the process, since implicit in the use of the term is the assertion that one entity changes into another at some level of description (e.g. the underlying /t/ phoneme becomes a /k/; or the consonantal occlusion is formed entirely, and for geminate duration, at the velar place of articulation.) One approach which has sought to contradict the conventional view is that of Articulatory Phonology [2], according to which the phenomenon under consideration, like all phonetic and phonological phenomena, is attributable not to an assimilatory change of identity of a phonetic or phonological object, but rather to a change in the relation between these objects. The relationship in question—"phase" in the terminology of the theory—is that governing the relative timing of two independent articulatory gestures, which are themselves the primitives of the representation. It follows from this that the theory proposes a simple and economical account of the so-called 'assimilation' phenomena, from which any notion of an arbitrary change in the nature of the representation is eliminated.

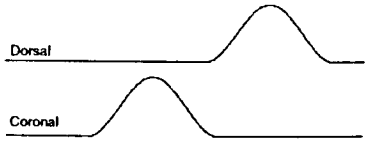


Figure 1. Schematic gestural trajectories for coronal and dorsal gestures in [nk] sequence.

The durational aspects of the data in question are crucial in evaluating the relative merits of the competing accounts. While the conventional 'assimilation' treatment predicts a significant durational difference between, e.g. the velar closure in a surface [tk] cluster (i.e. one in which the assimilation has not taken place) and a [kk] cluster resulting from such an assimilation, the Articulatory Phonology version of events predicts no such distinction, since the [tk] has been reduced to [k] by a shortening of the interval between the coronal and dorsal gestures, with the original [t] finally masked from the listener by the now-simultaneous velar closure. Under these conditions, the masked [t] is then subject to deletion, since both the onset and the offset of the closure phase of the consonant cannot, by virtue of the overlapping velar closure, be audible to the listener.

2. EXPERIMENTAL METHOD

The study reported here involved the acquisition of electropalatographic data from speakers of both English and Russian (5 English speakers, 3 Russian speakers), uttering carrier sentences involving target sequences of coronal plus velar consonants at controlled speaking rates, giving a range between slow, careful and rapid colloquial speech. EPG data was recorded at sample rates varying from 100 Hz to 220 Hz (the experimental was conducted in various laboratories in the UK, determined by availability subjects, particularly Russian speakers). From the raw EPG signal time varying plots corresponding broadly to gestural trajectories were calculated by summing contacts over regions of the palate corresponding to target contact regions for consonants in the denti-alveolar and velar regions.

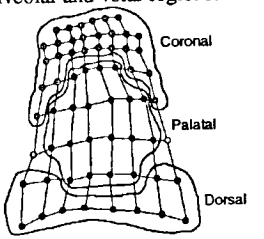


Figure 2. Sub-regions of the EPG palatal surface used for calculating coronal and dorsal (and, redundantly here, palatal) gestural trajectories.

This, I have argued, yields gestural plots of comparable quality to those derived from X-ray microbeam tracking data; the same might be argued for the comparison between EPG data and data from electromagnetic articulography. Figure 2 shows the delimited sub-regions of the artificial palate which were used for calculating coronal and dorsal gestures.

Figure 3. Schematic trajectories for coronal and dorsal gestures in underlying [nk] sequence pronounced as [ŋk]

The plots derived in this way were smoothed by the fitting of ninth-order polynomials over the data range, such that the coefficient r for the correlation between the data points and the smoothed curve was consistently greater than 0.97. The curve fitting served the dual purposes of smooth interpolation between the data points and the simple determination of onsets and maxima in the gestural trajectories; these tasks were performed by computer software written specially for the purpose, which permitted the results from a relatively large body of data to be accumulated. 60 tokens were considered from each of the eight speakers.

3. HYPOTHESES

The data collected was used to evaluate a number of competing hypotheses as to the patterns which might be discerned in the gestural trajectories for coronal-dorsal sequences.

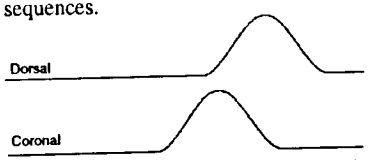


Figure 4. Schematic gestural trajectories for coronal and dorsal gestures in [nk] sequence with increased gestural overlap.

Hypothesis 1:

At increased rates of speech the coronal gesture is replaced by a prolonged dorsal gesture, with a closure duration for the stop comparable to that of a geminate sequence (Figure 3). This is the phonetic realisation predicted by the 'assimilation' analysis, and, if observed, would therefore support the view that at some cognitive level the speaker substitutes a velar for the original coronal consonant.

Hypothesis 2:

At increased rates of speech the coronal and dorsal gestures both remain intact, but the interval governing their relative timing is reduced (Figure 4). This is the pronunciation which the Articulatory Phonology account appears to propose. Under this view the perceived 'assimilation' is simply a consequence of the overlap in time of the two gestures.

Hypothesis 3:

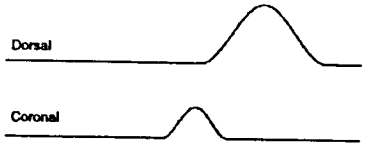


Figure 5. Schematic trajectories for coronal and dorsal gestures in [nk] sequence with coronal weakening.

At increased rates of speech the coronal gesture simply diminishes in magnitude (Figure 5). This has been proposed [3] as an alternative explanatory mechanism for the loss via apparent assimilation of syllable-final coronals. This view carries with it the prediction that the coronals will be subject to assimilation or loss in contexts other than before a heterorganic non-continuant consonant. This prediction, though, beyond the scope of the investigation reported here, does not appear to be born out by the general descriptive literature.

4. RESULTS

4.1 Sample of detailed results for a single speaker

The results obtained may best be illustrated by consideration of the patterns discernible in the articulatory activity of a single speaker. Figures 6–8 show the mean measured gestural trajectories over

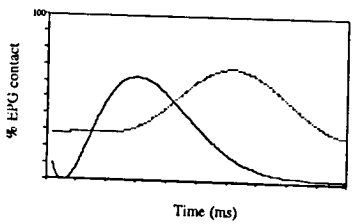


Figure 6. Mean gestural trajectories for [nk] clusters in slow utterances (Speaker DM)

twenty tokens for [nk] clusters uttered by one of the English speakers at the three self-selected speaking rates slowly, normal conversational rate and quickly.

It is evident from these figures that the data are consistent with none of the three hypotheses outlined above considered in

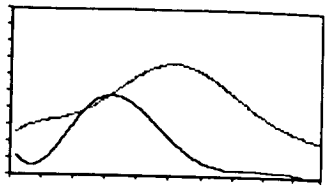


Figure 7. Mean gestural trajectories for [nk] clusters in normal conversational-rate utterances (Speaker DM)

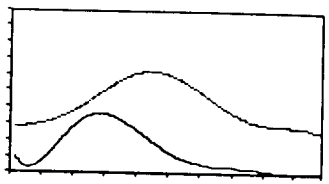


Figure 8. Mean gestural trajectories for [nk] clusters in fast utterances (Speaker DM)

isolation, but they appear to show tendencies reflecting both hypotheses 2 and 3: gestural overlap increases, and at the same time the magnitude of the coronal gesture reduces. Hypothesis 1, on the other hand, appears to receive less support from the data: in the general case, the duration of the dorsal gesture does not vary significantly across speaking rates, and there is no tendency exhibited for the velar gesture to lengthen in compensation as the supposed 'assimilation' takes effect. This last tendency is subject to certain exceptions, discussed below in 4.3.

4.2 General trends in the results

The tendency in the data for the speaker considered above to show compatibility with a hybrid version of hypotheses 2 and 3 (gestural overlap plus coronal reduction) was repeated for **all speakers** and for **both languages**. The question proposes itself: are there genuinely two independent forces at work in connected speech, inducing increased gestural overlap simultaneously with but independently from coronal reduction? This would, naturally, be an uncomfortable result, since (for the eight speakers and two languages considered here) the two phenomena appear to be closely linked, since we do not encounter either effect in the absence of the other.

A clearer picture may be obtained by observing the detailed pairwise correlation between the two dimensions being measured. Figure 9 shows a plot of measurements from all the tokens recorded in the study. The horizontal axis shows the interval between the peak value (corresponding to the mid-point in time) of the coronal gesture and the peak value for the dorsal gesture; thus the nearer to the left of the diagram, the shorter the phase-interval between coronal and dorsal gestures. The vertical axis shows the magnitude of the coronal gesture (as % of the maximum degree of lingual-(denti)alveolar contact for a fully-articulated stop closure). A point on the graph therefore shows the degree of coronal reduction for a given inter-gestural phase interval.

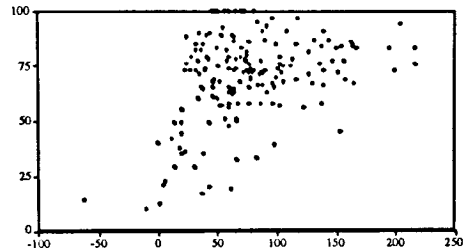


Figure 9. Plot of inter-gestural phase interval (in ms; horizontal axis) against magnitude of coronal gesture (in % of notional maximum; vertical axis) for all speakers.

The pattern of distribution observed here is once again broadly consistent across all speakers and both languages, and suggests that significant coronal weakening is encountered (predominantly, though not exclusively) in tokens where the interval between mid-points of the coronal and dorsal gestures is less than around 60–70 ms, in other words, in tokens in which the release of the anterior *closure* (and perhaps in some cases also its onset) is masked by the *closure at the velum*. This relationship perhaps allows one to argue that the motive force here is gestural overlap, in line with the Articulatory Phonology account; whereas coronal weakening is a subsidiary effect dependent on the overlapping of consonantal closures.

4.3 Exceptions in English

In the English data, though not the Russian, a significant subset of tokens did not follow the general pattern outlined above, but rather showed total or near-total reduction of the coronal gesture at the same time as *significant lengthening* of the dorsal gesture. These tokens, then, do indeed display compensatory lengthening in a process which is perhaps most aptly considered 'assimilation'. What appears to be the case is that English 'phonologises' the general phonetic effect (sc. of gestural overlap sometimes leading to loss of the coronal) observable in the English and Russian data, by making available to speakers a categorical phonological rule allowing the substitution of a geminate (or partial-geminate) velar for the original alveolar consonant—thereby forestalling the implementation of increased gestural overlap which would have led to a similar result, namely the percept of a velar-only consonantal sequence.

5. CONCLUSIONS

The data considered here give support to the principle embodied in the theory of Articulatory Phonology, that changes in inter-gestural phase relationships may be taken to be a motive force in the derivation of phonetic and phonological patterns. At the same time, it would appear that coronal weakening is licensed by the masking of one stop closure by another, again as a consequence of shifting phase relationships. English, at least, also allows the articulatory effects generated by these forces to play a role in the phonological system.

REFERENCES

[1] Nolan, F.J. (1992), "The descriptive role of segments: Evidence from assimilation", in Docherty & Ladd, *Papers in Laboratory Phonology II*, pp. 261–280. (Cambridge: CUP)

[2] Browman, C.P. & Goldstein, L. (1992), "Articulatory Phonology: An Overview", *Phonetica*, 49:155–180.

[3] Hayes, B. (1992), "Comments on the paper by Nolan", in Docherty & Ladd (*op. cit.*)

FILLED PAUSES IN SPONTANEOUS SPEECH

A. Batliner[1], *A. Kießling*[2], *S. Burger*[1], *E. Nöth*[2]
[1] *Institut f. Deutsche Philologie, L.M.-Universität München, München, FRG*
[2] *Lehrstuhl f. Mustererkennung (Inf. 5), Universität Erlangen-Nürnberg, Erlangen, FRG*

ABSTRACT

Filled pauses as, e.g., uh, eh, signal disfluencies, i.e. hesitations or repairs. They do normally not occur in read speech and were therefore up to now rather seldom investigated; they must, however, be accounted for in the (automatic) processing of spontaneous speech. We present descriptive statistics and the results of an automatic classification of filled pauses in the database of the VERBMOBIL project and discuss the relevancy of different prosodic features for the marking of different types.

INTRODUCTION

Filled pauses (henceforth FPs) as, e.g., *uh, eh*, signal disfluencies and can be classified into

(1) **Hesitations** (FPHs) that are due to planning, control of turn taking, or speaker idiosyncrasies. Functional equivalents are unfilled pauses and hesitation lengthening that is not caused by accentuation or normal preboundary lengthening.

(2) **Cue phrases** (edit signals) for repetitions or repairs of words and phrases, or for restarts of syntactic constructions (FPRs). Functional equivalents are words like *no, that means*, etc. Often, such disfluencies are not marked by cue phrases but only with prosodic means.

Basically, the processing of FPs in human perception/comprehension and in automatic speech processing is analogous: FPHs should be disregarded with respect to linguistic content, FPRs can be taken as cues for a new parse where not only the FP but the reparandum as well has to be disregarded. A full account of these phenomena is given in [1]. In word recognition, FPs are usually only modelled as a waste paper basket category and disregarded. They are often confused with other words. More important than an improvement of word recognition might, however, be the use of FPs for higher linguistic modules as indication of different kinds of phrase boundaries, as an indication for the necessity to start a new parse, etc. It is not likely that FPs can be classified reliably only with spectral features. Several prosodic features are, however, reported in the literature as being relevant for the marking of FPs in English, cf. e.g. the results of [3] and [4]: The F0 of FPs is lower than that of the context, the restart after a FPR is often more stressed than the reparandum before the FPR, FPs at major boundaries are longer than within syntactic constituents.

MATERIAL AND PROCEDURE

Our material was recorded at four different sites for the spontaneous German database of the VERBMOBIL project (domain of appointment scheduling [5]). Because of inconsistencies in the rest of the material, only data recorded at the two sites Karlsruhe and Munich will be used. In total, 2422 turns (339 minutes of speech) from 56 female and 81 male speakers were investigated.

In the basic transliteration, there are four different types of FPs with the following tokens given in SAMPA notation:

<äh>	6:, 6, E:, E, @:, @, 2, 9
<ähm>	6:m, 6m:, 6m, E:m, Em:, Em, @:m, @m:, @m, 2m:, 2m, 9m:, 9m
<hm>	hm, hm:, m, m:
<häs>	pu, pu:, f, f:, pf, pf:, ...

In the transliteration, FPRs can easily be distinguished automatically from FPHs because the disfluencies in their vicinity are labelled separately. The distribution of the four types of FPH and FPR within the 2422 turns is given in Table 1 together with their sum (FP) and, so to speak, their functional complement (*C*). There, either a <Z> denotes a lengthening of the final syllable in a word that is not only caused

Table 1: *Distribution of* FPs

	äh	ähm	hm	häs	FP	*C*
hesitations	471	368	59	70	968	964
repairs etc.	63	23	7	8	101	483

by a following higher syntactic boundary (i.e. 'regular' preboundary lengthening) or repetitions/repairs/restarts are found without FPs. 4% (35 cases) of the FPs are adjacent to <Z>, and 3% (25 cases) to pauses (<P>, 687 tokens) that are labelled if a clear silent interval of more than 0.5 sec can be perceived; 35% (337 cases) of the FPs are adjacent to breathing (<A>, 3001 tokens). 'Adjacent' in this context means 'strictly adjacent' i.e. not separated by any other event. Hesitations are thus almost always signalled either by FPH or by <Z> but not by both. Breathing cooccurs very often with higher syntactic boundaries and thus also with FPs at these boundaries. In the average, almost every second turn or every 19th sec, a FP can be observed. FPs amount to 2% of the vocabulary; in comparison, the most frequent word *ich* amounts to 3%; ca. 85% of the FPs are <äh> and <ähm>. FPHs are roughly ten times more frequent than FPRs. No gender specific difference could be observed as for average length of turns or overall frequency of FPHs or FPRs.

For the prosodic characterization, we used a large set of 47 syllable based features similar to those that proved to be relevant for the automatic classification of phrase boundaries and accents [2]: Duration: (**dur**) in ms and normalized (**durno**) as in [2]; for energy ("loudness"): mean (**enmean**), median (**enmed**), maximum (**enmax**), regression coefficient (**enreg**), and squared mean error of the regression coefficient (**enerr**); for F0, normalized with respect to range (logarithmized) and utterance (mean of utterance subtracted): mean (**F0mean**), median (**F0med**), maximum (**F0max**), regression coefficient (**F0reg**), squared mean error of the regression coefficient (**F0err**), minimum (**F0min**), onset (**F0ons**), and offset (**F0off**); length of pause (**pause**) before and after the (FPs). The features

Table 2: *Percentage of* FPs *at boundaries*

type	position	%
Wi	word internal	0
B0	any other word boundary	6
B1	constituent boundary	25
B2	weak/intermediate boundary	15
B3	strong/phrase boundary	31
Ti	turn initial	14
Tf	turn final	0
R	repair/restart/repetition	9

were extracted for three syllables before the FP (Index $\overline{31}$), the FP itself (Index $\overline{00}$), and three syllables after the FP (Index $\overline{13}$). These features are of course often highly correlated with each other. Their combined use, however, prevents from excluding features that are more relevant than those that might have been chosen by purely phonetic reasoning.

The position of syllable boundaries was computed by an automatic time alignment using a HMM based word recognizer. F0 and energy features were extracted automatically. For paradigmatic comparison, two control syllables with similar phonetic shape were processed as well: [vEm] in *November*, 125 tokens, and [vE:r] in *wär'*, 185 tokens.

RESULTS AND DISCUSSION

In the following, we will disregard the waste paper basket category <häs> because of its varying phonetic substance, and combine the remaining three types. Table 2 shows the distribution of FPs for different positions; the very few Wi and Tf types will be disregarded as well: B0, B1, and B2 constitute the class **FPHweak** at weak, B3 and Ti the class **FPHstrong** at strong boundaries [2]. These types were labelled manually in the transliteration, B2 e.g. in the vicinity of a comma, B3 in the vicinity of a period or a question mark. Final correction of the punctuation in the transliteration and of the labelling of FP types was done by one of the authors; even if these labels are not strictly based on a linguistic analysis, they are thus fairly reliable.

A thorough discussion of the results is beyond the scope (and especially space) of this paper. We will only present the most evident and important findings that are

Table 3: *Automatic classifications*

	constellation of classes	feat.	%
(1)	(B0)(B1)(B2)(B3)(FPR)	$\overline{31}$,$\overline{00}$,$\overline{13}$	46
(2)	(B0 B1 B2)(B3 Ti)(FPR)	$\overline{00}$,$\overline{13}$	55
(3)	(B0 B1 B2)(B3)(FPR)	$\overline{31}$,$\overline{00}$,$\overline{13}$	59
(4)	(B0 B1 B2 B3)(FPR)	$\overline{31}$,$\overline{00}$,$\overline{13}$	68
(5)	(B0 B1 B2)(B3)	$\overline{31}$,$\overline{00}$,$\overline{13}$	77
(6)	(B3)(Ti)	$\overline{00}$,$\overline{13}$	84
(7)	(B0 B1 B2 B3 FPR)(vEm)(vE:r)	$\overline{31}$,$\overline{00}$,$\overline{13}$	85
(8)	(B0 B1 B2 B3 FPR)(vEm vE:r)	$\overline{31}$,$\overline{00}$,$\overline{13}$	91

based on an automatic classification (linear discriminant analysis) where all features were used in a learn=test, forced entry design. Overadaptation takes place with learn=test, and the percent correctly classified can therefore not be taken as a realistic estimate for real life application. We can, however, estimate the relevancy of the features looking at their correlation with the discriminant function, and we can estimate the difference in predictability between those constellations that are given in Table 3 that shows classes to predict, features used (feat.), and percent correct (%). Chance level for the five classes in (1) is 20%, for three classes 33% and for two classes 50%. For Ti-FPs, preceding context, i.e. $\overline{31}$-features, are not available. It was therefore necessary to either exclude these features as in rows (2) and (6) or to exclude this class as in the other constellations from the analysis.

All results in Table 3 are well above chance level. Promising are the results of (7) and (8) because they show that prosodic features really can help in telling apart FPs from other syllables, the most important feature being **durno**, cf. below. In the other analyses, fewer classes result in better classification; that could be expected because the chance level increases as well. We can doubt whether in real life applications, different types of FPs can be told apart with a reasonably high probability but in the long run, not only the prosodic features used can be fed into the analysis but other features as well; e.g. the presence of breathing, cf. above, makes it more likely that a FP belongs to **FPstrong** etc. Even a rather simple language model might be very useful as well. Another factor might be that the database so far is relatively small; more data will hopefully result in a better statistical modelling and thus in better classification rates. (Note that the influence of random errors that always are contained in automatically extracted feature values diminishes if more data are used.)

We want to discuss row (3) in more detail, where (B0 B1 B2), i.e. **FPHweak**, (B3), i.e. **FPHstrong**, and FPR are contrasted. FPR tends to be confused more with **FPHweak** than with **FPHstrong** and vice versa, **pause** being more pronounced for **FPHstrong** than for the two other classes. In Table 4, mean values are presented for the most relevant four cover classes and for most of the features apart from **enmed**, **F0med**, **F0max**, and **F0min** where the relevant information is mostly encoded in other features (mean values or range). For convenience, energy values apart from **enreg** are divided by 10, and **F0range** is multiplied by 100. If we look at these mean values and at the correlation of the features with the canonical discriminant function, we can, with due care, assume that pause, energy and duration features (in this order) are most important for contrasting **FPHstrong** from the other two FPs on the one hand, and on the other hand, that energy and F0 features, esp. **F0reg**$\overline{13}$ and **F0reg**$\overline{31}$, are most important for contrasting FPRs from the two FPH classes. That means that prototypically, **FPHstrong** has longer adjacent pauses than **FPHweak** or FPRs and less energy on the preceding syllables; this finding is plausible as higher syntactic boundaries are expected to be marked with pauses and with a final energy decline. For FPRs, the F0 regression line on the preceding syllables is more falling, and the F0 regression line on the following syllables is more rising than in FPHs. The energy on the adjacent syllables is lower in FPRs than in FPHs. It might not surprise that energy on the following syllables is lower for FPRs than for FPHs even if usually, it is assumed that the reparandum is more stressed than other syllables: energy

Table 4: *Mean values of relevant features for four cover classes*

type	FPHweak			FPHstrong (B3)			FPR			vEm ∪ vE:r		
context	$\overline{31}$	$\overline{00}$	$\overline{13}$	$\overline{31}$	$\overline{00}$	$\overline{13}$	$\overline{31}$	$\overline{00}$	$\overline{13}$	$\overline{31}$	$\overline{00}$	$\overline{13}$
pause	225	—	301	429	—	791	292	—	242	22	—	35
dur	108	215	87	104	193	79	111	201	81	77	65	77
durno	.83	2.32	.11	.59	2.46	.01	.77	2.09	-.00	.01	-.12	.11
enmean	324	367	329	298	359	344	296	354	308	296	414	309
enmax	720	500	705	620	522	744	671	487	679	643	641	651
enreg	-4.49	37.64	16.34	-5.05	-3.00	22.79	-5.82	81.02	22.69	7.81	158.97	-14.54
enerr	1685	416	1416	1370	480	1512	1566	348	1345	1285	536	1301
F0mean	.027	-.042	.017	.005	-.040	.030	.013	-.060	.026	.023	-.008	.010
F0ons	.022	-.026	-.020	-.000	-.032	-.026	.034	-.046	-.041	.031	-.017	.017
F0off	.010	-.045	.030	.013	-.043	.063	-.032	-.060	.058	-.012	.012	.013
F0range	.272	.085	.268	.294	.073	.241	.306	.083	.289	.243	.110	.238
F0reg	-.000	-.078	.070	.036	-.052	.179	-.097	-.117	.226	-.083	.132	-.010
F0err	.574	.104	.462	.574	.090	.409	.636	.090	.519	.413	.095	.367

might be less important for accentuation than duration or F0 features, e.g. the rising F0 regression line after FPRs.

If we compare FPs with the control syllables, the most important feature is duration, regardless whether it is normalized or not. This might be due to the fact that the control syllables are intrinsically rather short, and that we simply have chosen "biased" control syllables. But even without all durational features, classification is only ca. 5% worse than with durational features. That means that the other features encode enough relevant information, most important being the adjacent pauses that are way shorter for the control syllables than for the FPs. F0 values are lower and F0 regression line is more falling in FPs; this finding corroborates the hypothesis that FPs behave like parenthetical chunks that have lower F0 than their surrounding.

CONCLUDING REMARKS

The results achieved for our spontaneous German database are similar to those of e.g. [3] and [4] where English material was investigated. (They are, however, not identical: in contrast to [3], **FPHstrong** is, e.g., not longer than **FPHweak**.) We didn't have a close "phonetically minded" look at some selected features but have tried to include a very large set of prosodic features. The picture that emerges from this data driven approach is possibly more complicated than expected; it is e.g. rather difficult to judge and to explain the relevancy of the different energy features. More data is needed and more space to disentangle matters. But we can expect that very large databases are available in the near future and we hope that with such an approach, the epistemological gap between knowledge based methods (phonetics) and statistically based methods (automatic speech processing) will diminish in the long run.

ACKNOWLEDGEMENTS

This work was funded by the German Federal Ministry of Education, Science, Research and Technology (BMBF) in the framework of the Verbmobil Project under Grants 01 IV 102 F/4 and 01 IV 102 H/0. The responsibility for the contents lies with the authors.

REFERENCES

[1] A. Batliner, S. Burger, and A. Kießling. Außergrammatische Phänomene in der Spontansprache: Gegenstandsbereich, Beschreibung, Merkmalinventar, Verbmobil-Report, Nr. 57, 1994.

[2] A. Kießling, R. Kompe, H. Niemann, E. Nöth, and A. Batliner. Detection of Phrase Boundaries and Accents. In Niemann, de Mori, and Hanrieder, editors, *Progress and Prospects of Speech Research and Technology*, infix, Sankt Augustin, pp. 266–269, 1994.

[3] D. O'Shaughnessy. Locating Disfluencies in Spontaneous Speech: An Acoustic Analysis. In *Proc. EUROSPEECH'93*, Vol. 3, pp. 2187–2190, Berlin, 1993.

[4] E.E. Shriberg and R.J. Lickley. Intonation of Clause-Internal Filled Pauses. *Phonetica*, Vol. 50, pp. 172–179, 1993.

[5] W. Wahlster. Verbmobil — Translation of Face–To–Face Dialogs. In *Proc. EUROSPEECH'93*, "Opening and Plenary Sessions", pp. 29–38, Berlin, 1993.

SUITABILITY JUDGMENTS OF PITCH AS A FUNCTION OF AGE

M. Biemans, R. van Bezooijen and T. Rietveld
University of Nijmegen, The Netherlands

ABSTRACT

Young, middle-aged, and older adult subjects were asked to select the most suitable mean pitch level for speakers of three adult age groups. The subjects chose a high pitch for young people, a low pitch for middle-aged people, and an intermediate pitch for older people. The rated differences between the age groups decreased as subject age increased. A relation is postulated between suitability judgments and stereotypical social characteristics of age groups.

INTRODUCTION

Mean pitch changes with age. This holds not only for childhood but also for adulthood. Various studies have been conducted into pitch change as a function of the age of the adult speaker (e.g. [1] [2] [3]). Most studies agree that the mean pitch of both male and female adult speakers lowers from the age of 20 until middle age. For older speakers the picture is less clear, but most studies seem to agree that the pitch of men rises slightly after they have reached middle age whereas the pitch of females remains constant or lowers slightly.

Pitch change in adult life often is explained by referring to physiological processes. Thick and flexible vocal folds produce a lower pitch than thin and stiff ones. The lowering of pitch until middle age could reflect an increase in the thickness of the vocal folds [2]. The slight pitch rise of older men can be explained by laryngeal processes of ageing which result in thinner and stiffer vocal folds [1] [2]. This ageing process also affects the pitch of older women, but hormonal changes in menopause may cause a lowering of pitch which counteracts this [2] [3].

In addition to being the result of physiological processes, pitch also has a meaning in a social context. There are several ways of describing the social meaning of voice qualities (including mean pitch) [4], of which the sociobiological and the social psychological are particularly relevant for the present study. A proponent of the sociobiological explanation is Ohala [5]. According to Ohala, pitch is used by animals to signal body size, strength, and dominance. A high pitch indicates small, weak, and submissive; a low pitch indicates large, strong, and dominant. This "frequency code" may also be used by humans. High pitch will then be associated with small, insecure, dependent people of relatively low social status, whereas low pitch will be associated with large, secure, independent people of relatively high social status. Graddol & Swann [4] argued that this fits well with a range of social psychological experiments in which higher pitched voices are heard as less competent or less "potent". In line with the scope of their book on gender they point out that such characteristics are components in more general perceptions of masculinity and femininity. These characteristics are also components in general perceptions of specific age groups. Harwood et al. [6] studied young people's impressions of the group vitality of young, middle-aged and elderly people in Hong Kong and California as measured on the dimensions status (e.g. sociohistorical, economic) and institutional support (e.g. in the media, education). They measured, for example, the degree to which the three age groups were perceived to have strength in : advancing knowledge in society; controlling the economy; local and national government; general status of the group in the past, present and future; wealth; home ownership; etc. A difference in ratings for the three age groups was shown: ratings of the middle-aged exceeded those of young and old across both cultures and the Californian subjects rated the elderly higher than the young.

Both pitch as a physiological process and the meaning of pitch in a social context can influence people's impressions of the most suitable mean pitch level for various groups of people. The aim of the present study was to examine which line of approach provides the best explanation for suitability judgments of the mean pitch level ("pitch" from now on) of various age groups. An experiment was conducted in which subjects listened to the same speech fragment at different pitches and indicated which pitch they found most suitable for a person of a certain age. The suitability judgments might vary as a function of the age group the subjects are in, so the experiment was done with adult subjects of various age groups.

METHOD

Speech material

Fifteen men and 15 women from three different age groups served as speakers (see Table 1). For each speaker seven seconds of speech material were selected from an interview about eating habits. Of each speech fragment three versions with different pitches were made by manipulating pitch by means of Linear Predictive Coding (LPC). The speech fragments of the male speakers were synthesized at 98 Hz, 117 Hz, and 137 Hz. For the female speakers the three mean pitch levels were 171 Hz, 195 Hz, and 221 Hz. The three pitch levels will be referred to as low, middle, and high pitch. The frequencies of the pitch levels were determined by first calculating the average mean pitch for all speech fragments of the male speakers and all speech fragments of the female speakers. These average pitches were 117 Hz and 195 Hz, respectively. These were taken as the middle pitches. To determine low and high pitches that were perceptually equidistant to the middle pitches, a psychoacoustic scale was used, the equivalent-rectangular-bandwidth-rate (ERB-rate) scale. This scale tries to approximate the way in which the human ear detects pitch [7]. The ERB values for the two middle pitches were calculated and the low and high pitches were determined by adding and subtracting 0.5 ERB.

Table 1. Speaker age groups.

	RANGE	MEAN	N
WOMEN	20-30	25	5
	48-52	51	5
	73-84	78	5
MEN	21-29	24	5
	47-55	50	5
	71-83	77	5

Subjects

Forty-five men and 45 women from three different age groups participated in the experiment (see Table 2).

Table 2. Subject age groups.

	RANGE	MEAN	N
WOMEN	20-31	24	15
	41-56	50	15
	62-83	73	15
MEN	20-30	24	15
	39-53	45	15
	62-74	66	15

Procedure

The subjects heard the three pitch versions of each speech fragment consecutively and had to indicate on an answering form which one they found most suitable for the speaker. The subjects were given the speaker's sex and age. The speech fragments were presented in two blocks, male and female

speakers separated. Within the blocks of male and female speakers the speakers were randomized. For each speaker the three versions of the speech fragment were also randomized.

RESULTS

The data resulting from the experiment were frequency data; information was available on how often the low, middle, and high versions of the speech fragments were chosen by the subjects. To analyze the data, log-linear analysis, more specifically logit analysis, was used. In this type of analysis the relative contributions of the (interaction between the) variables to the variance of the responses is expressed by R^2, a coefficient similar to the corresponding index of multiple regression, but only in a relative sense. R^2 can yield low values in spite of a strong connection between variables [8] [9]. All significant effects are listed in Table 3.

Table 3. Effects represented by significant parameters ($p<.05$) and their relative contribution to the variance of the responses (R^2). SpA = speaker age, SuA = subject age, SpS = speaker sex, SuS = subject sex.

EFFECT	R^2
SpA	.047
SuA	.003
SpA*SuA	.003
SpA*SpS	.001
SpA*SuS	.001
SuA*SpS*SuS	.001
SpA*SpS*SuS	.001
SpA*SuA*SpS*SuS	.003

Table 3 shows that effects with only age variables have relatively high indexes of association. Correspondingly, inspection of graphical representations of the effects in Table 3 revealed that the interaction between the two age variables (speaker age and subject age (SpA*SuA)) gave the most informative representation of the results of the experiment. The differences in ratings between the male and female subjects and the male and female speakers were negligible.

Figure 1 presents the interaction effect between speaker age and subject age. The scores of the subjects are represented by an index value, similar to indices used in other sociolinguistic research (e.g. [10]). The index value can vary between 0 and 100. Low values indicate that lower pitches are chosen more often, high values correspond with a more frequent choice of higher pitches.

In general the subjects chose a relatively high pitch for young speakers, a relatively low pitch for middle-aged speakers, and an intermediate pitch for older speakers. However, the three subject groups differed in the pitch they considered suitable for the middle-aged speakers; the younger the subjects, the lower the pitch chosen. For the young and the older speakers, the scores of the three subject groups are nearly identical.

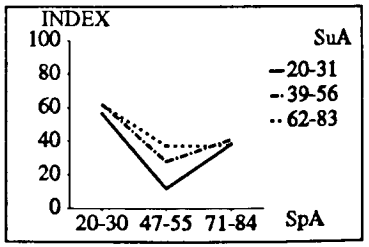


Figure 1. Interaction between speaker age (SpA) and subject age (SuA).

DISCUSSION

From the results of the experiment we conclude that suitability judgments are best explained by referring to the meaning of pitch in a social context. There are two reasons for this. First, the subjects did not differentiate between suitability judgments for men and women, whereas there is evidence that the pitch of older men and women develops differently. The suitable pitch of both older men and women in general was rated higher than the suitable pitch of middle-aged men and women. Because a high pitch is associated with, among other things, a relatively low status, this corresponds well with the study of Harwood et al. [6] in which young subjects rated older people lower on group vitality than middle-aged people (Harwood et al. did not include gender as a separate variable).

Second, the three subject groups differed in their choice of a suitable pitch for the middle-aged speakers. This can only be explained by elaborating on the social meaning of pitch. We hypothesize that when people age the assumed difference in vitality between middle-aged and older people disappears, and the contrast between younger and older people in this respect becomes less salient. The older subjects in the present study, however, were active people who lived entirely or for the greater part independently, so the results of the experiment cannot be generalized without due consideration. More research on the relation between vocal suitability judgments, vitality judgments, and age is needed to confirm the hypothesis.

ACKNOWLEDGEMENTS

The contribution by the second author to the present research has been made possible by a fellowship from the Royal Dutch Academy of Arts and Sciences.

We thank J. van Rie (Department of General Linguistics and Dialectology, University of Nijmegen) who compiled the speech corpus from which the speech fragments used in the experiment were selected.

REFERENCES

[1] Krook, M.I.P. (1988), Speaking fundamental frequency characteristics of normal Swedish subjects obtained by glottal frequency analysis, *Folia Phoniatrica*, 40, 82-90.

[2] Hollien, H. & Shipp, T. (1972), Speaking fundamental frequency and chronologic age in males, *Journal of Speech and Hearing Research*, 15, 155-159.

[3] Stoicheff, M.L. (1981), Speaking fundamental frequency characteristics of nonsmoking female adults, *Journal of Speech and Hearing Research*, 24, 437-441.

[4] Graddol, D. & Swann, J. (1989), *Gender voices*, Oxford: Blackwell Publishers.

[5] Ohala, J.J. (1983), Cross-language use of pitch: an ethological view, *Phonetica*, 40, 1-18.

[6] Harwood, J., Giles, H., Clément, R., Pierson, H. & Fox, S. (1994), Perceived vitality of age categories in California and Hong Kong, *Journal of Multilingual and Multicultural Development*, 15, 311-318.

[7] Hermes, D.J. & Gestel, J.C. van (1991), The frequency scale of speech intonation, *Journal of the Acoustical Society of America*, 83, 257-264.

[8] Haberman, S.J. (1978), *Analysis of qualitative data*, Orlando: Academic Press.

[9] Rietveld, T. & Hout, R. van (1993), *Statistical techniques for the study of language and language behaviour*, Berlin: Mouton de Gruyter.

[10] Trudgill, P. (1974), *The social differentiation of English in Norwich*, Cambridge: Cambridge University Press.

AN APPRAISAL OF RHYTHM AS A COORDINATOR OF TURN-TAKING

M. Bull, matthew@ling.ed.ac.uk
Department of Linguistics, University of Edinburgh, Scotland

ABSTRACT

This paper investigates the notion that perceptual isochrony may be used by participants in a dialogue as a method of timing and coordinating turn-taking. Results from the two experiments reported here provide no evidence for this notion. Instead it would appear that isochronic processes are at best heavily masked by pragmatic and cultural factors, and some form of linear cognitive processing.

INTRODUCTION

It has long been recognised that speech possesses a certain degree of rhythmicity. However, the extent, source and uses of this rhythmicity have over the years been subject to a great deal of debate. Earlier research [1, 8, 11] suggested that relatively strict productive isochrony existed. But these observations may have reflected a different and more plausible process, namely perceptual isochrony. Indeed, it is now beyond contention that listeners tend to impose a regularity on the rhythmic structure of an utterance even where no acoustically measurable regularity exists [4, 6, 7, 9]. The arguments in favour of a cognitive mechanism which organises raw acoustic data into perceptual chunks also appear not unreasonable [10].

Claims have also been made regarding the roles that isochrony might play in the interaction between speaker and hearer [5]. For example, Couper-Kuhlen has specifically cited turn-taking as a possible area for the use of perceptual isochrony. The coordination of turn-taking is extremely fine [12], and it has been suggested by Couper-Kuhlen that a purely linear model of the timing of turn-taking is insufficient to account for this. Instead, a hierarchic model is proposed, where there exists an unmarked case of turn-taking in which a hearer (H) would pick up on the rhythmic structure of a speaker's (S) utterance, such that the first relatively prominent syllable of H's turn would coincide with the rhythmic beat set up by S. Notice that this predicts that the duration of an inter-turn interval (ITI) would be a function of the perceived interval between stressed or prominent syllables in S's turn. However, Couper-Kuhlen's hypothesis remains largely speculative.

In a pilot study [3] I found that when recordings of exchanges with altered ITI's were presented to subjects, they could distinguish between ITI's which were longer or shorter than in the original recording, or of the same length. Having established that such differentiation was at least possible, the two experiments reported here were carried out to ascertain preferred ITI's. I reasoned that if the values chosen by subjects clustered about one or more points this would provide evidence that some very powerful mechanism was at work - one which could have been based on rhythmic principles.

Results from both experiments reported here do not seem to support this hypothesis, however. Instead, the picture appears to involve far more factors than can be accounted for by a rhythmic principle, and the role of isochrony in turn taking is at best one of several possible cues to the end of a turn.

EXPERIMENT I

Method

Twenty-five exchanges, consisting of a turn lasting a few seconds, a natural ITI, and a second turn lasting a few seconds, were selected from the HCRC Map Task Corpus [2], a corpus of spoken task-oriented dialogue. The exchanges chosen did not contain any major disfluencies, and appeared to involve a minimal amount of thinking time on the part of the hearer in the recording. The ITI's occurring in the original recording (hereafter Original ITI's) were well spread between 0 and about 1000ms.

The Original ITI's for each exchange were then altered, being replaced with ITI's between 0 and 1000ms, generated semi-randomly. The artificially generated ITI's consisted of low-volume 'noise', and 20ms of speech adjoining this noise were acoustically tapered to prevent a noticeable click which might otherwise have acted as an unwanted cue to the subjects.

The altered exchanges were presented to each of the subjects in a randomly generated order, over headphones. A transcript was provided to facilitate comprehension.

The subjects were instructed to listen to the presented exchange and to modify the ITI from the duration they first heard (hereafter the Start ITI) until they felt that the ITI was the length it would be in a natural discourse environment (Finish ITI). Subjects were able to make the ITI longer or shorter by pressing keys which altered the duration of the ITI either by 50ms or 150ms as often as they wanted, hearing the exchange with the altered ITI each time a key was pressed. They were also able to repeat the exchange with the unaltered ITI.

In designing the experiment it had been considered that subjects' responses might be influenced by Start ITI's. To account for this possibility, the subjects were split into two groups of fifteen, each group being given a different set of semi-randomly generated Start ITI's.

Results & Discussion

A multiple regression analysis was carried out on the data, indicating that a significant proportion of the variance in Finish ITI's could be accounted for by Start ITI's and Original ITI's (R^2 = 0.3766, F (2, 747) = 225.673, $p < 0.0001$). The results clearly show that Start ITI's had a significant influence on Finish ITI's ($\beta = 0.61$, $p < 0.001$). Original ITI's had no significant effect, however ($\beta < 0.01$, $p = 0.98$).

A comparison was also made between the choices made by the two groups of subjects. Each group was presented with exchanges which had a significantly different set of Start ITI's ($r = 0.02$, $p = 0.924$). One would have expected this difference to have been eliminated if subjects were to rely on a common rhythmic mechanism for determining ITI's. A simple regression analysis of the relationship between Finish ITI's of group A, and Finish ITI's of group B showed no significant correlation between the Finish ITI's of the two groups of subjects ($r = 0.035$, $p = 0.505$), indicating further that Start ITI's must have had a significant bearing on the results.

Finally, one pattern which emerged was the tendency for the Finish ITI's to be grouped within a tighter range than the Start ITI's (Start ITI mean = 499.2ms, sd = 293.75ms; Finish ITI mean = 521.06ms, sd = 248.93ms). It was found that Start ITI correlated with Start ITI less Finish ITI ($r = 0.574$, $p < 0.001$). That is, relatively large or small Start ITI's tended to yield greater alterations on the part of the subjects than did less extreme Start ITI's. Although the effect was not massive, it was large enough to conclude that subjects tended to choose Finish ITI's smaller in range than Start ITI's, and hence tended toward some 'average' ITI.

EXPERIMENT II

Method

The aim of Experiment II was to ascertain the effect that dialogue context has on the task in Experiment I. Instead of hearing only a few seconds of speech either side of the turn transition, subjects were presented initially with approximately ten seconds of speech either side of the turn transition, the exact amount depending on the details of each dialogue. A transcript of each dialogue was provided. The turns immediately surrounding the target transition point were highlighted, so that subjects knew which turn transition in the dialogue to pay attention to. Having heard the entire dialogue, subjects then heard only those turns immediately around the target transition, as in experiment I. Note that the Start ITI in both the first presentation of the dialogue and the first presentation of the target exchange were identical. At no stage was the Original ITI heard.

Results & Discussion

A multiple regression analysis was again carried out, again indicating that a significant proportion of the variance in Finish ITI's could be accounted for by Start ITI's and Original ITI's (R^2 = 0.4519, F (2, 747) = 307.919, $p < 0.0001$). Start ITI's had a significant influence on Finish ITI's ($\beta = 0.65$, $p<0.001$), and that Original ITI's had a

very small, yet significant, influence (β = 0.07, p = 0.01). It would seem, therefore, that given greater context subjects were influenced to a small extent by the Original ITI's, which at no stage did they hear. This suggests that an increased amount of context may yield a greater number of cues to some idealised ITI. In particular, the increased amount of conversation heard by the subjects may have given them a better impression of the rhythmic structure.

A comparison was again made between the choices made by the two groups of subjects. A simple regression analysis of the relationship between Finish ITI's of group A, and Finish ITI's of group B showed no significant correlation between the Finish ITI's of the two groups of subjects (r = 0.1, p = 0.053), although the correlation only just missed the standard 95% significance level. Therefore, while this result indicates as in Experiment I that Start ITI's must have had a significant bearing on the choice of Finish ITI, it also backs up the findings from the multiple regression analysis that Original ITI's had a greater influence than in Experiment I.

As in Experiment I, a tendency emerged for the Finish ITI's to be grouped within a tighter range than the Start ITI's (Start ITI mean = 499.2ms, sd = 293.75ms; Finish ITI mean = 476.38ms, sd=265.34ms).

It was also found that Start ITI correlated well with the difference between Start ITI and Finish ITI (r = 0.498, p<0.001). So, similar evidence emerged as in Experiment I that subjects tended, if only to a small degree, to choose Finish ITI's which were grouped about some 'central' range of ITI's.

GENERAL CONCLUSIONS & DISCUSSION

Experiments I and II lead one to conclude that subjects were primarily influenced by factors other than perceptual isochrony to judge the 'ideal' ITI in exchanges.

Experiment II did reveal a very small yet significant relationship between Original ITI's and Finish ITI's, indicating that a greater degree of contextual information has a small effect on the choice of Finish ITI. This finding is not surprising if one assumes that context plays an important role in all aspects of language. The problem here is to decide whether a greater amount of context would give subjects a better sense of the rhythmical structure of a dialogue, which would be necessary for the notion of rhythm-as-coordinator. But according to this notion, the only rhythmic structure that ought to be necessary are the few beats occurring before the turn transition. These were present even in the low-context situation in Experiment I, where Original ITI's had no significant effect on Finish ITI's. It should also be emphasised that any contextual effect was substantially smaller than the effect of Start ITI's.

One possible objection to these findings is that subjects, rather than altering the ITI's freely until they were completely satisfied that they had reached a 'natural' ITI, felt that they were under some time pressure and chose a value which was not vastly different from the Start ITI. Anecdotal evidence from the subjects would suggest however that this was not the case.

Also, an explanation for the lack of correlation between Original ITI values and Finish ITI values may have been caused by the non-spontaneous nature of the task. That is, that subjects were aware of an upcoming turn transition point through repeated exposure to an exchange, whereas in natural dialogue hearers would possibly be able to detect a turn transition point before it occurred through syntactic, pragmatic, intonational or other cues. Of course, if a rhythmic process were being used in natural dialogue, the results ought not be affected by repeated exposure, since in both natural and artificial situations the same rhythm would be timing the start of the second turn. But overriding these considerations is the observation that in both experiments the Original ITI was at best significantly less of a factor than Start ITI.

An interesting result emerged in the range of Finish ITI's for each experiment. While there were significant correlations between Start ITI's and Finish ITI's, there was evidence that longer Start ITI's produced Finish ITI's which were slightly shorter, and vice versa. Subjects seemed

to be sensitive to different ITI's, in that they were able to detect whether a given Start ITI was particularly long or short. They then attempted to adjust the ITI on this basis. The findings of [3] confirm this by showing a relatively broad tolerance of what constitutes a 'natural' ITI, yet with subjects' being able to recognise long from short ITI's.

Even if there exist rhythmic cues, the claim that they are used directly in timing entry to the floor does not seem to be borne out by this data. It is possible that there are rhythmic processes which facilitate the detection of turn-transition points, even if it is not in a precise enough manner to be detected by these experiments. However, these processes co-exist with a variety of cues signifying the imminent closure of a turn, and possible passing of the conversational floor.

One of the arguments used in favour of rhythmic coordination is that the mutually constraining principles of *earliest possible start* and *intelligibility* [12] would by default yield latching in conversation (latching being that situation where the close of a speaker's turn coincides with the start of a second speaker's turn). However, Couper-Kuhlen points out that this is not generally the case, and reasons that one of the causes for this is that the coordination process is determined rhythmically. But the reasons why participants in a conversation often overlap, or delay their entry to the floor by fractions of a second after the close of a previous turn are, as I hope to have pointed out here, apparently highly complex, involving cultural norms, pragmatic concerns and cognitive limitations, but not necessarily rhythmic factors. The evidence that perceptual isochrony plays anything more than a minor role in timing and coordinating turn-taking is currently thin.

REFERENCES

[1] Abercrombie, D. (1967), *Elements of General Phonetics*, Edinburgh: Edinburgh University Press.

[2] Anderson, A.H., Bader, M., Bard, E.G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H.S. & Weinert, R. (1991), "The HCRC Map Task Corpus", *Language and Speech*, vol. 34 (4), pp. 351-366.

[3] Bull, M. (1994), "Isochrony and the Rhythm of Conversation", Proceedings of the Edinburgh Linguistics Department Conference '94, pp. 5 - 16.

[4] Classe, A. (1939), *The Rhythm of English Prose*, Oxford: Basil Blackwell.

[5] Couper-Kuhlen, E. (1993), *English speech rhythm: form and function in everyday verbal interactions*, Amsterdam: John Benjamins.

[6] Darwin, C.J., & Donovan, A. (1980), "Perceptual studies of speech rhythm: isochrony and intonation", In *Spoken Language Generation and Understanding*, Proceedings of the NATO Advanced Study Institute. Dordrecht: D. Reidel.

[7] Donovan, A., & Darwin, C.J. (1979), "The perceived rhythm of speech", Ninth International Congress of Phonetic Sciences, vol. 1, pp. 268-274.

[8] Halliday, M.A.K. (1985), *An Introduction to Functional Grammar*, London: Edward Arnold.

[9] Lehiste, I. (1977), "Isochrony Reconsidered", *Journal of Phonetics*, vol. 5, pp. 253-263.

[10] Martin, J.G. (1972). "Rhythmic (hierarchical) versus serial structure in speech and other behavior", *Psychological Review*, vol. 79, pp. 487-509.

[11] Pike, K.L. (1945), *The Intonation of American English*, Ann Arbor, Mich.: University of Michigan Publications.

[12] Sacks, H., Schegloff, E.A. & Jefferson, G. (1974). "A simplest systematics for the organization of turn-taking for conversation", *Language*, vol. 50, pp. 696-735.

NON-NATIVE DUTCH: PHONETIC PROPERTIES AND EVALUATIVE JUDGEMENTS

Rianne Doeleman
Research Group on Language and Minorities, Tilburg, The Netherlands

ABSTRACT

This paper reports on the evaluation of the speech of Dutch first-generation adult immigrants by native Dutch listeners. An attempt is made to explain speaker evaluations on the basis of both suprasegmental deviancy from standard (non-accented) Dutch and ethnic group evaluations.

INTRODUCTION

The results presented here are part of the research project "Native judgements on non-native Dutch". The components in the project and the relations between them are represented in Figure 1. The present exploratory study is concerned with three questions. 1. To what extent do suprasegmental phonetic features in non-native Dutch deviate from native Dutch? (i.e. a partial description of the *input* in Fig. 1). 2. How do native Dutch speakers evaluate the personality characteristics of non-native speakers? (i.e. the *output*). 3. Are these speaker evaluations based on the suprasegmental deviations or on the social judgements and stereotypicalviews the native Dutch judges hold about ethnic groups? (i.e. what *triggers* the output).

METHOD

Speech material

Four ethnic groups were selected from the Dutch multi-ethnic society: three non-native groups and one native control group. Each non-native group was represented by two countries of origin. The native group contained both speakers with a regional accent and those speaking the standard variant (see table 1, next page). The speakers selected and interviewed were 18 to 35 years old male. They were attending (or had attended) higher education. They all spoke the language of their country of origin as their mother tongue (for the Moroccans and Surinamese in this study this was Moroccan-Arabic and Sranan).

For each speaker an identical text fragment of 15 sentences was selected from a reading text. Each fragment lasted approximately one minute. The text was about a family having a car problem.

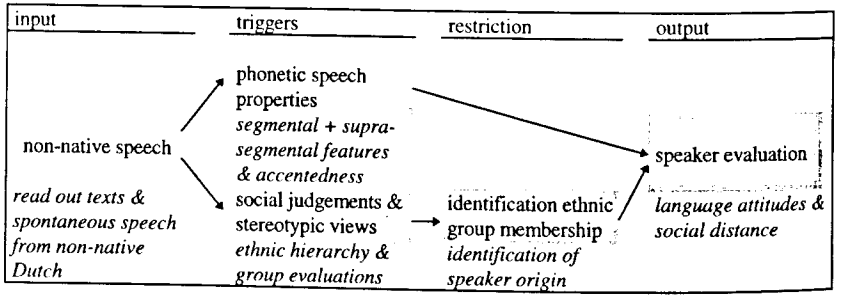


Figure 1. Components in native judgements of non-native speech.

Table 1. Speaker origin and number (N=number of speakers used in the present study).

	ethnic group	origin group (country)	N
non-native Dutch	Mediterranean	-Turkey	2
		-Morocco	2
	former colonial	-Surinam	2
		-the Antilles	2
	West-European	-England GB	2
		-Germany	2
native Dutch		-Brabant	1
		-Randstad	1
		-Standard	1

The (read out) speech fragments were identical at the morphological, syntactic, and lexical level. Thus, any differences between fragments as well as the evaluation of fragments can be attributed to the phonetic differences between speakers and speaker groups.

Judges and judgements

The speech fragments were presented to three types of judges. These represent the three central components in the Brunswikian lens model, used by Scherer to describe the operation of personality markers in speech: distal cues, percepts and attributions [1]. Firstly, three phonetic experts rated the phonetic features of the speech fragments at the suprasegmental level: the *distal cues*. The method used to describe the suprasegmental features was based on Laver's phonetic description of voice quality [2]. The experts made an auditory description of the settings for *pitch, tempo, loudness, articulation* and *voice* on a total of 25 bipolar scales. Two more scales were added to evaluate *intelligibility* and the strength of *non-native accent*.

Secondly, a group of 15 language students at the University of Tilburg judged the salience of several phonetic features: the *percepts* of the distal cues. Due to space limitations these results will not be presented here. However, they will be attended to at the poster.

The third group of judges were 67 other language students, who reported the *attributions* of personality characteristics on 13 bipolar semantic differential scales. The scales represent the three dimensions which, according to previous studies, underlie language attitudes: *attractiveness, status* and *social distance between speaker and rater* [3]. These judges also rated their social judgements and stereotypical views of non-native groups. The group evaluations were rated on the same differential scales as the speaker evaluations. Attractiveness, status and social distance were judged for the *average member of group x*. No stimulus tape was played. The ethnic groups filled in for 'group x' were those represented in the speaker evaluation: non-native Turks, Moroccans, Surinamese, Antilleans, Germans, and native Dutchmen. The English happened not to be represented here.

RESULTS

Suprasegmental phonetic properties

To attend the first question in this study, suprasegmental deviancy from standard, non-accented Dutch was computed. Mean deviancy scores were computed for intelligibility, non-native accent, pitch, tempo, loudness, articulation, and voice. As was expected, deviancy from standard Dutch was stronger for the non-natives than for the natives. Obvious differences were found for intelligibility, accent, pitch, tempo and articulation. Within the non-native groups suprasegmental settings seem to be personally defined, not depending on the mother-tongue of the speakers. The weak relation between origin and suprasegmental deviancy may be because each country of origin was represented by only two speakers

Speaker evaluation

The speaker evaluations were given by the third group of judges (lay people, n=67).

The speaker evaluations were performed to answer the second question in this study. It was found that differences between ratings of speakers' personality characteristics were significant. The results are presented in rankings on the underlying dimensions (attractiveness, status and social distance) in table 2.

Table 2. Rankings of speaker evaluation (most positive = 1, most negative = 15).

	attractive-ness	sta-tus	social distance
Turk-1	14	11	13
Turk-2	13	15	15
Moroccan-1	8	14	10
Moroccan-2	6	13	12
Surinam-1	7	6	8
Surinam-2	1	8	1
Antillean-1	5	10	5
Antillean-2	2	12	6
German-1	15	3	14
German-2	10	7	9
English-1	3	4	3
English-2	11	9	11
Brabant (South)	4	5	2
Randstad (West)	12	2	7
Standard Dutch	9	1	4

Within the native group status and attractiveness are opposites. This finding is in accordance with previous studies on language attitudes [3]. However, this is not found in the rankings of some non-natives; both Turks, for example, are judged negatively on all three dimensions.

It is also striking that the Dutch are *not* consistently rated more attractive and at closer social distance than all non-natives, which was expected on the basis of socio-psychological research [4]. The Surinamese and Antillean speakers are rated equally (or even more) attractive and close to the raters as the in-group: the Brabant natives.

Group evaluation

Differences between ratings on the three dimensions were significant; the results are presented in ranks in table 3.

Table 3. Rankings of group evaluation (most positive = 1, most negative = 6).

	attractive-ness	status	social distance
Turk	3	5	5
Moroccan	5	6	4
Surinam	1	3	3
Antillean	2	4	2
German	6	1	6
Dutch	4	2	1

The rankings show a clustering of evaluations of non-natives belonging to the same ethnic group (i.e. Mediterranean and former colonial). Non-native Dutch groups seem to be evaluated differently according to where they were born and raised. However, as table 3 indicates, the difference in ethnic group membership might be more important than the actual country of origin.

TRIGGERS IN SPEAKER EVALUATION

The third question to be answered in this study was (a) whether the suprasegmental deviations have an effect on the speaker evaluations, and/or (b) whether social judgements of ethnic groups determine the evaluation of a speaker of this group. The answer may be found in the correlation matrix, which is presented in table 4. (N.B. the suprasegmental scores of the English have not been incorporated here because there were no group evaluations on them.) It can be seen that attractiveness in speaker evaluation is closely related to attractiveness in group evaluation. In the evaluation of the speaker's status both the group evaluation of status and some of the suprasegmental features seem to be important (i.e. intelligibility, accent, pitch and tempo). The high correlation with social distance is caused by internal correlation of group evaluations. The evaluation of social distance appears to be related to the social distance towards the groups, and the ratings on intelligibility and strength of accent.

*Table 4. Correlation among speaker evaluation, suprasegmental deviancy, and group evaluation (two-tailed significance at *:5% and **:1% level).*

	attrac-tiveness	status	soc.dis-tance
intelligibility	.42	.56*	.62*
accent	-.06	-.69**	-.58*
pitch	.19	-.72**	-.22
tempo	.03	-.69**	-.39
loudness	-.41	-.03	-.20
articulation	.07	-.25	-.32
voice quality	.27	-.17	.12
group eval.:			
-attractiveness	.69**	-.18	.51
-status	-.21	.84**	.22
-soc. distance	.29	.61*	.73**

To gain more insight into the correlations, a regression analysis was performed. In the equation for speaker's attractiveness, only group attractiveness was included (R=.69). The equation for speaker's status included group status as well as intelligibility and voice ratings, resulting in a high multiple R (.97). (N.B. intelligibility was strongly correlated to accent, pitch and tempo, therefore these latter features are not included in the equation.) The regression equation of speaker's social distance included only social distance towards the group (R=.73).

DISCUSSION & CONCLUSION

The suprasegmental analyses of the read out text do not show that differences in suprasegmental deviancy from standard Dutch can be related to differences in mother-tongue of the non-native speakers. It might be that using deviancy from standard Dutch is too broad a measurement. At the poster a more detailed analyses of the exact scores on the suprasegmental scales will be presented also. On the other hand, it may be that most distinctive features between ethnic groups are to be found in the segmental analyses, which will be performed in the near future.

The present study indicates that the ideas judges have about the ethnic groups are most important in determining the ratings on social speaker evaluations. Suprasegmental deviancy does not seem to have a large influence on the speaker evaluations. However, it was found that ratings of speaker's status increase as intelligibility and voice quality get better.

In Figure 1 the restriction on using group attitudes in speaker evaluation is the identification of the speaker's origin. In a previous pilot it was found that the origin of speakers could be fairly well identified for the four ethnic groups (Mediterranean, former colonial, West-Europeans and native Dutchmen), but the actual country of origin was identified significantly less well [5]. It is now assumed that when speaker identification is easy, group attitudes form the basis for social speaker evaluations. The (suprasegmental) phonetic features probably influence speaker evaluation when identification of speaker origin is difficult. Suprasegmentals may also be the cause of differences in speaker evaluations between speakers from the same ethnic group.

REFERENCES

[1] Scherer, K.R. & Giles, H. (1978), *Social markers in speech*, Cambridge: Cambridge University Press.

[2] Laver, J. (1980), *The phonetic description of voice quality*, Cambridge: Cambridge University Press.

[3] Cargile, A.C., Giles, H., Ryan, E.B. & Bradac, J.J. (1994), Language attitudes as a social process: A conceptual model and new directions. *Language & Communication*, 14 (3), pp. 211-236.

[4] Hagendoorn, L. & Hraba, J. (1987), Social distance toward Holland's minorities: Discrimination against and among ethnic outgroups. *Ethnic and racial studies*, 10 (3), pp. 317-333.

[5] Doeleman, R. (1994), Identifying the origin of speakers, *Proceedings of the CLS Ph.D. Conference 1993*, pp. 21-36.

PROPERTIES OF FRENCH INTONATION AT FAST SPEECH RATE

Cécile Fougeron and Sun-Ah Jun°**
** Inst de Phon., CNRS URA-1027, Paris III, °* Dept. of Linguistics, UCLA, USA*

ABSTRACT

Effects of fast speech rate on French intonation in text reading are reported. Modifications at fast rate were found in both the shape of F0 curves and the prosodic organization of the text (phrasing and realization of underlying tones), with different patterns across speakers. The results suggest that F0 realization at fast rate is not a simple "speeding up" compared to normal rate but involves a prosodic reorganization.

INTRODUCTION

The realization of intonation patterns has been found to be influenced by several sources of variation, such as pitch range, sentence structure and length [5], or focus. However, very little is known about the effect of speech rate on intonation other than its effect on the number of intonational phrases [2, 8]. As far as we know, no systematic description has been done regarding the phonetic properties and phonological patterns of intonation when its physical domain of realization is shortened by an increase in rate.

So far, the observation of the effects of speaking rate has mostly been restricted to the segmental domain, showing modifications in the temporal, spectral, and articulatory organization of speech. Articulatory data have shown that there are different strategies for overcoming alteration in articulation time; mainly, target undershoot with reduction of the magnitude of movements, and/or increase in stiffness for faster transitions between targets [6, 1]. If we assume that a tune is composed of a sequence of underlying tonal targets [7], we can use the same analysis techniques as used in studies of segmental articulation to observe the effect of fast rate on intonation.

By examining F0 curves, we can test the hypothesis that there is a modification in the physical realization (F0 shape) of the underlying intonation structure as well as a reorganization of the prosodic structure.

METHOD

The text "La bise et le soleil" ("The wind and the sun", given in Table 1) was read by one male and two female Parisian French speakers at self selected normal and fast rates, with three repetitions at each rate. The full text was analyzed for one of the female speakers (1F). For the two other speakers (2F, 3M) the comparison was limited to the first half of the text where most variation between fast and normal rate was found for Speaker 1F.

Acoustic measurements were taken for each tonal pattern at the rising onset and falling offset (hereafter called F0 minima), and the peak (F0 maxima). Duration of accented syllables and intonational phrases were measured.

Qualitative comparison of the prosodic structure is based on the model of French intonation developed in Jun & Fougeron [3] where the lowest tonally defined prosodic level is the Accentual Phrase (/LHLH/), and the highest prosodic level is the Intonational Phrase.

RESULTS AND DISCUSSION

A. Differences depending on the position in the text.

The reduction of duration at fast rate for Subject 1F was about 33% compared to her normal rate. While this reduction is constant for the whole text, the rate acceleration seems to have different effects on F0 depending on the position in the text. In the first part of the text (see Table 1), rate acceleration induced reductions in the shape of F0 curve, as well as some changes in the prosodic organization of the phrases, while in the second part of the text we observed only phrasing differences.

In the first part (Fig. 1), F0 range is reduced at fast rate (32%) with the highest F0 value being lowered and the base line (lowest F0 value) being stable.

In addition, the displacements between L and H targets are also reduced by significantly ($p<.01$) lowering the maxima (17%) more than the minima (9%). Despite this overall lowering of the peaks, the peaks at major prosodic boundaries (#3, 10, 12, 19 in Fig. 1), critical points of prosodic organization, are not reduced. Comparison of F0 movement velocity (displacement Hz/time) shows very little difference between the two rates. The movements are not stiffer at fast rate, with the reduction of displacement being proportional to the reduction of duration.

In the second part of the text (Fig.2), rate acceleration induces a small reduction of F0 range (6%) but no change in displacement with small and equal degree of lowering of maxima (5.8%) and minima (5.5%). The velocity of F0 movements are not affected.

Considering the prosodic organization of the text, we found there is a change in the realization of tonal patterns at fast rate: in an Accentual Phrase /LHLH/ [3], the initial underlying high tone (initial accent) is often not realized (63% and 73% for 1st and 2nd part). We also found fewer prosodic groups at fast rate. Intonational Phrase boundaries were reduced to lower level boundaries (50% and 42% for the first and second part), and some of the lowest boundaries (AP) were deleted (22% for both the first and second part).

In sum, different strategies are adopted in the first and second parts of the text. When comparing the range of variation of F0 targets in the two parts, evidence is found for a saturation effect. In the first part, the F0 range is wide, which allows some variation in the displacement towards F0 targets without changing the prominence relations between successive targets. In the second part of the text, range of variation is reduced with a smaller F0 range (124Hz compared to 213Hz in the first part). A second indication of saturation is found in the fact that the lowering of the maxima at fast rate is proportional to their height at normal rate: as shown in Fig. 3, in both the first (empty square) and the second (circle) part of the text, the higher the peak is, the bigger the reduction is at fast rate (R^2=0.28 at both positions).

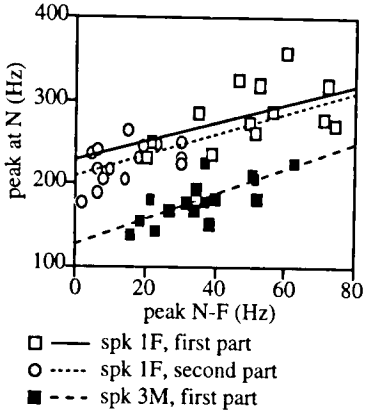


Figure 3: Lowering of F0 peaks between normal (N) and fast (F) rate depending of the height of the peaks at normal rate.

B. Differences between speakers

The three speakers show two patterns both in the way of reducing displacement in F0 movement and in prosodic organization.

The male speaker (3M) shows a pattern similar to Speaker 1F (presented above). With a reduction of 44% from the normal rate, this speaker significantly ($p<.01$) lowers his high and low targets. The lowering of the maxima is correlated with the height of the peak at normal rate (Fig. 3, R^2=0.56). Reduction of displacement is achieved by lowering the maxima (17%) more than by lowering the minima (5%). This speaker also shows a similar reduction in the number of phrases (55% IP and 28% AP reduction) and tonal realization (81% of the initial high tones are not realized).

Speaker 2F shows a totally different strategy. With a rate acceleration of 25% between her fast and normal rate, she doesn't change F0 maxima (2%, $p=0.3$) but significantly raises F0 minima (5%, $p<.01$), compared to the normal rate, as shown in Fig. 4. There is also no difference in velocity or pitch range. This speaker also differs from the others in that she doesn't modify the prosodic organization of the text. At

fast rate, the phrasing is exactly the same as in the normal rate, and the tones are realized as they are in the normal rate.

Thus, at fast rate, this speaker obtains displacement reduction by undershooting her low targets, keeping the high constant, while the other speakers reduce their displacement by lowering high targets.

When we compare the distribution of F0 targets within the range of the speakers, we can observe that for speaker 2F, the targets are equally distributed around the mid-range. For speaker 1F, the dispersion is more concentrated (74% of the L and H targets) in the lower part of her range. Since she uses the lower part of her pitch range more often, movements toward high targets located in the upper range require more energy than the displacement within the restricted lower region. Thus, when the time is limited, at fast rate, the targets at the extreme end of the upper range would be more likely to be reduced (cf. [4]). For speaker 2F, the interpretation is less clear; since she uses both parts of her pitch range equally, her target distribution doesn't provide any indication of a preference for reducing high or low targets. Therefore, there must be other factors to account for her raising of the low targets. Since high targets (mostly boundary tones) are linguistically more important in the prosodic grouping of our text, we can expect she may prefer to modify low targets. This factor may not be the main constraint for speaker 1F whose reduction of the high targets does not jeopardize their prosodic salience.

It should be also noted that the occurrence of target lowering is not determined by the extent of a speaker's pitch range. For example, although the male speaker 3M has a rather small pitch range (139Hz), he shows a high degree of displacement reduction when he accelerates his rate. In contrast, speaker 1F shows little reduction in the second part of the text, where her pitch range is reduced to a value (124Hz) similar to the male speaker's. Thus, the saturation effect we found for speaker 1F must be due to the reduction of her full pitch range at the end of the text, but not because of a small range itself.

CONCLUSION

In this paper, we showed that an acceleration of rate induces a reduction of pitch range and displacement between H and L targets, but no change in the velocity of F0 movements. Moreover, at fast rate, lower prosodic units are regrouped into higher prosodic units leading to fewer phrases, and some of the underlying tones are not realized. We also found that the effects of an acceleration of rate vary across speakers, and depend on the nature of the target tone (H or L), its position relative to the pitch range, its linguistic function (strength of boundary), its position in the text, and the distance between target tones.

ACKNOWLEDGEMENT

The first author is supported by Allocation de recherche M.R.T. granted to the D.E.A. de Phonétique de Paris.

REFERENCES

[1] Gay T. (1981). "Mechanisms in the control of speech rate", *Phonetica* 38: 148-158.

[2] Jun, S.-A. (1993) *The Phonetics and Phonology of Korean Prosody*. PhD. diss. The Ohio State Univ.

[3] Jun S.-A. & Fougeron C. (1995). "The Accentual Phrase and French Prosodic Structure". (this conference)

[4] Kuehn D.P. & Moll K. (1976). "A cinefluographic investigation of CV and VC articulatory velocities", *J. of Phonetics* 3:303-320

[5] Liberman M. & Pierrehumbert J. (1984). "Intonational invariance under change in pitch range and length", in M. Aronoff and R. Oehrle (eds.) *Language Sound Structure*, MIT Press.

[6] Lindblom B. (1963). "Spectrographic study of vowel reduction", *JASA* 35 (11): 1773-1781.

[7] Pierrehumbert J. (1980). *The Phonology and Phonetics of English Intonation*. PhD. diss. MIT.

[8] Vaissière J. (1986). "Variance and invariance at the word level" in Perkell J. & Klatt D. (eds.) *Invariance and Variability in Speech Processes*, L. Erlbaum Assoc., 534-539.

Table 1: Text "La bise et le soleil" and boundary codes used in figures 1, 2, and 4.
First part: La bise (1) et le soleil (2) se disputaient (3), chacun (4) assurant (5) qu'il était (6) le plus fort (7). Quand ils ont vu (8) un voyageur (9) qui s'avançait (10), enveloppé (11) dans son manteau (12), ils sont tombés (13) d'accord (14) que celui (15) qui arriverait (16) le premier (17) à le lui faire (18) ôter (19) serait (20) regardé (21) comme le plus fort (22).
Second part: Alors (23), la bise (24) s'est mise à souffler (25) de toutes ses forces (26), mais plus elle soufflait (27), plus le voyageur (28) serrait son manteau (29) autour de lui (30). Finalement (31), elle renonça (32) à le lui faire ôter (33). Alors (34), le soleil (35) commença à briller (36) et au bout d'un moment (37) le voyageur, réchauffé (38), ôta son manteau (39). Ainsi (40), la bise (41) dut reconnaître (42) que le soleil (43) était le plus fort (44).

Figure 1: spk1F, first part of the text

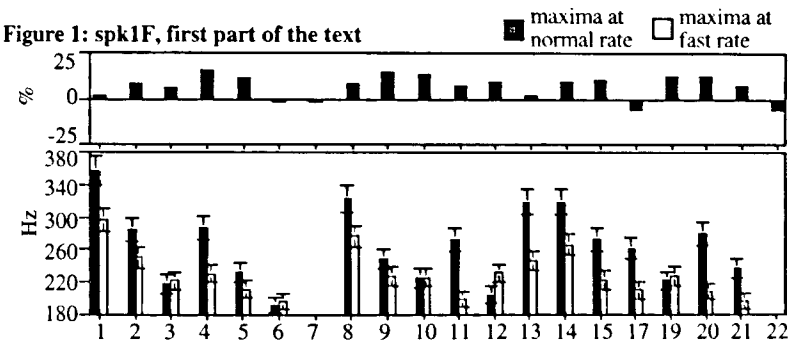


Figure 2: spk1F, second part of the text

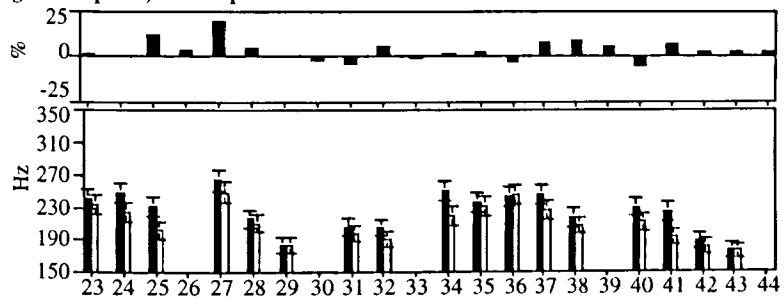


Figure 4: spk2F, first part of the text

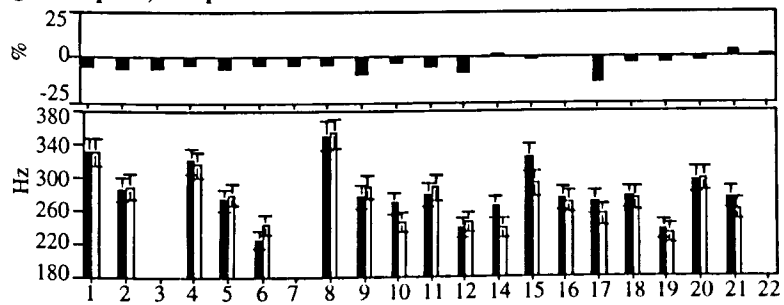


Figure 1, 2, and 4: In each graph, the lower part shows the difference (in Hz.) between normal (black) and fast (white) rate for F0 maxima. The upper part shows the difference (in %) of normal rate (N-F/N) for F0 minima (a positive value = a lowered minima at fast rate).

ACOUSTIC AND PERCEPTUAL EFFECTS OF LISTENER ADAPTIVE TEMPORAL ADJUSTMENTS IN DIALOGUE

S. Imaizumi[1], A. Hayashi[2], and T. Deguchi[2]

[1] RILP, University of Tokyo, [2] Dept. of Education, Tokyo Gakugei University, Japan

ABSTRACT

Effects of listener adaptive temporal adjustments in dialogue were investigated by analyzing vowel devoicing in the speech of teachers directed to hearing-impaired (HI) or normal-hearing (NH) children, and read speech (RD). The teachers did reduce the devoicing rate more in the HI vs. NH and RD samples in such a manner that contrasts between the highly devoicable mora groups versus the others are enhanced within phonological and phonetic constraints of Japanese.

INTRODUCTION

Listener-oriented adaptation of speaking style appears to affect various stages in speech production processes. Our previous analyses [1, 2] of dialogues between professional teachers and normal-hearing (NH) or hearing-impaired (HI) children found that the teachers tended to use simpler and shorter sentences for the HI children than for the NH ones. They also reduced their speaking rates by inserting longer pauses at phonological phrase boundaries and producing longer syllable durations. The teachers also reduce their vowel devoicing, probably to improve the listener's comprehension.

The purpose of the present paper is to elucidate acoustical and perceptual effects of listener dependent adjustments of speaking style by analyzing vowel devoicing in dialogue between teachers and NH or HI children. The main focus was put on the relations between listener dependent adjustments of speaking style and a phonological constraint.

METHOD

Recording of Dialogues

Dialogues were recorded during a simple picture-searching game through which a teacher attempted to assess the speech communication ability of a HI or NH child.

Two different panels were prepared, A or B, with each displaying 11 pictures (illustrations) of boys/girls labeled with their names. The A panel was set in front of the child and a copy of it in front of the teacher. The teacher instructed the child to point to a picture as fast as possible after a name was called out. The teacher randomly called out all the names one by one.

The question was fixed as "Donokoga /CVCVCV/ desuka?" (Who is /CVCVCV/ ?), where /CVCVCV/ represents the name of a picture. If the teacher mistakenly used a different form of question, the sample was not used.

Recording of Read Speech

To clarify the differences between dialogue and read speech, a read list was also recorded and analyzed. Six teachers read the target sentences "Donokoga /CVCVCV/ desuka?" five times at three tempo of fast (RDF), normal (RDN) and slow (RDS). The abbreviation RD is used to represent the read speech tokens.

Analyzed Samples

The names of the pictures consisted of target moras used to analyze the structure of dialogue. Each name consisted of three moras, that is, /$C_1V_1C_2V_2C_3V_3$/, where one mora is formed by one consonant C and one vowel V.

Six types of moras were analyzed, i.e., AffIni, FriIni, $StoMed_1$, AffMed, FriMed, and $StoMed_2$, which represent the manner of articulation of the component consonant (fricatives, affricates, or stops) and mora position (initial or medial). Accent was placed on the initial mora. AffIni and FriIni were the initial moras followed by /ki/, while $StoMed_1$ was /ki/ following the narrow vowels /i/ or /u/. Both the initial and medial moras can be devoiced for AffIni, FriIni, and $StoMed_1$. All the moras in AffMed, FriMed, and $StoMed_2$ were preceded by an open vowel /a/. $StoMed_1$ and $StoMed_2$ were treated separately because devoicing can be different depending on whether the preceding mora can be frequently devoiced ($StoMed_1$) or not ($StoMed_2$).

Subjects

Six professional teachers and seven corresponding HI or NH children participated in the test. All were speakers of the Tokyo dialect of Japanese.

Measurements

All the target moras, totally 2740, were examined using an acoustic analysis system[1]. For each target mora sample, M_m, the length of the unvoiced segment, U_m, and that of the voiced segment, V_m, and their sum, $L_m = U_m + V_m$, were measured. M_m was classified as "Voiced" unless $V_m = 0$. For each sample, the classification variable *Voice* was as either "Devoiced" or "Voiced."

Modeling

Four classification variables (*Mode*, *Mora*, *Teacher* and *Voice*) were defined as follows: *Mode* with five levels (HI, NH, RDF, RDN, RDS), *Mora* with six levels (AffIni, FriIni, $StoMed_1$, AffMed, FriMed, $StoMed_2$), *Teacher* with six levels (TE1 – TE6), and *Voice* with two levels (Devoiced, Voiced). Each mora sample was characterized by the continuous variables U_m, V_m, and L_m, and by the classification variables of *Mode*, *Mora*, *Teacher*, and *Voice*.

A four-dimensional contingency table, F_{ijkl}, was constructed first. F_{ijkl} represents the frequency of samples classified at the cell, C_{ijkl}, of the *i*-th *Mode*, *j*-th *Mora*, *k*-th *Teacher*, and *l*-th *Voice*. The devoicing rate, $Pr(C_{ijkl})$, was calculated as the ratio of the number of devoiced moras versus the total number of moras for each cell, i.e., $Pr(C_{ijkl}) = F_{ijk1} / (F_{ijk1} + F_{ijk2})$.

Two statistical models were constructed, i.e., a generalized linear model (GLM) describing the relationship between the mora length and classification variables, and a logistic regression model predicting the devoicing rate using the mora length and classification variables [3].

Perceptual Analyses

The perceptual characteristics of the tokens were analyzed using the semantic differential method. As previously reported, 24 pairs of adjectives were used as 9-point dipole rating scales. The listening subjects were 8 normal hearing students. The tokens used were 30 samples of /Donokoga hikita desuka?/ spoken by the 6 teachers in the five modes. Obtained rating scores were analyzed by a principal factor analysis, and then a regression analysis was carried out to extract any significant correlations with the temporal structure of the speech.

RESULTS AND DISCUSSION

Mora Length Adjustments

The ANOVA obtained by the GLM procedure showed that L_m was significantly affected by the four classification variables (*Mode*, *Mora*, *Teacher* and *Voice*).

Figure 1 shows bar plots for the total mora length, L_m, with respect to the *Mode* vs. *Mora*. As shown in Fig. 1, the teachers significantly lengthened the moras in speech directed to the HI children vs. the NH children and read speech. FriIni had the longest mora length which was significantly longer than the other mora groups. There was no significant difference in L_m between StoMed1 vs. StoMed2, AffIni vs. AffMed.

Devoicing Rate Variations

Figure 2 shows the predicted logistic regression curves for the devoicing rate of the HI, NH, and RD tokens. As L_m increases, the predicted devoicing rate clearly decreases more for HI than for NH, and even more than for RD, which confirms our previous report [1].

Some common tendencies, however, were observed regardless the modes. The accented initial mora groups, FriIni and AffIni, tended to have a lower devoicing rate than the medial

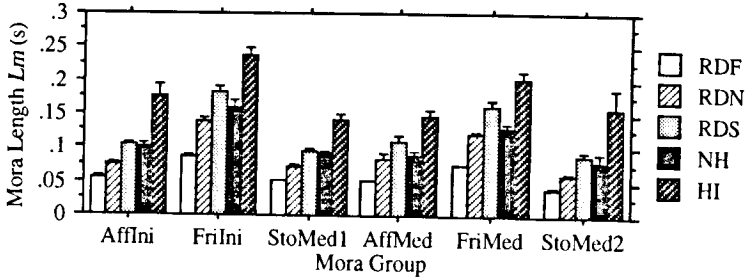


Figure 1. Interaction bar plot for mora length L_m between Mode vs. Mora.

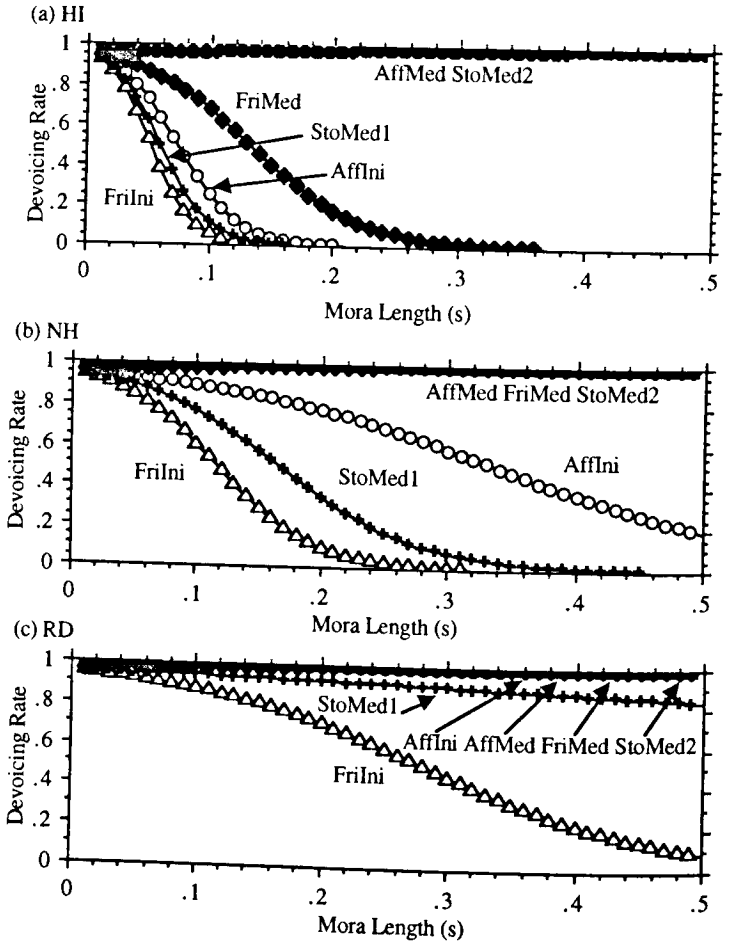


Figure 2. The predicted devoicing rate by the CATMOD logistic regression analysis for the modes of (a) HI, (b) NH, and (c) RD. The mode RD includes RDF, RDN and RDS.

mora groups, FriMed, AffMed and StoMed2. FriIni, the accented initial moras followed by /ki/, had the lowest devoicing rate, while AffMed and StoMed2, the medial unaccented moras preceded by an open vowel /a/, had the highest devoicing rate.

Furthermore, the devoicing rate was significantly different between $StoMed_1$ and $StoMed_2$. Both were the unaccented medial /ki/. $StoMed_1$ was preceded by a high vowel which was frequently devoiced, while $StoMed_2$ by an open vowel /a/ which was seldom devoiced. This significant difference in the devoicing rate cannot be accounted for by the mora length variation because there was no significant difference in L_m. This result suggests that the devoicing rate depends on the devoicing probability of the preceding vowel.

These common tendencies observed regardless the modes may be explained by a phonological rule proposed by Kondo [4]. She showed that vowels tended not to be devoiced consecutively over two moras to avoid creating series of consonant clusters on the surface level, which is not a favored structure in Japanese.

From our point of view, these common tendencies further suggest that the teachers reduced the devoicing rate more in the HI vs. NH samples, and even more in the HI vs. RD samples, within the phonological constrain. The mora groups, AffMed and StoMed2, which were highly devoicable from the phonological point of view were kept devoiced even when the teachers tried to talk carefully to HI children, while the others were highly voiced. The teachers did reduce the devoicing rate so as to enhance the contrasts between the highly devoicable mora groups versus the others within the phonological constraint of Japanese. Connecting the results of our previous report[1], listener-oriented adaptation of devoicing occurred within phonological and phonetic [2, 5] constraints of Japanese.

Perceptual Characteristics

The perceptual profiles of tokens could represented by four factors F1, F2, F3 and F4. F1 represents the contrast between discomfort ("Rough, Uneasy, Busy,") and pleasant ("Easy, Kind, Friendly, Restful, Polite"), corresponding to the perceptual difference between the RD and the other modes (NH and HI). F3 represents the contrast between "Slow, Stiff, Unnatural, Intelligible, Strong" and "Busy, Lifeless, Tense, Rough, Dull," corresponding to the differences between HI and the other modes (RD and NH). F2 and F4 could be interpreted as representing differences among the teachers. These results suggest that listener-oriented adaptation of speaking style produced significant perceptual effects.

CONCLUSION

The teachers did reduce the devoicing rate more in the HI than NH and RD samples in such a manner that contrasts between the highly devoicable mora groups versus the others are enhanced within phonological and phonetical constraints of Japanese. Listener-oriented adaptation of speaking style created significant acoustical and perceptual effects.

ACKNOWLEDGMENTS

Sincere gratitude is extended to all the teachers and students at Ohji Elementary School. This research was supported by a Grant-in-Aid for Scientific Research on Priority Areas of "Spoken Dialogue," Ministry of Education, Science and Culture, Japan.

References

[1] Imaizumi, S., et al. (**1993**). "Listener adaptive characteristics in dialogue speech," *Proc. of ISSD*, (Waseda University Printing, Tokyo), 279–282.
[2] Imaizumi, S., et al. (1995), "Listener adaptive characteristics of vowel devoicing in Japanese dialogue," *J. Acoust. Soc. Am.*, in press.
[3] SAS Institute Inc. (**1989**), *SAS/STAT User's Guide*, Ver. 6, Fourth Edition (Cary, NC, USA).
[4] Kondo, M. (**1994**). "Mechanisms of vowel devoicing in Japanese," *Proc. of ICSLP 94* (Yokohama, 1994), **1**, 61–64.
[5] Jun, S-A, and Beckman, M. (**1993**). "A gestural-overlap analysis of vowel devoicing in Japanese and Korean," *Annual Meeting of the Linguistic Society of America*, (Los Angeles, USA, 1993).

ON THE EFFECTS OF VOCAL TRAINING ON THE SPEAKING VOICE QUALITY OF MALE STUDENT ACTORS

Timo Leino & Päivi Kärkkäinen,
Institute of Speech Communication and Voice Research, University of Tampere, Finland.

ABSTRACT

Male student actors were given as an extra element in their ordinary voice training a special training period of eight months for strengthening of the overtones especially around 3.5 kHz. A spectrum analyzer was used troughout for visual feedback. The reading samples after training had a less steep slope in the long-term average spectrum and the peak at 3.5 kHz was in some cases more prominent. Samples were rated to sound better. After two years the changes still existed.

INTRODUCTION

In earlier studies Leino [1, 2] found that in the long-term average spectra (LTAS) from text reading samples good and poor male actors' voice qualities differed from each other the former having a less steep slope and a strong peak at 3.5 kHz. This peak was named an "actor's formant" by the author. Strengthening of the peak and all overtones was chosen for a goal of a special voice training period included in the ordinary voice training of male student actors. The present article summarizes the results of this project.

MATERIALS AND METHODS

After one year of ordinary voice training seven male student actors on a four year training course for professional actors received an extra 8 month voice training period. A training session of 20 - minutes was given once a week. The ordinary voice training continued at the same time. Special attention was payed to strengthening of the overtones especially around 3.5 kHz. Therefore a real time spectrum analyzer (Spectral Dynamics SD301/SD309) was used to give visual feedback throughout the training sessions. The vocal exercises consisted of nasal-vowel syllable strings produced aiming at a clear, bright, well projecting voice quality.

Before and after the training period the same prose extract of about one minute was recorded in a sound-treated studio using Revox A700 tape recorder and Electrovoice RE11 microphone 40 cm from the subject's mouth. The loudness was kept at normal reading strength on both occasions. Long-term average spectra were made from the text reading samples with a Hewlett-Packard signal analyzer (3561A). A four-hundred point narrow band FFT analysis was used. The frequency span was 10 kHz. Voiceless segments were excluded. The time record length was 40 ms. The display resolution was 25 Hz. The Hanning weighting window with the frequency band of 37.5 Hz was used.LTAS were made of individual samples and averages calculated from individual LTAS with a microcomputer. LTAS were compared according to the slope. For this purpose the spectra were studied on a relative scale where the strongest amplitude peak was given the value zero.

The samples recorded before and after training were played in random order to various groups of listeners including university students and theater and speech professionals.Text reading samples by the students were re-recorded and analyzed after two years of ordinary voice training, where no special attention was paid to strengthening of the overtones and the spectrum analyzer was no longer used in training.

RESULTS AND DISCUSSION

Figure 1 compares average LTAS (a) before and after special voice training and (b) before the special training period and two years after it was over and only ordinary voice training was given. The figures show only the frequency range 0-4 kHz, since no significant changes were observed above this. It can be seen that on average the spectral slope became less steep and the peak around 3.5 kHz became somewhat more prominent after the special training period (Fig. 1 a). After two years these characteristics were mainly still seen but weaker (Fig. 1 b).

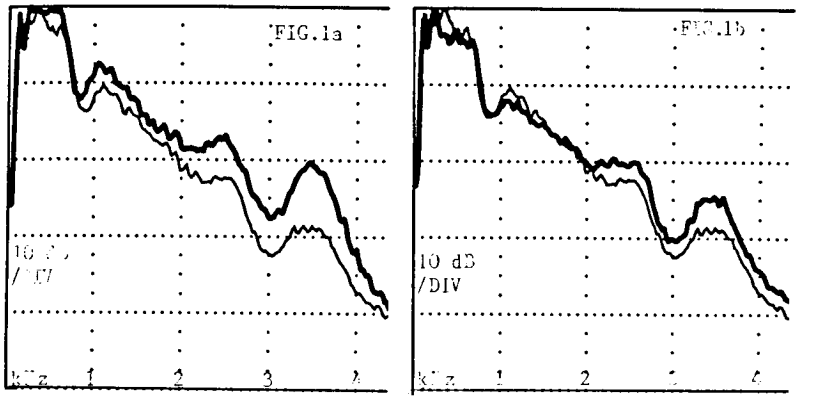
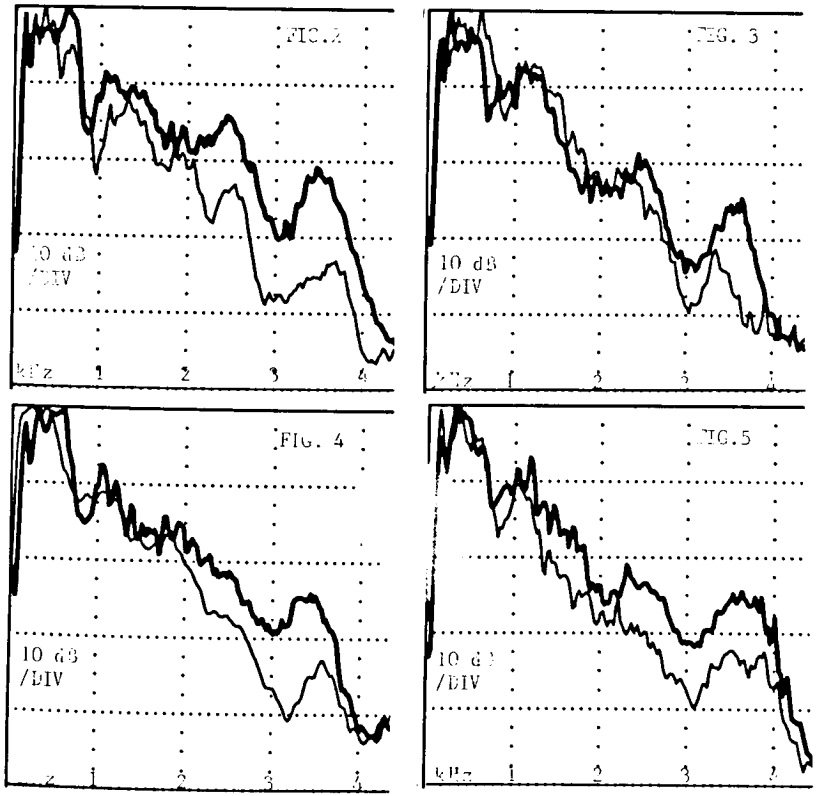


Figure 1. Average LTAS (a) before (thin line) and after (thick line) the eight-month special training, (b) before (thin line) and two years after the special training (thick line).



Figures 2-5. Changes in the LTAS of four individual student actors after the eight-month special voice training. Thin line = before, thick line = after.

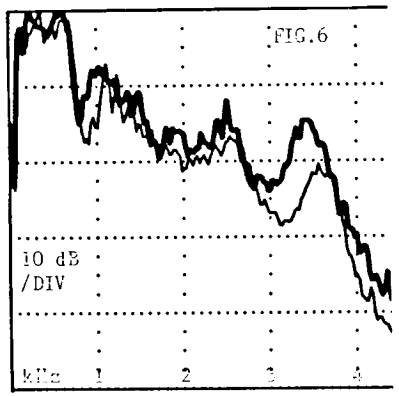
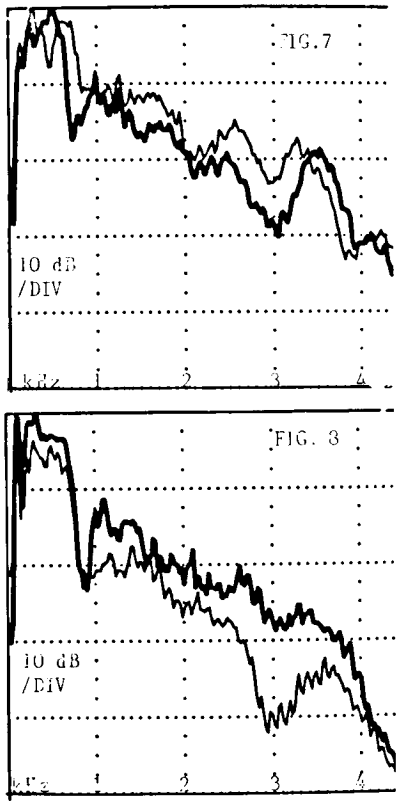


Figure 6. LTAS of one student, whose voice quality after the special voice training (thick line) received conflicting evaluations.

Figures 2-8 compare LTAS before and after the training period for each student. After the training in all but one case the slope became less steep and the peak around 3.5 kHz slightly more prominent in four cases. Only in two cases the peak of 3.5 kHz was as strong as the peak of 2.5 kHz or stonger than that, which according to the earlier findings of Leino [1]. seems to be one characteristic feature of a very good voice.

The text samples read after the training period were in the listening tests evaluated to sound better. Only in one case (Fig. 6) the listeners disagreed. Obviously that student, who had before the training period already had strong overtones and a prominent peak at 3.5 kHz, had tried too hard to make his voice even better with the result that the overtones became too strong for some of the listeners. In his LTAS the difference between the strongest spectral peak and the peak at 3.5 kHz was only 14 dB while for other students this difference was 20-25 dB after training.

Leino [2] has earlier found that if a speaking voice sample is altered by filtration, the voice quality is evaluated to be better if the amplitude difference between the 3.5 kHz peak and the strongest spectral peak is about 15-30 dB. The voice quality rating is impaired both when this difference increases and when it decreases.



Figures 7 and 8. LTAS of two students before (thin line) and after (thick line) the training period.

Another student (Fig. 7) behaved in the opposite way. He also had already rather strong overtones, but he did not try too hard to make them stronger. Instead, most likely he tried to change his voice quality in the exercises through resonatory changes by only increasing the prominence of the 3.5 kHz peak. Similarly the student whose LTAS can be seen in Fig. 8 seems to have changed his resonatory setting, which has contrastively led to the disappearance of the peak.

In general the spectral changes related to voice training and improvement of voice quality may be explained from the basis of both phonatory and resonatory changes. The spectral slope is known to be related to the glottal closing speed so that the increasing closing speed gives a less steep slope [3]. Also Frøkjær-Jensen & Prytz [4] and Wedin et al. [5] have found that vocal training decreases the spectral tilt and increases the perceptual "sonority" of the voice quality. However, the clear peak around 3.5 kHz which has been found to be one characteristic of a very good male voice quality [1-2], seems also to require resonatory bases. The valleys separating the peak from its surroundings suggest that it is formed by a formant or a cluster of two or more formants, most likely F4 and F5. This frequency range has been regarded as more prone to resonatory changes than phonatory ones [6]. Nolan has also found a clear peak at 3.5 kHz in his voice [7]. This peak was especially prominent in creak and creaky voice, largely absent in falsetto, in whispery voice as well as in raised or lowered larynx voice and totally absent in whisper. Nolan considers the possibility that this peak is a phenomenon similar to singer's formant, which, according to Sundberg, [8] is a result of laryngeal resonance arising when the cross-sectional area of the outlet of the larynx tube is sufficiently different from the cross-sectional area of the pharynx.

In conclusion, the results suggest that by vocal exercising it is possible along with consciously set goals to strengthen the overtones and also in some cases increase the prominence of the peak around 3.5 kHz, and that this change in voice quality is perceptually evaluated as positive. Spectrum analyzer seems to be a useful aid in visualising the aims of the exercises to the students. This may make the learning process faster and increase the motivation of the students. The results also show the limits beyond which the strengthening of the spectral peaks and overtones in general is no more perceptually acceptable and most likely also from the voice hygienic point of view questionable. The importance of individually set goals for every student must be emphasized.

REFERENCES

[1] Leino, T. (1976) Hyvän äänen spektripiirteitä. (In Finnish) (The spectral characteristics of good voice). Licenciate Thesis in Logopedics, Department of Phonetics. University of Helsinki.

[2] Leino, T. (1994) Long-term average spectrum study on speaking voice quality in male actors. In *SMAC93 Proceedings of the Stockholm Music Acoustics Conference July 28 - August 1, 1993*, A. Friberg, J. Iwarsson, E. Jansson & J. Sundberg, eds., pp 206-210.

[3] Gauffin, J. & Sundberg, J. (1980) Data on the glottal voice source behaviour in vowel production. *STL-QPSR (Speech Transmission Laboratory, Quarterly Progress and Status Report)*, Royal Institute of Technology, Stockholm, 2-3, 61-70.

[4] Frøkjær-Jensen, B., Prytz, S. (1976) Registration of voice quality. *Brüel & Kjær Technical Review* 3, 3-17.

[5] Wedin, S., Leanderson, R. & Wedin, L. (1977) Evaluation of voice training, In *Proceedings of the International Association of Logopedics and Phoniatrics, Ed. N.H.Buch, Copenhagen 1977*, pp. 361-381.

[6] Blomberg, M. & Elenius, K. (1970) Statistical analysis of speech signals.*STL-QPSR* 4, 1-8.

[7] Nolan, F. (1983) *The phonetic bases of speaker recognition.* Cambridge: Cambridge University Press.

[8] Sundberg, J. (1974) Articulatory interpretation of the singing formant. *Journal of the Acoustical Society of America* 55, 838-844.

THE EFFECT OF REGISTER VARIATION ON THE PERCEPTION OF THE FRENCH /w- ɥ/ DISTINCTION BY NATIVE SPEAKERS OF AMERICAN ENGLISH

Andrea Levitt
Wellesley College, Wellesley, MA, USA

ABSTRACT

The words *Louis* /lwi/ and *lui* /lɥi/ were produced by a female native speaker of French in three registers: Native Talk, Foreigner Talk and Child Talk. Perception of the /w-ɥ/ contrast by Americans who had never studied French was significantly worse in Child Talk than in the other two registers. Acoustic analyses of the stimuli suggest that these results may be due to significant F0 and formant differences in Child Talk as compared to the other two registers.

INTRODUCTION

Forms of speech that vary as a function of the addressee are frequently referred to as speech styles or registers. For example, speakers often modify their speech for listeners whose linguistic competence is in question. Such listeners include both young children learning their first language as well as older individuals learning a second language. Speakers frequently simplify grammar and vocabulary and also make prosodic and phonetic adjustments when addressing such listeners [1-4].

Although some researchers believe that such modifications can aid the language learner [5], few direct tests of the effects of register variation on language comprehension have been conducted [6], especially on the effects of register variation on the perception of phonetic contrasts, although there is at least one such study showing such an effect with infants [7].

The present study was thus designed to investigate the effects of speech style variation on a nonnative phonetic contrast that is normally difficult for adult Americans who are second-language learners of French. The phonetic contrast chosen was /w-ɥ/, as in the words *Louis* /lwi/ and *lui* /lɥi/. If, in fact, register variations do aid the language learner, then the discrimination of this contrast by nonnative adults ought to be better when the tokens tested are produced as Child Talk (CT) and as Foreigner Talk (FT) than when produced as Native Talk (NT). In Part I, we report the results of a perceptual test of this hypothesis. In Part II, we describe the results of an acoustic analysis that was undertaken in order to see which of the prosodic and/or formant features of the stimuli may have contributed to the outcome of the perceptual test.

PART I: PERCEPTUAL TEST

METHOD

Subjects

Twenty-four female native speakers of American English who had never studied French were paid $6 for their participation in the experiment.

Materials

Eight tokens each of the words *Louis* /lwi/ and *lui* /lɥi/ were embedded in a longer list of French words. These lists were read by a female native speaker of French three times, once as to another native speaker of French (Native Talk or NT), once as to a one-year-old child learner of French (Child Talk or CT) and once as to a nonnative, adult learner of French (Foreigner Talk or FT). The speaker was chosen from among a group of 10 talkers whose speech style variations on two read paragraphs had been acoustically analyzed previously [8].

All the tokens that were used to construct the three AXB tapes, one for each register, had been perfectly identified by three native speakers of French. In an AXB test, three stimuli are presented in sequence, the first (A) and the third (B) representing members of two different categories, here *Louis* and *lui*. The middle item (X) can be from either category, and the subject's task is to decide whether X is a member of category A or B. There were 48 AXB trials in each test, with an equal number of the four possible word orders: AAB, BBA, ABB, and BAA. The first two orders test for effects of primacy, which occurs when subjects perform better when X matches the initial item, whereas the latter two orders test for effects of recency, which occurs when subjects perform better when X matches the last item. The words in each trial were separated by 1 sec and trials were separated by 5 sec. There was a longer pause of 10 sec at the end of each block of sixteen trials.

Procedure

Subjects first filled out a language background questionnaire. Anyone with exposure to French was excluded from the study. Subjects were then told that the test had three parts. In each part, a speaker would pronounce sets of three words. In each set, the first and third words would always be different, even if they sounded very much alike. The middle word would be a member of the same category as either of the first or the last of the three words. The subjects were told to write a "1" on their answer sheets if they thought the middle word was a member of the same category as the first word and a "3" if they thought it was a member of the same category as the last word in the triplet.

The three tapes, CT, FT, and NT, were presented to subjects in a modified Latin square design to control for order effects. After subjects had listened to all three tapes, the experimenter then asked them which of the tapes they had found most difficult and why.

RESULTS

The data were analyzed in an ANOVA with one between group factor (Order) and two within group factors (Register and Primacy vs. Recency). The main effect of Register was significant [$F(2,42)=7.688$, $p=.0014$]. Subjects' responses were 86% correct for NT, 85% correct for FT, and 79% correct for CT. Post hoc tests (Newman-Keuls) showed that the results for CT were significantly different from those for the other two registers ($p<.01$), which did not in turn differ significantly from one another. There were no other significant main effects or interactions.

DISCUSSION

The results of the test of the effect of speech style variation on the perception of the /w-ɥ/ contrast by Americans who had never studied French indicate that, contrary to expectations, the CT tokens of *Louis* and *lui* were harder to categorize than the NT tokens. Furthermore, the FT tokens also did not improve subjects' ability to discriminate the contrast when compared with the results for NT tokens. The latter result suggests that the prosodic and phonetic modifications made in FT may not aid subjects' discrimination of difficult nonnative contrasts. However, the fact that subjects found the CT tokens significantly *more* difficult to identify is surprising. Subjects did comment that they found the CT tape more difficult because of the large F0 excursions associated with those tokens. An acoustic analysis of the prosodic and formant characteristics of all tokens used was conducted in order to verify subjects' impressions of the tokens and to see if there were other possible sources for their difficulty with the CT tape.

PART II: ACOUSTIC ANALYSIS

INTRODUCTION

Previous research has indicated that a major acoustic feature distinguishing /w-ɥ/ in French is F2 [9]. However, since it is nonetheless possible that concomitant prosodic and formant differences influenced the perception of this contrast by nonnative speakers, acoustic measurements were taken and submitted to statistical analysis. Our goal is to find a feature of the stimuli that is significantly different for *Louis* and *lui* in the native and FT registers, but not in the CT register.

METHOD

The prosodic measurements made on the stimuli included duration and mean, minimum and maximum F0. F0 range was calculated as the percent increase over minimum F0 represented by the difference in the minimum and maximum values. The first and second formant for each phonetic segment was also measured.

RESULTS

Separate Word (*Louis/lui*) by Register (NT, CT, FT) ANOVAs were run on the measurements for duration, mean F0 and percent increase in F0. (See Table 1). There was a significant main effect of Register in the duration analysis [$F(2,40)=41.39$, $p<.0001$]. Post hoc tests (Newman-Keuls,

p<.01) revealed that the duration of the words in NT was significantly shorter than in the other two registers, which did not differ from one another. There was also a significant main effect for Word [F(1,40)= 32.26, p<.0001], with *Louis* overall longer than *lui*. There were no significant effects or interactions in the mean F0 analysis. In the F0 range analysis, there was a significant main effect of Register [F(2,39)= 25.30, p<.0001]. Post hoc tests revealed that all registers were significantly different from one another (Newman-Keuls, p<.05).

Table 1. Mean values for prosodic measurements of Louis (1) and lui (2) in the three registers.

Register and Word	Duration in ms	Mean F0 in Hz	Percent increase in F0
NT 1	511	221	83
NT 2	417	212	71
FT 1	656	227	99
FT 2	571	209	111
CT 1	696	219	148
CT 2	585	220	166

Separate Word (*Louis/lui*) by Register (NT, CT, FT) ANOVA were run on the measurements for F1 for each segment. (See Table 2). For /l/, there was a significant Word effect, with the mean F1 for *Louis* higher than that for *lui* (240 vs. 224 Hz), [F(1,40) =5.047, p=.0303]. The values were somewhat lower than expected, perhaps because of the coarticulatory effects of the rounding of /w-ɥ/. For /i/, there was a significant effect of Register, with the means for CT, NT, and FT 330, 271, and 293 Hz respectively, [F(2,40)=5.79, p=.0064]. For the crucial /w-ɥ/, there was also a significant Word effect, with mean F1 again higher for *Louis* (346 vs. 269 Hz), [F(1,40)=34.086, p<.0001]. More interestingly, there was also a marginally significant Word by Register interaction [F(2,40)=2.854, p=.0694]. Post hoc simple effects indicated that the F1 for /w-ɥ/ was different for the two test words for NT and FT (p<.001), but not for CT, precisely the pattern that parallels the perceptual results.

Table 2. Mean F1 values in Hz for the three phonetic segments in Louis (1) and lui (2) in the three registers.

Register and Word	/l/	/w/ for 1 /ɥ/ for 2	/i/
NT 1	239	340	309
NT 2	214	253	278
FT 1	235	373	256
FT 2	233	266	287
CT 1	246	323	352
CT 2	226	289	307

For F2 for /l/, there was a significant effect of Register [F(2,40)=13.330, p<.0001] with means for CT, NT and FT 1900, 1997, 1793 Hz, all significantly different in Newman-Keuls post hoc tests (p<.05). There was also a significant effect of Word with F2 for *Louis* lower (1630 vs. 2163 Hz) [F(1,40)=271.223, p<.0001], and a significant interaction of Word and Register [F(2,40)=10.644, p=.0002]. But here, post hoc tests indicated that the F2 values for /l/ in *Louis* were different for the three registers but the same for *lui* (Newman-Keuls, p<.05). For /w-ɥ/, there was the expected significant main effect of Word with the F2 for *Louis* lower than for *lui* (1121 Hz vs. 2513 Hz) [F(1,40)=287.204, p<.0001]. The high values for F2, particularly for /ɥ/ (see Table 3), may have been due to the effect of the following /i/. There were no other significant main effects or interactions for /w-ɥ/ or /i/.

Table 3. Mean F2 in Hz for the three phonetic segments in Louis (1) and lui (2) in the three registers.

Register and Word	/l/	/w/ for 1 /ɥ/ for 2	/i/
NT 1	1831	1023	2645
NT 2	2164	2518	2589
FT 1	1445	1189	2706
FT 2	2139	2465	2598
CT 1	1614	1027	2691
CT 2	2185	2555	2706

DISCUSSION

It is not clear exactly what role each of the features showing a Word effect (duration, F1 and F2 for /l/, and F1 and F2 for /w-ɥ/) played in aiding subjects' discrimination of *Louis/lui*, although the F2 difference for /w-ɥ/, the traditional differentiating acoustic parameter [9], was undoubtedly important in contributing to subjects' above chance performance in all registers. Recall, however, that our goal is to find a parameter that shows a Word by Register interaction, with a significant difference for *Louis* and *lui* in NT and FT but not in CT, thus providing a possible explanation for subjects' lower performance with tokens from the CT register. Duration is not a good candidate for this parameter, because the main effects in the prosodic analyses for Register and Word do not explain the pattern of results across the three registers. Furthermore, although subjects claimed they were distracted by the F0 range in the CT tokens, and F0 variability is hard to ignore [10], the pattern of range differences also does not coincide with the register results.

In the F2 analyses, there was an expected significant main effect for Word for /w-ɥ/. The significant Word effect for /l/ was probably due to a coarticulatory influence of the F2 of /w-ɥ/. Neither effect, however, parallels the perceptual results across registers, which require a Word by Register interaction. Such an interaction was found for /l/, but post hoc tests showed that a pattern of significant differences emerged only for *Louis* across the three registers.

For F1, the Register effect for /i/, while it distinguishes CT from the other two registers, fails to discriminate between the words, and the significant F1 Word difference for /l/ is probably due again to the coarticulatory influence of the Word effect for /w-ɥ/, which in itself is not a traditional discriminating factor. Of particular interest, however, is the marginal interaction of Word and Register for F1, which in post hoc analysis fit our criterion of showing a significant difference for *Louis* and *lui* in the NT and FT, but not in CT. Thus, discrimination of the *Louis /lui* contrast, for which the F2 difference in /w/ and /ɥ/ is undoubtedly very important, may have been enhanced for nonnative listeners for the FT and NT tokens by a small concomitant difference in F1. Interestingly, this F1 formant difference for /w-ɥ/ led to a pattern in which the F1 transition for NT and FT was rising for /w/ into /i/ and falling for /ɥ/ into /i/, whereas in CT the F1 transition into /i/ rose for both /w/ and /ɥ/ (see Table 2). This transition pattern may also be relevant for explaining subjects' performance on the perceptual test, which, contrary to expectations, did not provide evidence for the hypothesis that speech styles addressed to language learners would increase the discriminability of this nonnative phonetic contrast.

ACKNOWLEDGMENT

Thanks to Doug Whalen for his comments and to Lana Makhanik for her help running subjects. This study was supported by grants from NIH (1 R15 HD28173-01) and NATO.

REFERENCES

[1] Ferguson, C. (1964), "Baby talk in six languages", *American Anthropologist*, vol. 66, pp. 103-114.

[2] Ferguson, C. (1975), "Towards a characterization of Foreigner Talk ", *Anthropological Linguistics*, vol. 17, pp. 1-14.

[3] Freed, B. (1981), "Foreigner talk, baby talk, native talk", *International Journal of the Sociology of Language*, vol. 28, pp. 19-39.

[4] DePaulo, B. and L. Coleman (1986), "Talking to children, foreigners, and retarded adults", *Journal of Personality and Social Psychology*, vol. 51, pp. 945-959.

[5] Krashen, S. (1980), "The Input Hypothesis", *Alatis*, pp. 168-180.

[6] Long, M. (1985), "Input and second language acquisition theory", In *Input in Second Language Acquisition*, ed. by S. Gass and G. Madden. Rowley, MA: Newbury House. pp. 377-393.

[7] Karzon, R. G. (1985), "Discrimination of polysyllabic sequences by one- to four-month-old infants", *Journal of Experimental Child Psychology*, vol. 39, pp. 326-342.

[8] Levitt, A. (in preparation), "An analysis of the prosodic characteristics of register variation in read speech in French."

[9] Chafcouloff, M. (1980), "Les caractéristiques acoustiques de /j,ɥ,w,l,ʀ/ en français, *Travaux de l'Institut Phonétique d'Aix*, vol. 7, pp. 7-56.

[10] Wood, C. (1974), "Parallel processing of auditory and phonetic information in speech discrimination", *Perception and Psychophysics*, vol. 15, pp. 501-508.

PHONOLOGY OF NON-NATIVE ACCENTS IN ENGLISH: EVIDENCE FROM SINGAPORE ENGLISH

Paroo Nihalani
National University of Singapore

ABSTRACT

In most Commonwealth countries, it has been fashionable to promote the use of English that has a native-speaker base with everyone being encouraged to speak like a native speaker (Smith 1985). Therefore most of research on nonnative varieties (Bansal 1966; Tiffen 1974; Tay 1982) has sought to identify, in the past, the ways in which a non-native accent **deviates** from a native accent. This paper considers Singapore English (a non-native accent) in its own right, and sets out to attempt a **scientific** description of vowel system in Singapore English (hereinafter referred to as **SSE**) by means of studying the visual sound patterns produced with the help of a DSP sonograph. A comprehensive picture of the acoustic characteristics of vowels in SE based on the quantitative and qualitative analysis of the data will be presented, and some of the areas of its application will be discussed.

1. INTRODUCTION

The consonant system of English is relatively uniform throughout the English-speaking countries. Accents of English mainly differ in terms of their vowel systems as well as in the phonetic realisations of vowel phonemes. Singapore English is not monolithic; it is actually a gradient ranging from speech forms like those of standard English, the *Acrolect*, through the medium range, the *Mesolect*, and to the 'lowest' variety, the *Basilect* (Platt 1977). The variety of Standard English spoken in Singapore has few lexical and syntactic characteristics that set it apart from the Standard English used in England. SSE is, however, spoken with an accent that is slightly different from any other accent of Standard English. This paper deals with Standard Singapore English. The speaker of Standard Singapore English is typically one who has studied in English medium school up to at least GCE 'A' Level, and uses English as his predominant language both at home and at work.

2. TEST MATERIALS

The data on vowels in Singapore English has been collected from 8 subjects (4 Chinese, 2 Malays and 2 Indians) who represent fairly well the proto-typical speaker of Standard Singapore English. The subjects chosen are adult male Singaporeans between twenty and twenty five years of age. Each speaker was asked to read a list of words in the carrier frame "Say C-V-C again" where C represents a consonant and V represents a vowel. The list contained words representing 10 relevant vowels as given below:

1. PETE	6. PUT
2. PART	7. POT
3. PIT	8. PUTT
4. PET	9. PORT
5. BOOT	10. PAT

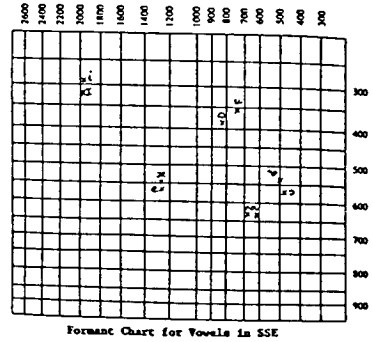
It is hoped that the carrier frame will provide a context and ensure that speech resembles **natural** spoken language. The recording was done under ideal lab conditions in a sound-proof Recording Studio. The subjects were advised to read in their most natural way and at their normal conversational speed. Each speaker read the list of words, repeating each phrase three times. As a result, there were three tokens for each vowel for each of the speakers.

3. INSTRUMENTATION

The use of the sound spectrograph in describing the vowels enables reliable and objective measurements of the vowels based on formant frequencies. Descriptions of vowel quality based on auditory perceptions discussed by Brown (1988) are impressionistic and rather subjective. The two features of tongue height and backness are best 'defined in acoustic terms' (Ladefoged 1982:207).

4. RESULTS AND DISCUSSION

Table 1 shows mean values of F1 and F2'. These mean frequencies of F1 and F2'(the distance between F2 and F1) were computed for all tokens of 10 vowels for all the subjects and have been plotted on the logarithmic scale with F1 on the ordinate, reading downwards on the vertical axis, and F2' on the abscissa, reading right to the left as shown below in the vowel formant chart.



Formant Chart for Vowels in SSE

A vowel is identifiable by its F1 and F2' frequencies. A close examination of the vowel formant chart clearly points to the phenomenon of conflation of some pairs of vowels such as [i] and [ɪ], [e] and [æ], [ɑ] and [ʌ], [ɒ] and [ɔ], and [ɑ] and [u] in SSE. Since vowel segments of each pair tend to cluster together, there seems to be hardly any significant qualitative difference among these pairs of vowels. No wonder, pairs of words like *beat* and *bit*, *set* and *sat*, *cot* and *caught*, *but* and *bart* and *should* and *shooed* very often sound indistinguishable from each other in SSE. Vowel length however is one of the features used, though not consistently, to distinguish these pairs of vowels.

Based on the acoustic results, the vowels in SSE can be classified as follows:

VOWEL	DESCRIPTION	
i/ɪ	high	front
e/æ	low-mid	front
ɑ/ʌ	low	back
ɒ/ɔ	low-mid	back
ɑ/u	high-mid	back

5. CONCLUSION

The present acoustic study, though small in its sample size, provides enough evidence that an SSE speaker fails to maintain sufficient perceptual distance between two vowels in each pair. In English, each of these pairs has a high functional load. If a speaker of SSE fails to maintain this distinction, it could cause a lack of *'comfortable'* mutual intelligibility when a SSE speaker interacts with speakers of other varieties of English.

An acoustic analysis of vowels of SSE is useful in areas such as the codification of Singapore English and Speech Therapy. Besides, language trainers could profitably use these insights in the preparation of teaching meterials and language planning.

REFERENCES

[1] Bansal, R.K.(1966). The intelligibility of Indian English. Ph.D. dissertation, University of London.

[2] Brown, Adam. (1988). Vowel differences between Received Pronunciation and the English of Malaysia and Singapore: which ones really matter. In New Englishes: The case of Singapore, Joseph Foley (ed.), 129-147, Singapore: Singapore University Press.

[3] Ladefoged, P. (1982). A Course in Phonetics. New York: Harcourt Brace Jovanovich, Inc.

[4] Platt, J.T. (1977). The sub-varieties of Singapore English: their sociolectal and functional status. In The English Language in Singapore, W.J. Crewe (ed.) 83-95. Singapore: Eastern Universities Press.

[5] Smith, Larry,E. (1985). EIL Versus ESL/EFL: What's the difference and what difference does the difference make? In English Teaching Forum, 1985.

[6] Tay, Mary. (1982). The phonology of educated Singaporean English. In English World-Wide, Volume 3(2):135-145.

[7] Tiffen, B. (1974). The intelligibility of Nigerian English. Ph.D. Dissertation, University of London.

Table 1: Mean Values of F1 & F2' Frequencies for SSE Vowels

Speakers		1	2	3	4	5	6	7	8	MEAN
i	F1	260	300	300	300	290	250	313	333	293
	F2'	1960	2573	1933	2240	1733	1967	1780	1647	1979
ɪ	F1	270	367	373	307	270	333	280	393	324
	F2'	2033	2427	1887	2500	1780	1853	1847	1633	1995
e	F1	600	560	573	553	547	533	560	580	563
	F2'	1287	1187	1300	1193	1220	1340	1320	1247	1262
æ	F1	607	513	573	553	553	553	533	520	550
	F2'	1227	1260	1247	1373	1247	1273	1280	1280	1273
ɑ	F1	593	613	673	633	640	653	600	667	634
	F2'	540	487	653	613	667	600	620	680	607
ʌ	F1	570	540	707	703	660	620	613	667	635
	F2'	673	647	687	593	717	683	688	720	676
ɒ	F1	580	553	573	540	533	567	507	500	544
	F2'	393	407	547	520	487	587	440	627	501
ɔ	F1	633	560	620	540	580	553	527	520	566
	F2'	347	380	487	400	607	580	487	600	486
ʊ	F1	410	430	390	370	333	360	353	407	381
	F2'	787	813	920	827	747	907	820	807	828
u	F1	360	347	380	373	333	300	320	407	352
	F2'	807	747	673	647	720	867	833	840	766

Phonological rules modelling style variations in French Parisian spontanous speech for text-to-speech synthesis

Péan Vincent

LIMSI-CNRS BP133, F 91403 Orsay-Cedex, France

ABSTRACT

Keywords : phonological variability, text-to-phonemes conversion, speaking styles, spontanous speech, speech synthesis.

The study presented in this paper has been carried out in the framework of phonological variability in French, with applications to automatic speech processing in mind. The specific aim of this study is to both characterize and model intra-speaker segmental variants (at and within word boundaries) in two speaking styles. Data have been collected from casual and careful speech corpus. Examples of phonological rules are given here.

INTRODUCTION

When a speaker tries to change his way of speaking from casual to careful speech, intra-speaker segmental variants can be observed which could be modelized by phonological rules. This implies the collect and the study of both casual and careful speech data: the ICY database (for the study of inter and intra-speaker variability and the characterisation of speaking style) is first described.

The methodology used to analyse data is described. A segmental analysis of the data is then provided in terms of statistical quantification. A comparative and qualitative study of segmental strategies both between two styles for a given speaker and between different speakers for one given style is also presented. Finally, we present examples of phonological rules, with consequences of their modification on synthesized speech.

PRESENTATION OF THE DATABASE

Corpus, task and speakers

Style concept implies choice between several possibilities. Thus, in a given setting and for a given speaker, a modification of the speaker's intention could lead to style variation.

The ICY database [1] has been developped to study inter- and intra-speaker variability which occurs when a speaker try to speak more carefully.

ICY has been recorded to collect three different styles of speech: two spontaneous and one read. Spontaneous speech is considered as non read speech. Here a remark may be noted: the structure of speech of a speaker vary with the context of discourse (setting) and with his psychological state. Thus a lot of different spontaneous speech exist, and to collect speech in a laboratory in a specific context for a specific goal gives one of them. With a view to collecting the data (and to generate a modification of the speaker's performance corresponding only with style variation) a methodology has been developped: the speaker's task is a description of two drawings which differ in some of their parts. Each speaker has to describe each object which differed from one drawing to another, with its colors and spatial positions. A lot of phonological contexts (i.e. where phonological variation could occur) are obtained by constraining the speaker to pronounce them in his description: each object which differ in the two drawings may be, with the constraints imposed during the task, described with groups of words which contain phonological context at word boundaries (for example: robe bleue, context of gemination /bb/). The phonological contexts choosen are: gemination, palatalisation, nasalisation, voicing, devoicing, and schwa.

With a view to obtaining the three different styles, a goal is given to the speaker for all of recordings: this consist of making recordings to help hard of hearing children to learn lip-reading. The speaker goes throught the task three times. The casual speech is collected first when the speaker describes the four drawings just to "rehearse". The careful speech is obtained when the speaker does the "real" recording in front of a camera.
The results presented concerned only four speakers: three female (RF, GS, GM) and one male (BP).

THE SEGMENTAL STATISTICAL ANALYSIS OF VARIANTS

The variation studied

The study is about the phonological variation which occurs in two speech styles (casual and careful) between different speakers. The phonological variation considered corresponds to the variation which leads to a complete modification (i.e. insertion, deletion or substitution) of one or more segmental units which constitute a phonological system of reference of Parisian's speech: GRAPHON[2].

The use of a reference is imposed because to make a straight out comparison between two speakers, the linguistic content of their two recordings must be the same, but that is not the case here because the speech is spontaneous. Thus, by first using a comparison with a reference the results obtained on each of the two speakers could be used to compare them.

The methodology

The analysis method is to do ortographic transcription of the recordings for each speaker, and to use it to obtain by GRAPHON and a specific automatic treatment a homogeneous translation grapheme to phoneme with pauses, word boundaries, and syllable boundaries within word, which will be called the "ideal" phonemic transcription for a given speaker. Then a correction of the ideal phonemic string is done according with what the speaker has really pronounced, by listening and using acoustical representation. Then, the corrected phonemic transcription, which will be called the "real" phonemic string, is compared automatically to the ideal transcription. In this way a characterization of the phonological variation as compared to the reference GRAPHON is obtained for each speaker. For example: the speaker says "il y a un nuage jaune sur le dessin de droite"; by GRAPHON and others semi-automatic treatments the ideal phonemic string obtained is:

#IL#I#A#<#N^AJ$#JON$#SYR#LE#D(~S<#DE#DRWAT$#;

the corrections give the real phonemic string:

#Y#A#<#N^AJ$#:ON#SYR#L#D(~S<#D#DRWAT$#.

The string comparison leads to specific information files[3]. The sharp sign '#' mark the word boundaries; the tilda '~' the intraword syllable boundaries; and the '$' the graphemes 'e' corresponding with linguistic E caducs.
Then the resulting information files are semi-automatically analysed as follow.

The data analysis

Different kind of events are obtained from the automatic analysis: insertion, deletion or substitution of one or more segmental units. To each event may correspond a specific phonological event. For example the deletion of 'E' in the monosyllabic word 'DE' is in fact a schwa deletion.

The results present here concern the deletion of schwa (i.e. linguistic E caduc); and the substitution of one consonant by another consonant correspondind to voicing; devoicing; and palatalisation. For example 'grande table', translated in #GR*D$#TABL$#, may lead to #GR*T$#TABL$#; thus we obtain the substitution of 'D' by 'T' which is in fact a devoicing event in a regressive form (the second phoneme influences the preceding one); and between word.

Some results about the schwa have been given [3] but here the name of schwa is given to linguistic E caduc, which corresponds to a graphemic 'e' in the And in the semi-automatic treatment a distinction have been done between different realisations of E.

For example in the utterance **'Euh** ... ce film(e) tchèqu(e)', four timbres of E caduc may be distinguished. **Euh** is an hesitation vowel; (c)e is a 'true' E caduc (i.e. a linguistic E caduc); (film)e is a non-linguistic E caduc; and (tchèqu)e is a E caduc links to the pronunciation of a final consonant before a pause. [4] [5].

The results given here concern only the 'true' E caduc, but a more complete study will be done on different kinds of E caduc. [6].
The devoicing analysis is about the substitution of one of the consonant {/B/, /D/, /G/, /V/, /Z/, /J/} by the unvoiced corresponding one from the set {/P/, /T/, /K/, /F/, /S/, /X/}. This event is considered in all devoicing context between word (i.e. of type: voiced consonant#unvoiced consonant).
The voicing analysis concerns the substitution of one of the unvoiced consonant {/P/, /T/, /K/, /F/, /S/, /X/} by the voiced corresponding one from the set {/B/, /D/, /G/, /V/, /Z/, /J/}. This event is studied in all voicing context between word (i.e. of type: unvoiced consonant#voiced consonant).

The palatalisation analysis concerns the substitution of one dentale fricative consonant /S/ or /Z/ by the palatale fricative consonant /X/ or /J/. This event is observed in all 'palatalisation context' between word (i.e. S#X; S$#X; etc.).

For these three analysis, two different kind of set of contexts have been considered: a) of type consonant$#consonant (where '$' is a potential schwa); and b) of type consonant#consonant.
Thus for one given analysis and one given type of context, six events are considered.

For voicing regressive inter-word context c1$#c2 (e.g. xxxF$#Dxxx): 1)voicing (i.e. c1 is substitued by c1' which is voiced; e.g. F substitued by V); 2) progressive devoicing (i.e c2 is substitued by c2' which is unvoiced; e.g. D substitued by T); 3)[E]-E caduc 'insertion' (i.e. $ is substitued by E; e.g. F$#D becomes FE#D); 4)nothing (i.e. c1$#c2 is not modified); 5)[E:]- hesitation insertion (i.e. $ is substitued by E: which represents an hesitation realized on linguistic E caduc; e.g. F$#D becomes FE:#D); 6)[p]- empty pause insertion (i.e. there is a pause insertion in the context c1$#c2); but this last event may correspond in fact to several sub-events as 6)a)only pause insertion (i.e. c1$#c2 becomes c1$#p#c2; e.g. F$#D becomes F$#p#D); 6)b)[&]- E caduc 'pre-pausal' insertion (cf. above the E caduc of the word 'tchèqu(e)'); (e.g. F$#D becomes F&#p#D); and 6)c)[E4]- hesitation pre-pausal

insertion (e.g. F$#D becomes FE4#p#D). The distiction between the sub-events have not been done: only pause insertion is considered in the three cases.

The same considerations are done for devoicing regressive inter-word context c1$#c2 (e.g. D$#F) except for the points 1) and 2): 1)devoicing (i.e. is substitued by c1' which is unvoiced; e.g. D is substitued by T); 2) progressive voicing (i.e. c2 is substitued by c2' which is voiced; e.g. F is substitued by V).

The two first points change also for the palatalisation analysis with progressive or regressive inter-word context c1$#c2 (e.g. X$#S or Z$#J): 1)progressive or regressive palatalisation (i.e. c1 (or c2) is substitued by c1' (or c2') which is palatale; e.g. X$#S (or Z$#J) becomes X$#X (or J$#J)).

In the same way, for voicing regressive inter-word context c1#c2 (e.g. xxxF#Dxxx): 1)voicing (i.e. c1 is substitued by c1' which is voiced; e.g. F substitued by V); 2) progressive devoicing (i.e c2 is substitued by c2' which is unvoiced; e.g. D substitued by T); 3)[E3]- non-linguistic E caduc (cf above the E of the word 'film(e)') 'insertion' (e.g. F#D becomes FE3#D); 4)nothing (i.e. c1#c2 is not modified); 5)[E2]- non-linguistic hesitation insertion (E2 represents an hesitation realized on non-linguistic E caduc; e.g. F#D becomes FE2#D); 6)[p]- empty pause insertion (i.e. there is a pause insertion in the context c1#c2); but this last event may correspond in fact to several sub-events as 6)a) only pause insertion (i.e. c1#c2 becomes c1#p#c2; e.g. F#D becomes F#p#D); 6)b)[&2]- non-linguistic E caduc 'pre-pausal' insertion (e.g. F#D becomes F&2#p#D); and 6)c)[E5]- non-linguistic hesitation pre-pausal insertion (e.g. F#D becomes FE5#p#D). The distiction between the sub-events have not been done: only pause insertion is considered in the three cases. And again, the same considerations are done for devoicing regressive inter-word context c1#c2 (e.g. D#F) except for the points 1) and 2): 1)devoicing (i.e. is substitued by c1' which is unvoiced; e.g. D is substitued by T); 2) progressive voicing (i.e. c2 is substitued by c2' which is voiced; e.g. F is substitued by V).

The two first points change also for the palatalisation analysis with progressive or regressive inter-word context c1#c2 (e.g. X#S or Z#J): 1)progressive or regressive palatalisation (i.e. c1 (or c2) is substitued by c1' (or c2') which is palatale; e.g. X#S (or Z#J) becomes X#X (or J#J)).

Results

The first results presented concerned the schwa (i.e the linguistic E caduc defined before) and four speakers: three females (GS, GM, RF) and one male (BP). To illustrate the great variability that occurs for the schwa between different in a given style, the percentage of deletion of E caduc obtained by comparison between the real phonemic string of each speaker for each picture and the corresponding ideal phonemic string obtained by GRAPHON have been plotted. (see figures 1 and 2).

figure 1

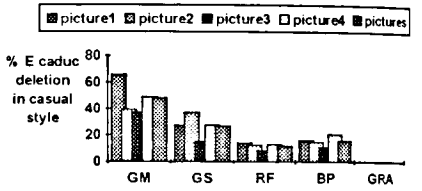
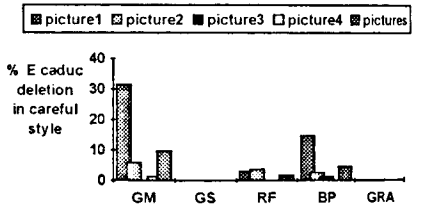


figure 2



The comparison above between two styles for a given speaker shows that a fixed description (given by the rules of GRAPHON here) is not enough to describe speech communication. The GRAPHON rules on the schwa seem to be more appropriate to describe the careful style.

It seems again that the casual style for each speaker is marked by a more important percentage of E caduc deletion than careful style. But the percentage between speakers are very different, thus the phonological rules that will govern this event will be variable.

The second results concerned the voicing defined before. For the contexts c1$#c2 in the two styles (see figures 3 and 4).

figure 3

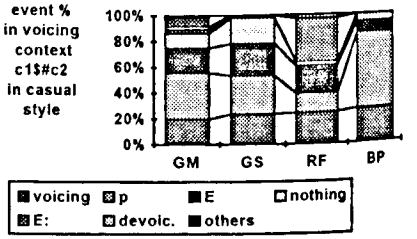
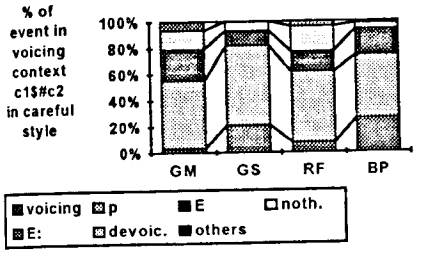


figure 4



Then figures 5 and 6 represent the c1#c2 contexts.

figure 5

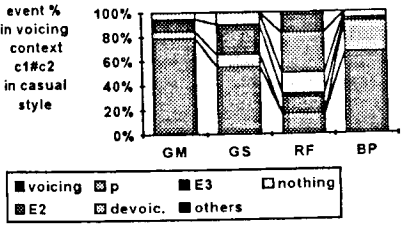
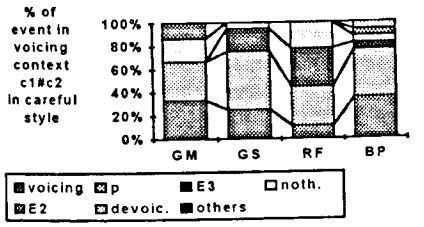


figure 6



The same variability between speakers and styles have been obtained for devoicing and palatalisation contexts.

Phonological rules

The results show in all cases that for a given the phonological behaviour is very different between the two styles studied. Moreover, in a given style, different strategies can be observed.

The role played by E caduc and pauses seem very important to distinguish the two styles. Moreover presence or absence of potential E caduc ('$') implies different strategies for a given speaker in a given style (see figure 3/ figure 4).

These results led us to test on the KTH synthesizer some phonological rules using pause and hesitation insertions and schwa deletions to characterize both a given speaker and a given style. (See figure 7).

Figure 7: examples of phonemic rules.
(1) if casual style
F$#D ---> **VS#D**; F#D ---> **FE2#D**;
(2) if careful style
F$#D ---> **F$#p#D**; F#D ---> **FE3#D**.

CONCLUSION

The study presented here have shown that to generate phonemic rules modelling strategies used by different speakers in different styles it seems to be necessary to take into account phonotactic constraints and specific phonological events.

In futur, perception tests on speech synthesis from the kth synthesizer will be developp to test intelligibility and naturalness involve by this type of rules.

REFERENCES

[1] **Péan V., Williams S., and Eskénazi M.**, 1993, "The design and Recording of ICY, a corpus for the study of Intraspeaker variability and the Characterization of Speaking stYles", Eurospeech, vol 1, pp 627-630, Berlin.

[2] **Prouts B**, 1980, "Contribution à la synthèse de la parole à partir du texte; transcription graphème-phonème en temps réel sur microprocesseur", thesis, France.

[3] **Lacheret-Dujour A., Péan V.**, 1994, "Towards a prosodic cues-based modelling of phonological variability for text-to-speech synthesis", ICSLP, pp 1763-1766, Yokohama.

[4] **Léon P. R.**, 1987, "E caduc : facteurs distributionnels et prosodiques dans deux types de discours", XIth ICPhS, pp 109-112, vol 3, Tallinn, Estonia, USSR.

[5] **Hansen A. B.**, 1991, "The covariation of [ ] with style in Parisian French: an empirical study of 'E caduc' and pre-pausal [ ]", Proceedings of the ETRW 'Phonetics and Phonology of Speaking Styles', pp 30_1-30_7, Barcelona.

[6] **Péan V.**, 1995, "Phonological rules modelling style variations of 'e caduc' in Parisian French spontanous speech for text-to-speech synthesis", Eurospeech, Madrid.

AN ACOUSTIC ANALYSIS OF HESITATION PARTICLES IN GERMAN

M. Pätzold and A. Simpson
IPDS, Kiel, FRG

ABSTRACT

A spectral analysis of the vocalic portions of hesitation particles produced by three speakers of North German was undertaken. For two of these speakers vocalic portions of similar quality in lexical items were also analysed and found to be significantly different from the vocalic portions of hesitation particles.

INTRODUCTION

During the course of an interactional exchange movements in a speaker's vocal tract are fulfilling linguistic, interactional and primary biological functions. These functions can be carried out in temporal overlap, e.g. German *ja* produced on a pulmonic ingressive airstream is doing both linguistic and interactional work; counting out loud can be carried on while inhaling, allowing talk to continue while a primary biological function is performed.

In the majority of cases it is possible to assign the phonetics being produced in talk to one of these functional categories or, if need be, tease the various components apart if two or more functions are being accomplished simultaneously.

For a small number of items in German, however, the assignment of the phonetics to one or the other category is not always transparent. Hesitation particles in German are an example of this. By hesitation particles we mean a syllable comprising a vowel (plus bilabial nasal) employed by some speakers at trouble spots in talk, often represented in conversational transcripts with *uh(m)* (English) or *äh(m)* (German). While it is clear that hesitation particles only do interactional work, they do consist of a vowel (plus nasal for some speakers) and the question arises as to whether the phonetics which make up these particles are correlates of the same phonological systems and structures which make up lexical items.

Levelt [1,2] suggests that both possibilities must be entertained: while the vowel in hesitation particles may represent the neutral position of the oral cavity for many languages, the [ɛ] quality found in hesitation particles in Swedish may be one consequence of acquiring a form of derived lexical status [1:74].

In this paper we would like to provide tentative acoustic evidence from German that shows hesitation particles to be phonetically different from lexical items, i.e. although they make use of the facilities provided by the vocal tract hesitation particles do not take part in the phonology of the language. However, we will also distance ourselves from Levelt's claim that the vowels in these items represent the neutral position of the oral cavity.

We examine the acoustic details of the vocalic portions in the hesitation particles of three speakers of German and for two of these speakers we compare the vocalic portions in hesitation particles with those found in a selection of lexical items. We show that the vocalic portions in hesitation particles are significantly different from those of lexical items having similar quality.

The phonetic description of hesitation particles has received relatively little attention and has been largely restricted to Fo and duration [3,4]. The treatment of qualitative aspects is rare and restricted to brief impressionistic description [5:246].

DATA AND METHOD

The material analysed here was collected at the IPDS Kiel as part of the Verbmobil project [6]. Speakers were required to arrange a number of appointments within a two-month period displayed on a sheet they had before them. Appointments could not be made on certain days in the two-month period. These days were indicated by shaded areas which were different for each speaker. The speakers communicated via headsets and had to press a button if they wanted to talk, at the same time blocking the channel for the other speaker. This set-up elicited spontaneous data which excludes turn overlap, back channel responses, etc. For more details on the technical set-up used, elicitation materials, etc. see [7].

The hesitation particles produced by three speakers (henceforth TIS, OLV, GEP) were subjected to an LPC spectral analysis and the first two formant frequencies at around the mid-point of the vocalic portion were measured.

For two of the speakers (TIS and OLV) values for the first two formants were also obtained at the midpoint of vocalic portions in lexical items considered to lie in the vicinity of the vocalic portions in hesitation particles. The lexical vowels chosen for comparison are ɛ (e.g. *fest*), ɑ (e.g. *Tage*), ə (e.g. *bitte*) and ɐ (e.g. *wieder*). We will use ɜ to represent the vowel found in hesitation particles under investigation, although, as with the symbols being used for the other vowels, ɜ is not meant to directly represent the phonetic qualities found.

One of the problems of making a comparison of the vocalic portions in hesitation particles with those in lexical items are the considerable durational differences. The vocalic portions of hesitation particles are very long, having mean durations more than twice that of long open vowels in lexical items. So, whereas any contextual effects produced by neighbouring are likely to be negligible the same can not be said for vocalic portions in lexical items. In an attempt to minimize these contextual effects only vocalic portions in lexical items with a duration greater than 80ms were analysed.

RESULTS

Both GEP and OLV produced hesitation particles comprising only vocalic portions. These had central quality both in the height and front-back dimensions. The hesitation particles produced by TIS consist of a half-open central vocalic portion followed by a bilabial nasal. The vocalic portions in all cases were monophthongal.

Table 1: Mean and standard deviations of F1 and F2 of vocalic portions in hesitation particles for the three subjects. Measurements made at around the mid-point of the vocalic portion.

	F1		F2		
	$\bar{x}$	*s*	$\bar{x}$	*s*	n
OLV	520	38	1556	93	38
TIS	569	35	1273	46	40
GEP	377	28	1475	95	30

As we can see from Table 1 the impressionistic differences are supported by the inter-speaker differences in the formant values. TIS has a higher F1 and lower F2 indicating a vocalic portion more open in quality than those found for OLV and GEP.

Table 2: Mean and standard deviations of F1 and F2 of the vocalic portions in hesitation particles and a selection of vowels of similar quality. Measurements made at around the mid-point of the vocalic portion.

a) TIS

	F1		F2		
	$\bar{x}$	*s*	$\bar{x}$	*s*	n
ɛ	501n.s.	92	1797***	169	9
ɐ	556n.s.	77	1371***	120	21
ɔ	438***	62	1387**	183	18
ɑ	695***	46	1289n.s.	108	67
ɜ	569	35	1273	46	40

b) OLV

	F1		F2		
	$\bar{x}$	*s*	$\bar{x}$	*s*	n
ɛ	561***	44	1680***	155	26
ɐ	616***	81	1353***	133	34
ə	461***	53	1569n.s.	230	16
ɑ	659***	51	1236***	87	70
ɜ	520	38	1556	93	38

The comparison of the vocalic portions in hesitation particles with those in a selection of lexical items is shown in Table 2 for two speakers. For TIS the quality of the vocalic portions in hesitation particles is closest to that found for ɐ, for OLV it is the phonetic realisation of ə which is closest. Although in all cases either F1, F2 or both formant values are significantly different from those found in the vocalic portions of lexical items.

DISCUSSION

We set out to show that the phonetic quality of the vocalic portions in hesitation particles is significantly different from that of vocalic portions in lexical items. Our results show that for two speakers of German this would seem to be the case. The vocalic portions of hesitation particles have their own quality suggesting that they are phonetic correlates of a phonological system which is different from those employed in lexical items. This result is hardly surprising when one considers that these particles serve to indicate linguistic trouble, a function which can be successfully fulfilled by being different in form from surrounding linguistic material.

Our results in part lend support to Levelt's claim that hesitation particles are different from words. But Levelt's claim is stronger than this. He suggests that the vowel of a hesitation particle is a neutral sound which varies phonetically with "the neutral position of the oral cavity from different languages" [1:74], a vowel which he slightly later refers to as schwa. This claim is far harder to substantiate since it is not clear how one would go about ascertaining the "neutral position of the oral cavity" for a language, or even an individual speaker. Levelt is presumably referring to a cavity position which is dependent upon the articulatory setting [8] of a language and not to an independently motivated articulatory (e.g. [9:137]) or acoustic construct [10:49f]. Levelt's later use of schwa is equally problematic as it can only be referring to the auditory product of the neutral position and not to the phonetic vowel category [ə] or the phonetic correlate of a phonological item such as /ə/ often proposed for languages such as English and German.

On the basis of our data, we would like to make a claim which can be tested and refuted. The vowel quality found in hesitation particles is different from vowel qualities found in lexical items. A consistent difference is maintained, although the exact nature of this difference varies from speaker to speaker. The three speakers we investigated all produced vowels which were central, but with considerable inter-individual variation in height.

Indeed, it would be possible, to test our claim on a language, such as Swedish in which the non-central [ɛ] quality of the vowels in hesitation particles led Levelt to claim that such items were taking part in the phonology of the lexis. It would be interesting to see if the vowel quality found is different from that found in lexical items such as *lära, läkare*.

Problems of Comparability

One of the biggest problems we encountered in this study was the degree to which the items we are comparing are indeed comparable. Hesitation particles are in general prominent, brought about by factors such as length and loudness. We would therefore have been justified in comparing hesitation particles with vowels from prominent syllables in lexical items, i.e. those which are stressed. Although this comparison would have been simple, we wanted to go one step further and demonstrate that the vowels of hesitation particles are even different from those of central vowels in the language. This causes a serious problem since the central vowels in German (ɐ and ə) are always unstressed and the quality varies greatly, not least because of the often short duration in various consonantal and vocalic contexts. We therefore imposed a minimum duration of 80ms in an attempt to minimize these effects, while still being able to get a sufficient number of tokens. However, the setting of a lower duration is also not without problems as it almost exclusively returns central vowels in open syllables, which in the case of ə were also utterance final. One way of overcoming this problem may be to record the same speakers producing spoken prose. This would allow the structures and frequency of occurrence to be controlled.

Other aspects of hesitation particles

In the course of analysing the vowels of hesitation particles, we made a number of other observations which suggest that hesitation particles make use of a different set of phonetics from lexical items. So, for instance, the hesitation particles produced by TIS consisted of a vowel plus nasal sequence, however, the vowel was rarely nasalized and the soft palate was often lowered shortly after bilabial closure was made for the nasal, leading to nasal plosion. What is of interest is that this complex was different from similar complexes in lexical items, i.e. either vowel-stop-nasal sequences (e.g. *eben*) or vowel-nasal sequences (e.g. *gemeinsam*).

We suspect that as the amount of phonetic material gathered on hesitation particles in German and other languages grows, so too will the catalogue of differences between them and linguistic items.

ACKNOWLEDGEMENT

The data collection was partially funded by the German Federal Ministry for Education, Science, Research and Technology (BMBF) under grant 01IV101M7. The responsibility for the contents of this study lies with the authors.

REFERENCES

[1] Levelt, W.J.M. (1983), "Monitoring and self-repair in speech", *Cognition*, vol. 14, pp. 41-104.

[2] Levelt, W.J.M. (1989), *Speaking: from intention to articulation*, Cambridge: MIT Press.

[3] O'Shaughnessy, D. (1992), "Recognition of hesitations in spontaneous speech", *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 1, pp. 593-596.

[4] O'Shaughnessy, D. (1992), "Locating disfluencies in spontaneous speech: an acoustical analysis", in *Proc. European Conf. on Acoustics on Speech Communication and Technology*, vol. 3, pp. 2187-2190.

[5] Wittkop, E.-M. (1988), *Gliederungspartikeln im Dialog*, Munich: iudicium.

[6] Karger, R., Wahlster, W. (1994), *VERBMOBIL Handbuch*, Verbmobil Technisches Dokument 17, Saarbrücken: DFKI.

[7] Pätzold, M., Scheffers, M., Simpson, A., Thon, W. (1995), "Controlled elicitation and processing of spontaneous speech in Verbmobil", *Proc. XIIIth ICPhS.*

[8] Honikman, B. (1964), "Articulatory settings", in D. Abercrombie, D.B. Fry, P.A.D. MacCarthy, N.C. Scott, J.L.M. Trim (eds.) *In Honour of Daniel Jones*, London: Longman Green, pp. 73-84.

[9] Laver, J. (1994), *Principles of phonetics*, Cambridge: CUP.

[10] Fant, G. (1960), *Acoustic theory of speech production*. The Hague: Mouton.

TEMPORAL-BASED SPEAKER SEX DIFFERENCES IN READ SPEECH: A SOCIOPHONETIC APPROACH

Sandra P. Whiteside
Speech Science, University of Sheffield, S10 2TA, England, UK.

ABSTRACT

This paper summarises the findings of a study investigating temporal-based speaker sex differences. Measures taken include sentence durations, syllable rates, consonant elisions, vowel reductions, VOTs of plosives and durations of fricatives. The findings are discussed within a sociophonetic framework.

INTRODUCTION

Speaker sex differences have been shown to exist in the acoustic signals of read speech. Some studies have looked at fundamental frequency differences and formant frequency differences [1, 2, 3]. Other studies have investigated differences in the glottal source [4] and first formant bandwidths and amplitudes [5]. However, there is also some evidence to suggest that differences also exist in the temporal domain [6]. For example in the TIMIT database women tend to speak more slowly than men, men tend to reduce their vowels to [ə] more often than women and women tend to release sentence-final plosives more often than men. These findings were statistically significant [6].

This paper presents the results of a preliminary acoustic-phonetic investigation into speaker sex differences in the temporal domain. It focuses on the read speech data of three men and three women speakers with a British General Northern accent.

METHOD

Subjects

Three male and three female adult native speakers of English served as speakers. All speakers came from North of England and represented a British General Northern accent which can be defined as a non-rhotic accent of Standard English characterised in the vowel system by COULD/CUD and GAS/GLASS rhyming and a tendency to retain strong vowels where RP shows weakening e.g. computer /kɒmpjuːtə/ [7].

Speech material

Ten repetitions of five sentences were read by three men and three women speakers. This made a total of three hundred sentences (150 for the men and 150 for the women). Both the sentences and the speakers formed part of the APLAWD [8] speech corpus of British General Northern (GN) accent speakers. The sentences are as follows:

Sentence One.: *George made the girl measure a good blue vase.*
Sentence Two. : *Why are you early you owl ?*
Sentence Three .: *Cathy hears a voice amongst Spar's data.*
Sentence Four. : *Six plus three equals nine.*
Sentence Five. : *Be sure to fetch a file and send theirs off to Hove.*

Recording procedures

High quality recordings were made in a sound proof studio at the University of Leeds, using recording procedures described in [8]. High quality audio cassette copies were used to digitise the speech samples onto a Macintosh LCII computer using a Farallon™ Macrecorder and Signalyze™. A sampling rate of 11kHz was used.

Analysis

Using Signalyze™ [9] the durations of each of all the three hundred sentences were measured. It must be noted that these durations did not include any of the pauses observed in the sentences as the observed pauses were subtracted from overall sentence duration. However, the incidence of pauses was noted and some of these observations are discussed below. A summary of the sentence duration results can be found in table 1. Syllable rates (syllables/second) for each of the sentences were also calculated. A summary of these results can be found in table 2. In addition each group of sentences (60 data items for each sentence) was examined for specific linguistic and acoustic phonetic phenomena as outlined below. Observations were tested for statistical significance using a statistical package (Statview ™). The results of these tests are summarised in tables 3 to 7 where an asterisk (*) indicates statistical significance.

Sentence 1 (/ ˈʤɔːʤ ˈ meɪd ðə ˈ gɜːl ˌmeʒə(ɹ)ə ˈgʊd ˈbluː ˈvɑːz/)

i) The occurrence of schwa elisions in 'measure a' was examined. Whether the speakers realized the utterance as [ˈmeʒə ə], [ˈmeʒɹə] or [ˈmeʒəɹə] was noted using auditory and acoustic analysis; ii) In addition the occurrence of pauses after 'girl' was investigated. It was predicted that the occurrence of pauses would coincide with a lengthening in the duration of 'girl'; iii) VOT values were taken for /g/ in 'girl' and 'good'; iv) the duration of [ʒ] in 'measure' was measured and whether it was voiced or devoiced was noted and v) dB differences between the amplitude peaks of the vowels /ɔ/ in 'George' and /ɑ/ in 'vase' were noted. See table 3 for statistical analyses.

Sentence 2 (/ˈwaɪ ˈɑ/ə juː ˈɜːlɪ juː ˈaʊl/)

i) Whether 'are' was fully represented as the vowel [ɑ], reduced as the schwa [ə] or elided altogether was noted. Auditory and acoustic analysis was used to make these decisions. The presence of a schwa was noted if there was a separate intensity peak in the speech pressure waveform; ii) The incidence of pauses after 'early' was noted together with iii) dB differences between the peak amplitude of /aɪ/ in 'Why' and the peak amplitude of /aʊ/ in 'owl'. See table 4 for statistical analyses.

Sentence 3 (/ˈkæθɪ ˈhɪəz ə ˈvɔɪs əˈmʊŋst ˈspɑːz ˈdeɪtə/)

i) Pausing after 'voice' was noted; ii) In addition the duration of 'voice' was measured for each speaker; iii) The duration of the /sts/ cluster in 'amongst Spar's' was measured; d) Whether speakers realized this cluster as a reduced form ([s s], [s]) or as a full representation([sts]) was also noted. The criteria used for these judgements included auditory and acoustic analysis. Speech pressure waveforms were used for the acoustic criteria where: [sts] was realized as fricative followed by a closure phase and a subsequent pulse/transient which was followed by a fricative; [s s] was realized as two fricatives separated by a reduction in amplitude in the speech pressure waveform and [s] was realized as a single fricative; iv) VOTs were measured for /k/ in 'Cathy' and /d/ in 'data'; v) the duration of /z/ in 'hears' was measured, it was also noted whether this segment was devoiced or not; vi) the duration of /s/ in 'voice' was measured and vii) dB differences between /æ/ in 'Cathy' and /eɪ/ and /ə/ were also measured. See table 5 for statistical analyses.

Sentence 4 (/ˈsɪks plʊs ˈθriː iːkwəlz ˈnaɪn/)

i) The vowel pair in /ˈθriː ˈiːkwəlz/ was examined to see whether speakers had fully realized the vowels acoustically or reduced them to a single vowel. Auditory analyses were also used in this procedure; ii) the duration of /ˈθriː ˈiːz/ was also measured; iii) the occurrence of a pause after 'three' was examined for each of the 60 sentences; iv) duration measurements were made for: initial and final /s/ in 'six', the word 'nine', /s/ in 'plus' and /z/ in 'equals'; v) whether /z/ in 'equals' was devoiced was noted; vi) dB differences between the vowel peaks in 'six' and 'nine' were measured and vii) whether speakers showed pre-plosive glottalization in 'six' was also noted. Statistical analyses for this sentence are given in table 6.

Sentence 5 (/bɪ ˈʃɔː tə ˈfeʧ ə ˈfaɪl ən(d) ˈsend ˈðeəz ɒf tə ˈhəʊv/)

i) Whether speakers paused after 'file' was noted; ii) the duration of the word 'file' was measured for each of the 60 sentences using the speech pressure waveform and auditory analysis; iii) duration measurements were taken for /ʃ/ in 'sure', /ʧ/ in 'fetch', /s/ in 'send' and /z/ in 'theirs'; iv) it was also noted whether /z/ was devoiced or not; v) dB differences between the vowel peaks of /ɔ/ in 'sure' and /əʊ/ in 'Hove' were measured and vi) the occurrence of glottalization in 'fetch' was noted. See table 7 for statistical analyses.

RESULTS AND DISCUSSION

Table 1 shows that sentence durations are longer for the women versus the men, with the women showing larger standard deviations. These results agree with those of Byrd [6]. The women also show lower syllable rates than the men as shown in table 2.

Table 1. Mean sentence durations and standard deviations of sentences 1 to 5 by speaker sex

	Men		Women	
Sent.	Mean (ms)	s.d. (ms)	Mean (ms)	s.d. (ms)
1	2479.1	113.9	3045.2	483.2
2	1420.6	94.1	1872.2	259.1
3	2325.9	131.2	2978.4	493.0
4	1706.6	177.2	2124.0	236.7
5	2705.1	224.0	3442.0	541.3

Table 2. Mean syllable rates (syllables per second) and standard deviations (s.d.) of sentences 1 to 5 by speaker sex

	Men		Women	
Sent.	Mean	s.d.	Mean	s.d.
1	3.853	0.264	3.234	0.393
2	4.735	0.434	3.806	0.51
3	4.313	0.242	3.44	0.592
4	3.144	0.14	2.844	0.341
5	4.466	0.378	3.584	0.644

Table 3. Statistical analyses for Sentence 1

Single factor ANOVA & X2 test results
speaker sex (SS) & sentence durations F=38.997, p =0.0001*
schwa elisions & sentence duration F=28.17, p=0.0001*
SS & schwa elisions F=0.62, p=0.4343
pausing after 'girl' & SS χ^2 =13.469, p=0.0002*
pausing after 'girl' & longer durations of 'girl F=136.017, p=0.0001*
VOT /g/'girl' & SS F=0.0004, p=0.957
VOT /g/ 'good' & SS F=8.341, p=0.0055*
Duration (ms) [ʒ] 'measure' & SS F=65.354, p=0.0001*
Duration [ʒ] 'measure' & voiced/ devoiced F=29.54, p= 0.0001*
Voiced/devoiced [ʒ] 'measure' & SS χ^2=23.72, p=0.0001*
dB difference /ɔ/ in 'George' & /ɑ/ in 'vase' & SS F=4.867, p=0.0313*

Table 4. Statistical analyses for Sentence 2

Single factor ANOVA results
speaker sex (SS) & sentence durations F=80.529, p= 0.0001*
SS & schwa elisions F=16.789, p = 0.0001*
SS & full representation of 'are' F=12.069, p = 0.001*
vowel reductions & sentence duration F=39.576, p = 0.0001*
schwa elisions & sentence duration F=8.511, p = 0.005*
pauses after 'early' & SS F=29.696, p = 0.0001*
dB differences 'Why' & 'owl' F=1.361, p=0.2481

For Sentences 1 to 4 (tables 3 to 6), this small group of speakers shows a link between speaker sex and pausing. In addition, there is also a link between the occurrence of pauses and longer duration values when words precede a pause. The findings here differ from those of Byrd [6] who found that there was no link between speaker sex and the occurrence of pauses. However what is interesting to note is that the findings here mirror some of the evidence of previous research which has shown that men tend to pause less frequently than women during a conversational speech setting. By not pausing men tend to dominate a conversation as this reduces both turn taking and any interruptions. Conversely women tend to pause more thus allowing themselves to be interrupted more frequently [10 &11].

The differences between Byrd's [6] findings and those here could be due to cultural differences between British English and American English speakers. That cultural differences exist in conversational style has been reported elsewhere [11]. However it is also possible that these differences are purely a result of individual speaker variation in the speakers investigated this study.

Table 5. Statistical analyses for Sentence 3

Single factor ANOVA & X2 test results
speaker sex (SS) & sentence durations F=49.087, p = 0.0001*
SS & pausing after 'voice' F=11.505, p= 0.0013*
SS & duration of 'voice' F=14.132, p=0.0004*
SS & /sts/ cluster durations F=25.67, p=0.0001*
SS & cluster reductions F=34.208, p = 0.0001*
/sts/ reductions & /sts/ durations F=47.823, p = 0.0001*
VOT /k/ 'Cathy & SS F=3.451, p=0.0683
Voiced/devoiced /z/ 'hears' & SS χ2= 30.0, p=0.0001*
Duration (ms) /z/ 'hears' & SS F=10.529, p=0.002*
Duration (ms) /s/ 'voice' & SS F=22.348, p=0.0001*
VOT /d/ 'data' & SS F=3.924, p=0.0529
Voiced/devoiced /z/ 'hears' & duration (ms) /z/ F=32.492, p=0.0001*
dB differences between /æ/ in 'Cathy' & /eɪ/ in 'data' & SS F=14.007, p=0.0004*
dB differences between /æ/ in 'Cathy' & /ɔ/ in 'data' F=38.545, p=0.0001*

Tables 6 and 7 showed no speaker sex differences in glottalization and few differences in VOTs. Tables 3 to 7 however, do show that the men in this study tended to either elide or reduce both vowels and consonants which contributed to shorter sentence durations. Conversely the women showed a tendency to realise speech segments more fully. This therefore meant

Table 6 Statistical analyses for Sentence 4

Single factor ANOVA & X2 test results
speaker sex (SS) & sentence durations F=59.753, p = 0.0001*
SS & vowel pair reductions F=39.479, p = 0.0001*
pauses after 'three' & duration of 'three' F=34.952, p = 0.0001*
SS & pauses after 'three' F=13.576, p = 0.0005*
Duration (ms) initial /s/ in 'six' & SS F=9.371, p=0.0033*
Duration (ms) final /s/ in 'six' & SS F=82.119, p=0.0001*
Duration (ms) of 'nine' & SS F=17.413, p=0.0001*
Duration (ms) /z/ in 'equals' & SS F=116.391, p=0.0001*
Voiced/ devoiced /z/ in 'equals' χ^2=16.463, P=0.0001*
Voiced/ devoiced /z/ in 'equals' & its duration value F=23.572, p=0.0001*
Duration (ms) /s/ in 'plus & SS F=14.288, p=0.0004*
dB differences between 'six' & 'nine' & SS F=1.599E-17, p=1
Glottalization in 'six' & SS All speakers showed glottalization effects

Table 7 Statistical analyses for Sentence 5

Single factor ANOVA & X2 test results
speaker sex (SS) & sentence duration F=34.952, p = 0.0001*
speaker sex & pauses after 'file' F=2.61, p = 0.1116
pauses after 'file' & duration of 'file' F=16.859, p = 0.0001*
Duration (ms) /ʃ/ in 'sure' & SS F=40.972, p=0.0001*
Duration (ms) /ʧ/ in 'fetch' & SS F=15.064, p=0 0003*
Duration (ms) /s/ 'send' & SS F=51.992, p=0.0001*
Duration /z/ 'theirs' & SS F=36.727, p=0.0001*
Voiced/ devoiced /z/ 'theirs' & SS χ^2=23.254, p=0.0001*
Duration /z/ & voiced/devoiced F=9.998, p=0 0025*
dB differences between /ɔ/ in 'sure' & /əʊ/ in 'Hove' & SS F=27.608, p=0.0001*
Glottalization in 'fetch' & SS All speakers showed glottalization effects

that the women's speech segments were on average longer than those of the men. These results agree with the finding that women enunciate more clearly than men [10]. We can argue that enunciating more clearly requires greater articulatory effort. From this we can suggest that the significant differences (tables 3 to 7) in the dB ratios for sentence-initial and sentence-final syllable-nuclei for the men and women speakers reflect greater articulatory effort by the women speakers, who had lower dB ratios. The findings may also reflect the different strategies men and women adopt in a conversational setting. However this is pure conjecture at this stage and reflects the need for further research.

The results of this preliminary investigation provide some acoustic-phonetic evidence that the men and the women in this data sample realise sentences differently when they are read in a controlled laboratory situation. Further research is planned using another British English database.

REFERENCES

[1] Peterson, G. E. and Barney, H. L. (1952). Control methods use in the study of vowels, *Journal of the Acoustical Society of America, 24,* 175-184.

[2] Wu, K. and Childers, D. G. (1991). Gender recognition from speech. Part I: Coarse analysis, *Journal of the Acoustical Society of America, 90(4),* 1828-1840.

[3] Childers, D. G. and Wu, K. (1991). Gender recognition from speech. Part II: Fine analysis, *Journal of the Acoustical Society of America, 90(4),* 1841-1856.

[4] Fant, G. (1979b). Temporal fine structure of formant damping and excitation, *Speech Communication Papers, Acoustical Society of America,* 161-165.

[5] Karlsson, I. (1988). Glottal Waveforms for normal female speakers, *Speech Transmission Laboratory-Quarterly Progress and Status Report, Royal Technical Institute, Stockholm,* 31-36.

[6] Byrd, D. (1992). Preliminary results on speaker-dependant variation in the TIMIT database, *Journal of the Acoustical Society of America, 92(1),* 593-596.

[7] Wells, J. C. (1982). *Accents of English 2.* Cambridge University Press.

[8] Lindsey, G., Breen, A & Nevard, S. (1987). Spar's Archivable Actual-Word Databases, *UCL Report on SPAR Project.*

[9] Keller, E. (1992). *Signal Analysis for Speech and Sound.* InfoSignal™ Inc.

[10] Elyan, 0. (1978). Sex differences in speech style, *Women Speaking, 4,* 4-8.

[11] Tannen, D. (1992). *You just don't understand,* Virago Press.

THE PHONETIC OF IBN DURAYD

S. I. Sara, S.J. and A. O. Zawawi
Georgetown University

ABSTRACT

Arabic sources since the eighth century have provided treatises on the phonetics of Arabic. Ibn Durayd, in the process of composing his lexicon of Arabic pre-pended the lexicon with a treatise on the phonetics of Arabic . In it he classified the sounds of Arabic according to three types of articulatory criteria that yield multiple sets of features, and which in turn distinguish each letter, and group different letters into distinct subgroups.

IBN DURAYD

Ibn Durayd (223-321H/ 838-933A.D.)[1] is an Arab essayist, poet, lexicographer and linguist. He was born in Basrah, Iraq (223/838), grew up in Oman and died in Baghdad in (321/933). Among his many teachers are listed nineteen prominent savants of his time, and among his students are listed forty five influential thinkers who shaped the development of Arabic studies. His biography is related in thirty seven biographical and historical records. Of his works eight have been published, and nineteen others have been mentioned in the sources, though not yet published.

INFLUENCES

Even though Ibn Duryad was an independent innovator and thinker, in composing his own lexicon he refers specifically to the book of Al-Khalīl(101-175/719-791)[2] who had provided the first model for the study of the science of lexicography and the science of Arabic phonetics. Ibn Duryad would re-arrange the lexical entries of the Arabic lexicon according to a new organizational principle . He grouped together all the lexical items that shared the same number of radicals,i.e. all the bi-radicals together, all the tri-radicals together, etc. He made other innovations into which we can enter here. In addition, he pre-pended to this massive undertaking an introduction that included a treatise on phonetics explaining the sounds of Arabic, just as Al-Khalīl had done with his Kitāb Al-ʕayn, the first Arabic Dictionary.

PHONETICS

The treatise on phonetics of Ibn Durayd keeps alive the tradition of explaining to the user of the dictionary the basic elements i.e. the letters, of the lexical items, the arrangement of the lexical items, and the manner in which the letters are produced. What is of interest in this treatise on phonetics, is not only that it maintains the tradition of Al-Khalīl, but it has new groupings and new terminology that is not found in Al-Khalīl. However, the arrangements that the author discussed, had only limited lexicographical function, since he ignored the phonetic order of the letters and reverted to the traditional order of the letters in his dictionary. One can only conclude that this was a mere courtesy by the new author to the first lexicographer by keeping the tradition alive. The author stated that he was aware of the work of other linguists, but he was explaining the phonetics of the language in his own way for the benefit of the user of the dictionary.

MUŞMATAH & MUðLAQAH

The first task of the user is to know the letters of the dictionary, since they are the poles around which the words are constructed. Hence the reader must know their exits (maxārig), their progressive stages (madārig), their remoteness from each other (tabʃāud), their closeness (taqārub) to each other and what may or may not co-occur (taʔāluf) with each other, and the reasons for such allowance or disallowance. According to Ibn Durayd, the letters of Arabic are of seven types that are grouped under two major headings: Twenty two letters are /muşmatah/ 'silent', three of which are weak, and nineteen are strong; the other six letters are /muðlaqah/ 'edge letters'. They are schematically arranged in the following Chart I:

Class		type	letter
/muşmatah/ 'silent'	1.	throat	ʔ, h, ħ, ʕ, ɣ, x - أ,هـ,ح,ع,غ,خ
	2.	lowest part of the	q, k, g, ʃ - ق,ك,ج,ش
	3.	tongue	s, z, ş - س,ز,ص
	4.	middle of the tongue	t, ţ, d - ت,ط,د
	5.	nearest in the mouth nearest upper concavity	đ, θ, ð, ḏ - ظ,ث,ذ,ض
/muðlaqah/ 'edge'	6.	labial	f, b, m - ف,ب,م
	7.	tip of tongue	r, n, l - ر,ن,ل

Chart I: First Binary classification of the letters of Arabic

Even though the above chart is binarily conceived, it parallels the classification of Al-Khalīl in dividing the vocal tract into eight subdivisions except that 2 & 3 subdivisions are a conflation of three locales in Al-Khalīl, Sara[3]. While Al-Khalīl emphasized the divisions of the upper perimeter of the vocal tract, Ibn Duryad's emphasis was more on the active articulator, the tongue, and the lower perimeter of the vocal tract.

EXITS

For Arab linguists, the term /maxrag/'exit' is a description of the narrowing of the vocal tract. It corresponds, in a broad sense, to the 'point of articulation' in our current use. Each segment or group of segments were characterized by their appropriate 'exit'. The following is Ibn Durayd's classification of the letters of Arabic according to sixteen exits as in chart II:

Cavity		Exit	letter
THROAT	1.	Lower part	h, ʔ, A- ه,ء,أ
	2.	Middle part	ʕ, ħ - ع,ح
	3.	Upper part	ɣ, x- غ,خ
M	4.	Farthest	q, k- ق,ك
	5.	Uvula	g, ʃ - ج,ش
O	6.	middle of tongue	y - ي
	7.	side of tongue/upper incisors	s, z, ṣ,- س, ص, ز
U	8.	right edge of tongue	n - ن
	9.	Right edge close to /n/	l - ل
T	10.	Close to /n/ but inner	r - ر
	11.	Edge of tongue, base of incisors	t, d, ṭ - ط,د,ت
H	12.	Inner lower lip	f - ف
	13.	between the lips	w, b, m - م,ب,و
	14.	light /n/	n - ن
	15.	Edge of tongue/edge of incisors	ð̣,θ, ð, - ظ,ث,ذ
	16.	Middle of the tongue/ right edge	ð̣ - ض

Chart II: Classification of the letters according to exit

The above Chart with its sixteen exits reflects the organization of Sībawayh in his treatment of Arabic sounds [4]. Sībawayh had arranged his sounds according to 16 exists. There are to be sure points of difference, but the general organization is similar. To be noticed again, Ibn Durayd's emphasis on the tongue and the lower perimeter of the oral cavity, while Sībawayh gave equal recognition to the palate and the upper perimeter of the cavity.

FEATURES

There is yet another classification that the Ibn Durayd provides when discussing the letters of Arabic. He noticed that though the letters may have different exits, they still may have features in common. 'Soft' letters may be found in the throat region or the mouth region. Consequently he regrouped the letters according to these common features as in chart III.

Feature	Letters
Mahmūs 'muted'	h, ħ, k, x, s, ʃ, θ, ṣ, t, f
Maghūr 'loud'	ʔ, A, ʕ, ɣ, q, g, y, ð̣, l, n, r, z, d, ð, ṭ, ð̣, b, w, m
Rixwah 'soft'	h, ħ, k, x, s, ʃ, ʕ, ɣ, ṣ,ð̣, ð, ð̣, θ, f, z
Madd& Layn 'length'	w, y, A
Muṭbaqah 'covered'	ṣ, ṭ, ð̣, ð̣
Shadīdah 'tight'	ṭ, ʃ, g, etc [ʔ, q, k, l, r, n, d, b, m].

Chart III. Classification of the letters according to features

Several comments are in order when one reflects on the above chart of features. The features are identical with those found in Sībawayh, even though not all the features of Sībawayh are accounted for. There are also variations in the selection of letters that share the same feature. First the sequencing of letters is not identical in the two sets, even though the letters are the same. For the feature Shadīda 'tight' he did not list all the letters, but only a sample of three. The balance of letters that share this feature is supplied here from Sībawayh and included between[]. There are, however, two significant deviation in the above organization: first /k/ is included with the Rixwah 'soft'. This is completely contrary to the features of this letter. As one notices that all the letters under this feature are of continuant type, and they have been so classified by the other linguists of the time. In a similar manner, he grouped /ʃ/ with the Shadīdah'tight'. The subgroup of letters that are listed as Shadīdah are all of closure type. By putting /ʃ/ with Shadīdah it effectively puts it with both Rixwah and Shadīda, i.e. 'continuant' and 'interrupted' types. Since he did not list all the letters that share this feature, and what is supplied above is from Sībawayh, whom he seems to follow so closely, this classification is again out of character. It is not easily explainable why /k/ and /ʃ/ are grouped under these features, i.e. out of their natural classes. We have no reasonable explanation at this time except to say that this may have been an error due to the nature of the composition of the dictionary, which was dictated by the author. Needless to say Ibn Durayd did not employ all the features that were available to him, and were readily available in the literature. One can only conclude that he was doing this by way of example, and that he was not necessarily giving an exhaustive listing of all the phonetic/ phonological lore of his time. Even with this brief outline of the sound system of Arabic, he was able to point reasons why certain letters do not co-occur within the same word, and why some regional dialects substitute one letter or sound for another in their speech due to the proximity and the articulatory congruity of the confused or substituted letters. There is a great deal in this treatise that is of historical and linguistic value.

BIBLIOGRAPHIY

[1] Ibn Durayd. 1987 Kitāb Jamharat al-Lugha. Ed. R.M. Baʕlabakki. Beirut:Dār al-ʕilm lilmalāʔīn. vol 1. Pp 39-47..

[2] Al-Khalīl Al-Farāhīdi. Ed. Makhzumi. Kitab Al-ʕAyn. Baghdad: Manshurāt Wazārat al-Thaqāfa wa al-ʕilm. Vol 1. Pp. 47-60.

[3]Sara, Solomon. 1993."Al-Khalīl the First Arab Phonologist".International Journal of Islamic and Arabic Studies: vol viii. Pp.1-59.

[4] Sībawayh.1898. Al-Kitāb. Egypt: Bulāq. Press. Vol. 2. Pp. 404-431.

IBN JINNI'S CONTRIBUTION TO PHONETICS

Muhammad Hasan Bakalla
King Saud University, Riyadh, Saudi Arabia

ABSTRACT

The tenth century A.D. witnessed a tremendous surge and expansion in avenues of Islamic scholarship both quantitatively and qualitatively. Arabic linguistic and phonetic scholarship is no exception. Towards the middle of the century, Ibn Jinni (d.1001), a Greco-Roman Arab pioneer in linguistics and phonetics advanced various ideas in these fields including their descriptive-interpretive techniques, methodology, technical terminology, definitions, as well as statements of universal validity. Most of these and other notions are scattered in some fifty works of his, especially ***Sirr Ṣinā'at al-Ʃrāb*** or ***SS*** [1] and ***Khaṣā'iṣ*** [2]. Only some general phonetic issues are being discussed below.

DEFINITION OF LANGUAGE

Being aware of the numerous languages of the Islamic Empire of his time, Ibn Jinni (IJ) defines the human language as follows : "Language is a set of sounds which are used by each community to express their ideas and intentions". Although modern linguists usually emphasize the arbitrary nature of the sound system of all languages, IJ appears to be aware of this fact and the social role of language too [2].

ORIGIN OF LANGUAGE

IJ presents the three theories about the origin of language, which are current amongst his contemporaries. (1) Language is of divine origin; (2) Language is a convention and social agreement between two or more people in one and the same community; and (3) Language origin is based on onomatopoeic consideration where the adherents to this theory advocate the idea that languages started as a result of imitating the sounds in nature such as the roaring of the wind, the crashing of thunder, the murmuring of water, the braying of donkeys, the cawing of crows, the neighing of horses, etc. [2]. He seems to subscribe to each of the three theories, though he acknowledges that most learned men adhere to the second one.

SOUND SYMBOLISM

IJ is considered to be the founder and proponent of a theory known in Arabic morphology as the ***major etymology***. He observes that certain words clearly show a kind of natural association between sounds and meaning. According to him no matter in what order their radicals may occur, certain consonantal roots are connected with a main concept or common meaning. However, he admits that only a small portion of the vocabulary displays this symbolism [3].

DEFINITION OF VOCAL SOUND

IJ defines the vocal sound as "a perceptible though fleeting event which accompanies the pulmonic air-stream and lasts as long as it continues. Whenever the air passage is obstructed in the throat, mouth, or lips, the sound resulting from the obstruction by the articulators is called ***ḥurūf***. The sounds or timbres of ***ḥurūf*** differ from each other according to their places of articulation" [1]. Here he alludes to the two classes of speech-sounds, namely the vowels (V's) whose production is characterized by having free air passage; and the consonants (C's) which have varying degrees of obstruction as treated separately by him later. Also, reference is made here to the various points of articulation along the vocal tract, which will be dealt with below. Underlying his distinction between ***ʔaṣwāt*** (sounds) and ***ḥurūf*** (phonemes) is what in modern phonetic terminology is referred to as allophones and phonemes, respectively. This explanation can be substantiated by a newly-coined term he introduces for the first time, viz. ***ʿilm ʔaṣwāt wa ḥurūf***, meaning. the science of phonetics and phonology [1].

SPEECH AND WRITING

IJ clearly points out the precedence of speech over writing. According to him, speech had interceded writing in existence and, thus, the latter is considered secondary and ancillary to speech. He also makes a clear distinction between speech-sounds and letters of the alphabet. In his presentation of the sound system of Arabic, he discusses some sound variants which have no written symbols at all [1].

THE SPEECH APPARATUS

Prior to and during the tenth century there exists an abundant literature on the anatomy and physiology of the human body in general, and the organs of speech in particular [4]. The influence of such works on Arabic phonetics cannot be denied, as the latter draws a host of physiological and perceptual terms from the former. Here I like to refer to the earliest ever recorded diagram of the vocal tract which appears in a twelfth century linguistic work by Sakkaki [5].

SOUND AND MUSIC

In his definition of the vocal sounds, IJ appears to recognize the fact that the human vocal apparatus can produce innumerable varying sounds.This idea can also be supported by the comparison he draws between the human speech organs and musical instruments such as the ***nay*** and the ***lute*** for their similarity in making varying sounds (or notes). He maintains that the vocal tract which can produce various speech sounds in a language as a result of being articulated at different places, may be likened to the ***nay***. When blown unchecked, it produces a simple long sound just like the simple long vowel *a:* which is produced by an unchecked vocal tract. Alternating the fingers, while blowing the ***nay***, can produce various sounds in the same way as the organs of speech can make different sounds at different points of articulation. Similarly, IJ also maintains that the ***lute*** may be likened to the speech apparatus. When an unchecked string is plucked, it produces a simple sound. But the checked strings of the ***lute*** can produce various sounds when plucked at different places with varying states of the strings. Here, the lute string is likened to the human vocal tract; and plucking at various places is compared to the articulation at different points. Although he admits that musicology is not relevant to his book [1], he maintains that the science of phonetics is relevant to music in so far as sound-making is concerned. It is interesting to note that Dr. William Holder, an early English phonetician and music theorist (d. 1698) makes similar comparison with the lute, horn, cornet and trumpet [6].

CONSONANTS AND VOWELS

IJ recognizes the two main classes of sounds for Arabic and other languages namely the C's and V's. This is based on his criterion of the free vs non-free air passage along the vocal tract, as stated earlier. He discusses these classes in two different sections of the ***Introduction*** to ***SS***. Furthermore, in his analysis V's are unlike C's in terms of lengthening. V's can be freely and naturally prolonged [1].

SOUND PATTERNS

Speech sounds are patterned in every language in accordance with its phonological and phonotactic rules. IJ makes numerous statements regarding the arrangement and combination of the Arabic C's and V's. His discussions concerning their distribution, co-occurrences, possible / impossible and acceptable /non-acceptable combinations, syntagmatic / paradigmatic relations are quite intelligent and interesting even for a modern phonologist. For instance he remarks that Arabic has no word-initial vowels or clusters, unlike other languages, such as Persian [1], [2] and [7].

PHONETIC ORDER OF ALPHABET

Although many of IJ's statements concern Arabic, their implications are universally valid. He presents a list of the Arabic letters based on the ascending phonetic order. That is to say, the 29 letters are ordered according to their points of articulation starting from the larynx, through the various areas of articulation along the vocal tract, until finally ending with the labials. In this ordering, he differs slightly from his predecessors [7].

POINTS OF ARTICULATION

IJ acknowledges 16 points and, again, he lists them in the ascending phonetic order as follows :

A. *GUTTURALS*
1. Laryngeals *ʔ, aː, h*
2. Pharyngeals *ʿ, ḥ*
3. Uvulars *ʀ, χ*
4. Post-Uvulars *q*

B. *SOFT AND HARD PLATE*
5. Velars *k*
6. Palatals *j, y, ʃ*
7. Palato-Lateral

C. *ALVEOLARS*
8. Lateral *l*
9. Nasal *n*
10. Trill *r*
11. Alveolars *t, d, ṭ*

D. *DENTALS*
12. Dento-alveolars *s, z, ṣ*
13. Inter-Dentals *θ, ð, ḍ,*
14. Labio-Dental *f*
15. Labials *b, m, w*
16. Homorganic Nasals *ŋ, ɴ*

He is also aware of the active and passive articulators as can be deduced from his description. This arrangement is both interesting and practical, though his inclusion of *a*: amongst the laryngeals, and the homorganics is not accepted in modern phonetic practice.

MANNER OF ARTICULATION

Afterwards IJ classifies the Arabic phonemes in terms of the manners in which they are produced. The phonemes are grouped under one or more than one of the following classes :

1. Voiced or Voiceless
2. Plosive or Fricative
3. Emphatic or Non-Emphatic
4. Raised or Lowered back of tongue
5. Lateral or not
6. Trilled or not
7. Strongly-released plosive or not
8. Nasal or not

DISTINCTIVE FEATURES (DF's)

Underlying this description there seems to be a kind of a binary DF analysis. There are many statements which may lead to this deduction. To exemplify, we know that velarization or emphaticness is a DF in Arabic. IJ states: "without emphaticness *ṭ* would become *t; ṣ* would become *s*; and *ð̣* would be *ð*" [1].

EXPERIMENTATION

The phonetic description of the speech sounds as presented by IJ is based on a thorough knowledge of the structure of the Arabic language, a very good observation of its sound mechanism, and an underlying theoretical approach to the phonetic phenomena. In addition, the phonetic descriptive techniques are enhanced by empirical methodology. In his analysis, IJ resorts to experimentation whenever it is possible in order to support his data or arguments. This can be shown from the following statement concerning the nasals: "the voiced *m* and *n* are articulated in the mouth as well as in the nose and hence they are characterized by nasality. To demonstrate, if you hold your nose while uttering them you will not be able to produce them" [1].

Another example demonstrating his inclination to experimental method can be inferred from the following quotation from *SS*: "If you intend to get the right sound of a consonant, you must pronounce it alone, unfollowed by a vowel since this can change its quality. But since Arabic does not allow initial clustering, you add the sequence *ʔi-* before the consonant. Thus you can say: *ʔik, ʔiq, ʔij*, etc." Although this shows a phonological influence on the analysis, yet it is clear here that IJ tries various empirical techniques to get the correct results.

CONCLUSIONS

The presentation above is both brief and incomplete. Yet the intention here is to demonstrate IJ's contribution to general phonetics. The ideas or statements extracted from his works are arranged herein in such a way as to reflect genuinely the organization of the data and information in his own works. This paper has only touched on some of the salient issues in the present discussion. There remain some other notions such as pause, lengthening, syllable, speech defects, and suprasegmental features which deserve further consideration. Ibn Jinni's works and contribution to phonetics and linguistics deserve to be acknowledged in any scientific account of the history of linguistics and phonetics.

REFERENCES

[1] Ibn Jinni, ʿUthman (1985), *Sirr Ṣināʿat al-I ʿrāb*, edited by Hasan Hindawi, Damascus: al-Qalam Publishers.

[2] Ibn Jinni, ʿUthman (1952-56), *al-Khaṣā'iṣ*, 3 vols, edited by M.A. al-Najjar, Cairo: Dar-al-Kutub Press.

[3] Greenberg, J.H. (1953), "Historical linguistics and unwritten languages", ***Anthropology Today***, edited by A. Kroeber, pp. 256-86, Chicago: University of Chicago Press.

[4] Koning, Pieter de (1903), ***Trois traités d'anatomie arabes***, par Muhammad Ibn Zakariyya al-Razi (d. 925), Ali ibn al-ʿAbbas al-Majusi (d. late 10th century), et Abu Ali Ibn Sina (d. 1037), Leiden: E.J. Brill.

[5] Bakalla, M.H. (1994), "Arab and Persian phonetics", ***The Encyclopedia of language and linguistics***, edited by R.E. Asher and J.M.Y. Simpson, Oxford: Pergamon Press, Vol. 9, pp. 187-191.

[6] Holder, William (1669), ***Elements of Speech : An essay of inquiry into the natural production of letters***, London : T.N. for Martyn Printer. Reprinted, New York : AMS, Inc., 1975. Also cf. Peter F. Ostwald (1973), ***The Semiotics of Human Sound***, The Hague: Mouton, pp. 32ff.

[7] Bakalla, M.H. (1982), ***Ibn Jinni : An early Arab Muslim Phonetician. An interpretive study of his life and contribution to linguistics***, London and Taipei : European Language Publications.

PROTOCOL ANALYSIS OF THE PROCESS OF TRANSCRIPTION

J. Fokes and Z.S. Bond
Ohio University, Athens, Ohio, USA

ABSTRACT

In the first experiment, the phonetic transcription of three undergraduate students was subjected to protocol analysis to determine whether proficient vs. mediocre transcribers used different strategies. A second experiment compared the transcription strategies used by two students early and again later in their training.

INTRODUCTION

In spite of its importance, the process of learning to transcribe phonetically has received little empirical attention. We investigated the hypothesis that proficient vs. mediocre students apply different strategies when faced with a transcription task and that student strategies change with practice. Protocol analysis [1], a procedure devised to investigate problem solving, was adapted to discover strategies used in phonetic transcription.

EXPERIMENT I

Three undergraduate students in introductory phonetics, one earning an A (LF), and two who were less skilled (LD, ST) transcribed a 170 word passage. They were instructed to talk their thoughts out loud while transcribing. All their comments were recorded. They reported no difficulty in verbalizing and found it to be quite natural.

Student comments were written as protocols. Analysis yielded the following classification which represents the expected FLOW of the process of transcription:

1) SCAN: preliminary reading of a sentence or phrase.

2) RECOMBINATION: grouping previously transcribed material with new material.

3) FOCUS: attention in the attempt to transcribe a unit: phrase; word; partial word; syllable; consonant; or vowel.

4) METHOD OF ATTACK: repetition of a unit; blind repetition with no variation; systematic repetition with changes in pronunciation; memory aids or other devices for transcription; orthographic cues.

5) DECISION: exit, final unit uttered signifying completion of transcription; evaluation, comments about transcription.

LF, the proficient student, completed the transcription in 15 min. with 34 errors. The mediocre students required more time; LD required 30 min. with 154 errors and ST 37 min. with 94 errors.

Analysis of the protocols showed three systematic differences between the proficient and the mediocre transcribers: 1) their initial approach to the task, 2) the primary units on which they focused during transcription, and 3) the method of attack.

LF SCANNED up-coming phrases before she attempted to focus on a unit for transcription. This preliminary scan was rarely used by the other two transcribers.

LF used RECOMBINING more than twice as often as LD and ST. LF read lengthy phrases such as "By way of introduction, I'd like..." Almost all LF's efforts were preceded by some type of scan before focusing on a unit for transcription. In contrast, both LD and ST limited their recombinations to two or three words, and sometimes even partial words.

Initially, the students FOCUSED on a unit and subsequently experimented with the details of the unit. They differed in the size of the unit for focus. LF dwelt on relatively large units such as phrases or words. LD initially focused on words but quickly fragmented words into syllables and vowels. When LD did focus on words, she selected short words such as *the*, *to* or *have* while LF concentrated on multi-

syllabic words such as *united*, *immigrant*, or *introduction*. ST typically focused on segments or syllables. The numerical differences of units of focus are given in Table 1.

ST's transcription of *problems* contained two vowel errors as might be expected from her piecemeal approach.

After selecting a unit for focus, the subjects employed a METHOD of attack for transcription. Initially, a REPETITION seemed to be used to replay or recheck a word. LF and LD rechecked a unit more than twice as often as ST. LF rarely used BLIND REPETITION whereas LD and ST repeated words without any variation in pronunciation.

Table 1. Units of focus selected in transcription.

	Phrase	Word	Word-part	Syllable	Vowel	Consonant
LF	73	70	15	37	5	2
LD	14	80	12	83	60	8
ST	16	46	27	116	67	15

Table 2. Methods used in transcription.

	Repetition	Blind-rep	Sys-rep	Mem aids	Orthography
LF	68	9	144	26	3
LD	59	75	20	19	32
ST	25	103	24	7	2

In a few instances, LF focused on smaller units but with considerable recombining, as in the following sample:

51. by way of introduction
52. [Introdʌkʃən]
53. by way of [Intrədʌkʃən]
54. [Introdʌkʃən]
55. [Intrədʌkʃən]
56. [ʃən]
57. [Intrədʌkʃən].

LD initially focused on words and then fragmented the word into smaller and smaller units:

88. [Intrədʌkʃən]
89. [Introʊ]
90. [dʌ]
91. [dʌk]
92. [ʃən].

ST rarely focused on words as units but instead attacked syllables and sounds. It was sometimes difficult to determine just which word she was working on:

196. [pra]
197. [pʌ]
198. [pɹa]
199. [pɹa]
200 [bəɪɛmz].

On occasion, subjects used MEMORY AIDS about transcription. These were helpful when correct, but comments were sometimes erroneous. LD used ORTHOGRAPHY as a cue. The numerical differences given in Table 2 point out preferences in method.

LF used SYSTEMATIC REPETITION, varying her pronunciation to determine the most appropriate for her transcription. The following shows her approach to transcribing "immigrant families".

33. [Imɑgɹɪnt fæmlɪz]
34. [tu ɪmɑgɹənt]
35. [Im ɪgɹɪnt]
36. [Imɑgɹɪnt]
37. unstressed
38. ['Imɑgɹənt]
39. [Im'ɪgɹənt]
40. ['Imɑgɹənt fæməlɪz]

41. [fæmlɪz]
42. You don't say [fæmllɪz]
43. [fæməlɪz]

In contrast, LD repeated blindly. In her attempt to transcribe each, she confused orthography with phonetic symbols.

464. [itʃ]
465. I don't hear the a
466. [itʃ]
467—470. [i]
471. [æ]
472. [itʃ]
473. Why don't I hear a in these words?
474—475. [itʃ]

ST also used blind repetition of parts of words. In working on *promised*, she pronounced the word only when her transcription was complete. Her repeated vowel was [a] although she wrote [ɔ] as in

259. [pi]
260. [pɔ]
261—264. [a]
265. [ɜ]
266. [a]
267—268. [I]
269. [Ist]
270. [pramIst]

In addition, ST used memory aids such as *cow* to remind her of the diphthong [aʊ] more often than the other two transcribers.

The subjects used different patterns to indicate that they were satisfied with the results of transcription. LF used repetitions, usually recombining them with upcoming material. For example, after transcribing *united*, she combined the word with the next item *states*. Both LD and ST tended to repeat only the target word when completing transcription.

In summary, the proficient transcriber attacked the problem of transcription in a different manner from the two less competent transcribers. She initially scanned a portion of the material to be transcribed before focusing on units for transcription. Her units of focus were typically phrases or words. Her predominant method of attack was systematic repetition. She combined previously transcribed material with upcoming material before focusing on the next unit. In contrast, the mediocre transcribers used scanning to a limited degree and tended to focus on small units. LD focused first on words and then fragmented them into syllables and sounds. ST worked from a sound-up direction and often did not pronounce the whole word. Repetitions by LD and ST did not seem to be experimental but simply another attempt to hear the word.

The resulting question is whether a proficient transcriber uses more advanced strategies because of competence or is competence a result of advanced strategies? If students are provided with additional training, will changes in strategies occur? The second experiment is designed to answer these questions.

EXPERIMENT II

In the second experiment, two students (LH and KT) were asked to transcribe the original passage midway in their phonetics course (T1) and another 154 word passage six weeks later (T2). Protocols describing their thoughts while transcribing were obtained at both times. The protocols were analyzed as in the first experiment.

KT, the proficient student, transcribed the T1 passage in 25 min. with 15 errors and LH, the mediocre student, in 30 min. with 49 errors. At T2, both students required the same amount of time as in T1. Both error counts were reduced: KT's transcription was nearly perfect with three errors while LH made ten errors, a vast improvement in accuracy over her first attempt.

Both transcribers SCANNED words and phrases, both scanning phrases more often at T2 than T1. Only LH scanned sentences at T2. KT used recombining of transcribed material with new material, often as a check of her previous work. LH recombined syllables to build up words at T1 but abandoned this procedure by T2.

Both transcribers FOCUSED on words and syllables as units at T1 with LH focusing on syllables more often than KT. By T2, LH changed her unit of focus from syllables to words and phrases, indicating an ability to handle larger units. The changes in focus for LH and KT at T1 and T2 are given in Table 3.

LH and KT used many Systematic Repetitions as a METHOD OF ATTACK in transcribing words, especially those which they found difficult. KT increased the number of times she used systematic repetition from T1 to T2 while LH used approximately the same number. KT monitored her transcription more than LH did, offering comments such as "I hear a schwa". KT referred to orthography surprisingly often, with observations such as "That's not a t" in transcribing a word with initial [ð].

Table 3. Units of focus selected by LH and KT at T1 and T2.

	Phrase	Word	Word-part	Syllable	Vowel	Consonant
LH (T1)	11	107	10	106	17	6
LH (T2)	20	77	13	40	2	4
KT (T1)	8	109	17	84	18	3
KT (T2)	12	63	13	82	6	5

Both LH and KT repeated words or phrases in EVALUATING their work, indicating satisfaction with the transcription. KT sometimes made evaluative comments when terminating transcription. For example, she would reread as in this example, "...or at least reduce static electricity in your body...yeah!" Both made extraneous comments or sighed audibly, particularly when faced with difficult stretches of transcription.

KT was a fairly accurate and proficient transcriber at T1. Since she was doing well, she may not have seen any need to make major changes in her approach to transcription, either in her initial scans, her units of focus, or in her methods of attack. LH was probably aware of some need to improve her performance. At T2, she began to scan longer units and to focus on words and phrases rather than syllables and segments. The accuracy of her transcription improved at the same time.

In conclusion, we believe the observed evidence between strategies and performance in transcription is compelling. Although our data are based on association rather than a cause-and-effect relationship, the differences in strategies between proficient and mediocre students reflect greater accuracy in transcription. Changes are evident with additional training as well. We plan to incorporate some explicit instruction in transcription strategies in the future. Training will be based on an organized FLOW in which the initial SCAN on mega-units rather than micro-units is emphasized to the extent that students can handle larger units. Scanning will be followed by FOCUS on the unit to be scrutinized for transcription. SYSTEMATIC REPETITION of the focal units is the METHOD recommended in determining sound-symbol relationships. RECOMBINING as a strategy throughout will provide for check on old material and the approach to new material. Final EVALUATION of the unit under attack will provide greater assurance of accuracy in the transcription process.

REFERENCES

[1] Hayes, J.R. (1981), The Complete Problem Solver, Philadelphia: Franklin Institute Press.

WHAT DO TRANSCRIPTION AGREEMENT INDICES SAY ABOUT TRANSCRIPTION ACCURACY?

Catia Cucchiarini
Centre for Language and Migration, Louvain, Belgium

ABSTRACT

This paper deals with the drawbacks of a common measure of transcription agreement, percentage agreement. It is argued that this metric does not give a realistic representation of transcription similarity and that it can be easily inflated by adopting a higher level of abstraction (one involving fewer categories) than the one recorded in transcriptions, when calculating agreement. An alternative measure of transcription (dis)similarity is presented and its advantages over percentage agreement are discussed.

INTRODUCTION

In the last few years the issue of transcription reliability has received considerable attention in the literature (for a review see [1]). Since it is known that phonetic transcriptions tend to contain an element of subjectivity, it is now common practice to check the objectivity and accuracy of transcription data before using them for research. To give an indication of the accuracy of the transcriptions on which their findings are based, researchers usually provide so-called transcription reliability or agreement indices. The most used measure for this purpose is percentage agreement, which is computed by comparing two transcriptions symbol by symbol and by taking the percentage of identical symbols in the two strings. Although this index in reality expresses agreement between transcriptions, the term reliability index is often used instead [1]. However, this is not correct given that phonetic transcription involves classification into categories (phonetic symbols) which are not ordered. In other words, the variables have the properties of measurement at the nominal level. At this level there can be no "proportionality of ratings" [2], a notion which is crucial to reliability. For these reasons, the term agreement index will be used in the present paper (for further details, see [2] and [3]).

In general, no standards or levels of significance are available for transcription agreement indices. Although it seems that indices should be as high as 75% [4] or 85% [1, 5] in order to be acceptable, the plausibility of these agreement values has never been considered, let alone demonstrated. Consequently, it is not clear whether high percentages of agreement really correspond to high degrees of transcription accuracy. This point is addressed in the following section.

THE IMPACT OF CHANCE AGREEMENT

One of the things that tend to be overlooked in the literature is that the value of an agreement index does not only depend on the degree of accuracy of the transcriptions in question, but also on the number of categories on which two transcribers, or one and the same transcriber on different occasions, have to agree. The number of categories involved in the judgement partly determines the impact of chance agreement, that part of agreement that is determined by chance alone. It can be stated that agreement indices tend to be higher for simple judgements such as correct / incorrect than for more complex decisions involving a greater number of

categories like, for instance, the precise description of the vowels produced by a speaker. Consequently, agreement indices are expected to be higher for broad transcriptions than for narrow transcriptions. This means that providing an agreement index is not sufficient to give a precise idea of the degree of accuracy of transcriptions. One should also mention the number of categories out of which transcribers could choose (level of abstraction).

Unfortunately, the impact of chance agreement on transcription agreement is often overlooked in the literature (see for instance [1], [4]), with the result that in computing agreement researchers often reduce the number of categories in order to achieve the longed for 75% or 85% of agreement. However, as has been pointed out [6] "higher coefficients of agreement that result from use of simpler observation codes do not guarantee that observer recordings are accurate".

The fact that agreement indices are so sensitive to the number of categories involved has to do with the way in which transcription differences are treated, when it comes to determining the degree of agreement. In general it is assumed that certain transcription deviations are more serious than others. For instance, transcribing [b] instead of [p] would be considered to be less serious a mistake than transcribing [l] instead of [p]. Similarly, differences concerning diacritical marks are assumed to be less serious than differences concerning basic symbols.

However, these differences in gravity are usually neglected when transcriptions are compared symbol by symbol. The only thing researchers look at is whether the symbols and the accompanying diacritics are identical in the two transcriptions or not. The result is that any difference will affect the agreement index in the same way, regardless of its degree of gravity. In turn the agreement index will be extremely sensitive to the degree of detail recorded in the transcriptions. This also means that this kind of index can easily be inflated by reducing the degree of detail, not when making the transcriptions, but when calculating agreement.

In addition to making percentage agreement so subject to manipulation, this procedure is also unrealistic. It is obvious that a measure of transcription agreement should take account of the various degrees of (dis)similarity between speech sounds. Moreover, when diacritics are present, one should not merely check whether the same diacritic is used or not, as is sometimes done [1]. As a matter of fact, a diacritic is an integral part of the phonetic symbol, since it partly determines its meaning, so it would be wrong to consider them as separate elements. Furthermore, different diacritics used with different basic symbols could represent very similar speech sounds. For example, the two vowel symbols [o] and [ɔ] can be made more similar by adding appropriate diacritics for 'height' properties as follows: [o̝] [ɔ̝]. The higher degree of similarity between these two transcriptions would not be reflected by percentage agreement. In fact, in this metric the two differences would be combined thus obtaining a very low agreement index.

AN ALTERNATIVE APPROACH TO TRANSCRIPTION EVALUATION

In the previous section I have argued that percentage agreement is no adequate measure of transcription similarity, because it is too sensitive to the level of abstraction of transcriptions and because it treats agreement between phonetic symbols in an all-or-none way. In an attempt to overcome these problems, an alternative measure of transcription (dis)similarity was developed, which does take account of the various degrees of similarity (or difference) between speech sounds and of the effect of

diacritics on basic symbols [3]. This metric is called average distance because it gives an indication of the mean distance between the vowels and/or the consonants of two transcription strings.

The average distance is based on the feature matrices defining vowels and consonants that are presented in [7]. These matrices were obtained by combining results of experiments on proprioceptive speech sound dissimilarity with phonetic knowledge. The values contained in these matrices make it possible to express the degree of dissimilarity between all possible pairs of sounds in numerical form. Each speech sound is assigned a numerical value for each of the defining features in the matrices. Dissimilarity values for pairs of speech sounds can be determined by calculating city-block distances between them. This is done by comparing two speech sounds feature by feature and by summing the individual differences. Overall dissimilarity values for the vowels and consonants contained in each transcription pair are obtained by computing the mean for all vowel and consonant pairs, respectively.

One of the advantages of this method is that it gives a more realistic impression of the degree of (dis)similarity between two transcriptions. For instance, with this metric it is possible to indicate that there is more similarity between [b] and [p] than between [l] and [p]. In other words, this metric goes beyond the mere appearance of phonetic symbols (are they identical or not?) and takes account of their meaning (which speech sounds do they represent and how are they related to each other?). Moreover, in this metric it is possible to discount the impact of diacritics on basic symbols before computing the distance between two corresponding symbols. Also in this case the meaning of diacritics is considered (what is the effect of adding this specific diacritical mark to this basic symbol?) and not merely their presence or absence.

This brings us to another advantage of this metric, namely that in computing agreement one can take account of all the details that have been recorded in transcriptions in a realistic way. It does not make sense to carry out transcriptions at a certain level of abstraction, for instance narrow transcription, and then compute agreement at a higher level of abstraction, i.e. broad transcription, in order to achieve acceptable percentages of agreement. If researchers make narrow transcriptions there must be a reason for this, i.e. the details are relevant to their research. It is therefore important to know to what extent transcribers agree at this level of specificity. It is obvious that the more details transcribers record, the less likely they are of agreeing with each other. However, one should avoid using a measure such as percentage agreement which penalizes detailed transcriptions in an unwarranted way.

That the average distance is a more appropriate measure than percentage agreement was also revealed by the results of an evaluation test described in [3]. For 50 transcription pairs the overall dissimilarity between vowels and consonants was computed by means of the two metrics, i.e. the average distance and percentage disagreement, the complement of percentage agreement. The values thus obtained were compared with the dissimilarity judgements expressed by 19 experienced phoneticians for the same transcription pairs. The phoneticians were asked to assign a mark varying between 1 (no similarity) and 10 (no difference) to the vowels and consonants of each transcription pair. The reliability coefficient computed for these judgements (formula for composite ratings with raters as a random factor) appeared to be high (0.97).

It turned out that the average distance better reflected the phoneticians' judgements than percentage agreement. As a matter of fact, the correlation coefficient was higher in the former case ($r = -0.86$, $df = 48$, $p < 0.01$) and lower in the latter ($r = -0.68$, $df = 48$, $p < 0.01$). The two coefficients also appeared to be significantly different ($t_{47} = -2.94$, $p < 0.01$). It should be noted that the correlation coefficients are negative in both cases because the phoneticians' judgements indicate similarity and the other two measures dissimilarity.

On the basis of this test it seems that when phoneticians judge the degree of (dis)similarity between pairs of transcriptions, they do not limit themselves to establishing whether the symbols in the two strings are identical or not, but try to determine to what extent they are similar. Apparently, phoneticians consider agreement between phonetic symbols to be gradual, not all-or-none. This is precisely what happens when the average distance is calculated (for a fuller account of this method and of its advantages over percentage agreement, the reader is referred to [8]).

CONCLUSIONS

The new measure of transcription agreement proposed in this paper, the average distance, differs from the more common percentage agreement, because it makes it possible to indicate different degrees of (dis)similarity between corresponding phonetic symbols. Only two of the advantages of this method are discussed here. First, the average distance gives a more correct representation of transcription (dis)similarity because it takes the meaning of phonetic symbols and diacritics into account. Second, since different degrees of (dis)similarity between phonetic symbols can be distinguished, this measure is less sensitive to the level of abstraction of transcriptions. For these reasons it seems that the average distance provides a more appropriate measure of transcription agreement than percentage agreement. This was also confirmed by experimental results.

ACKNOWLEDGEMENTS

The present research was supported by the Linguistic Research Foundation, which is funded by the Netherlands Organization for Scientific Research, NWO. I am indebted to M. Biemans for the realisation of the transcription evaluation experiment described in this paper.

REFERENCES

[1] Shriberg, L.D. & Lof, L. (1991), "Reliability studies in broad and narrow phonetic transcription", *Clinical Linguistics and Phonetics*, vol. 5, pp. 225-279.

[2] Tinsley, H.E.A. & Weiss, D.J. (1975), "Interrater reliability and agreement of subjective judgments", *Journal of Counseling Psychology, Vol.* 22, pp. 358-376.

[3] Cucchiarini, C. (1993), *Phonetic transcription: a methodological and empirical study* (PhD thesis, University of Nijmegen).

[4] Henderson, F.M. (1938), "Accuracy in testing the articulation of speech sounds", *Journal of Educational Research*, vol. 31, pp. 348-356.

[5] Pye, C., Wilcox, K.A. & Siren, K.A. (1988), "Refining transcriptions: the significance of transcriber 'errors'", *Journal of Child Language*, vol. 15, pp. 17-37.

[6] McReynolds, L. & Kearns, K.P. (1983), *Single-subject experimental designs in communicative disorders* (Austin, TX: Pro-Ed).

[7] Vieregge, W.H., Rietveld, A.C.M. & Jansen, C.I.E. (1984), "A distinctive feature based system for the evaluation of segmental transcription in Dutch", *Proceedings of the Xth International Congress of Phonetic Sciences*, Utrecht.

[8] Cucchiarini, C. (1996), "Assessing transcription agreement: methodological aspects", to appear in *Clinical Linguistics and Phonetics*, vol. 10.

Some figures concerning the transliteration of the Dutch Speech Styles Corpus

Els den Os, Marian Ellens, Cor in 't Veld, and Lou Boves
SPEX, Leidschendam, The Netherlands

ABSTRACT

On the basis of orthographic transliterations of monologues, picture descriptions and read short texts of in total 127 speakers (in total more than 19 hours of speech), we present data concerning speech rate, hesitational sounds, clitic groups, and verbally deleted words (repetitions, repairs at the beginning of an utterance (false starts), and repairs later in an utterance).

INTRODUCTION

The Dutch Speech Styles Corpus was collected to investigate the voice quality of speakers of standard Dutch. The speech material was designed by R. van Bezooijen and the speech recordings were made by J. van Rie and R. van Bezooijen. The corpus contains three different speech styles: spontaneous speech (monologues), semi-spontaneous speech (picture descriptions), and read speech. The speech was always recorded in the presence of a female 'interviewer' of about 30 years old. In all three styles the speech contents refer to domestic topics, eating habits, and food.

There are 127 speakers, in three age categories: 30 speakers (17 males and 13 females) from 10 to 20 years old, 45 speakers (19 males and 26 females) from 20 to 60 years old, and 52 speakers (24 males and 28 females) between 60 and 86 years old. The total duration of speech is 19 hours and 10 minutes (4 hours and 40 minutes of monologues, 10 hours and 20 minutes of picture descriptions, and 4 hours and 10 minutes of read speech). The total number of *word forms* that were transcribed in the corpus amounts to about 118,000. There are about 6,300 *different* word forms.

The whole corpus has been transliterated. Among other things, these transliterations offer the possibility to study disfluencies in several speech styles. The goal of this paper is to compare disfluencies in two types of spontaneous speech and read speech. By disfluencies we mean in this paper specifically hesitational sounds (filled pauses like 'uh'), and verbally deleted words, i.e. words spoken, but superseded by subsequent speech.

TRANSLITERATION

The transliteration of the corpus is a word level transcription of what the speakers said. The standard spelling of Dutch is used. This necessarily implies a compromise between the sounds heard and what has to be written down.

Because it could be expected that some reduced forms of words (mostly containing schwa's) would occur more often than the full forms, it was allowed to write these reduced forms down in a sometimes non-standard way; see [1] for a more detailed description of the transliteration of the corpus.

The Speech Styles Corpus has been labelled at the utterance level (i.e. a time stamp between utterances is provided to allow access to the speech files). The notion of what constitutes an utterance in spontaneous speech is necessarily an arbitrary one. An utterance was defined as a number of words being semantically consistent and containing at least a subject and a verb. In addition, this string had to be preceded and followed by a clear acoustic pause.

Clitization

Clitization which resulted in syllable deletion was indicated and for this we could not always use existing spellings. However, we decided to mark these forms, since we wanted to know how often these forms occur in spontaneous Dutch speech. These forms are of significance in relation to automatic segmentation programmes and training of speech recognizers.

MARKINGS

Next to the orthographic transcriptions, conventions were used to indicate all audible events that occur during speaking. These conventions consist of different kinds of brackets with or without additional information, see also [2] and [3] for comparable markings used in the ATIS and Switchboard corpus. Only those conventions will be presented below that will be discussed in the following.

Hesitational sounds

A distinction was made between hesitational sounds which were uttered in isolation and those which were uttered connected to the preceding word. There were four types of hesitational sounds pauses: [uh], [um], [mm], and [naa]. For a hesitational sound to be isolated, a silent pause has to occur before and after it.

Words spoken by the interviewer

Words spoken by the interviewer are indicated with curly brackets { }. The interviewer interfered most in the monologues (3700 words) opposed to 1021 words in the picture descriptions and 1 word in the read texts. Speech rate per style was calculated on the basis of the total number of words (those by the speakers as well as those by the interviewer). This procedure had to be followed, since we could only use the total duration per style per speaker to calculate speech rate. Since the interviewer was the same person most of the time, we think the used procedure is justified.

Verbally deleted words

Words verbally deleted by the subject are enclosed in angle brackets. Verbal deletions are words spoken by the speaker but which are superseded by subsequent speech. This can occur explicitly (<at> <the> <grocery> <I> <mean> at the bakery.....) or implicitly (<at> <the> <grocery> at the bakery...). Both can occur at the beginning of an utterance (false start) or later in an utterance. Verbally deleted words can be literally repeated or can be repaired. Word fragments are also indicated by angle brackets (<ba> bakery).

After the transliterations were completed, the utterances containing angle brackets were selected, and a classification was made in different types of verbally deleted words:

(1) *repetitions*: literally repeated words, word groups or word fragments. They do *not* occur at the beginning of an utterance.

(2) *false starts*: the speaker starts an utterance producing a word, word group or word fragment, but he/she decides to start all over again. The beginning of an utterance was defined as the first two words.

(3) *repairs later in the utterance*: the speakers interrupts a word, word group or word fragment and continues the utterance in a different way.

SPEECH RATE

The overall speech rate measured in words per minute was somewhat higher for reading texts (156 words per minute) than for monologues and picture descriptions (136 and 129 words per minute respectively). No difference was observed between the speaking rates of male and female speakers. Nor was any great difference observed between the various age categories, although in the

monologues the young speakers spoke somewhat more slowly (123 words per minute) than the older speakers (143 words per minute). It must be noted that in this calculation the time spent in pausing was included.

CLITIZATION

The total number of clitic forms that resulted in syllable deletion amounted to about 450. This is only 0.4% of the total number of words in the corpus. Especially the forms including the personal pronoun *ik* 'I', the verb form *is* 'is', and the personal pronoun *het* 'it' involve syllable deletion.

FILLED PAUSES

Filled pauses only occurred in the monologues and in the picture descriptions. There were no differences in the number of filled pauses between these two styles.

Related to the total number of words produced per speech style by male or female speakers of one of the three age categories, the percentage of filled pauses is between 2.6% (picture descriptions by male speakers of 20 to 60 years of age) and 7.5% (picture descriptions by male speakers under 20 years of age). The younger speakers produced more filled pauses (on average 7%) than the older ones (on average 5%). In both styles the relative number of connected filled pauses was about the same as the relative number of isolated filled pauses (2.8% and 2.3%, respectively). In Dutch, speakers often connect pauses to function words, e.g. *en[uh]* 'and[uh]'. We only observed very few instances in which the filled pause was connected to a content word.

VERBALLY DELETED WORDS

In the following, we present the percentage of words that are themselves verbally deleted or are involved in a verbal deletion. The data are given for male and female speakers, and for the age groups separately. We only present data for the monologues and picture descriptions, because verbally deleted words did almost not occur in the read texts.

Repetitions

In table 1 it can be observed that there are no clear differences between the age groups, sex, or styles for the percentage of repetitions. The percentages range from 0.4 to 1.2.

Table 1: Percentage of repetitions for Female (F) and Male (M) speakers for the three age categories (1=<20, 2= between 20 and 60, and 3=>60 years of age), for monologues and picture descriptions.

	Monologues	Picture description
F1	0.8%	1.0%
F2	0.8%	0.7%
F3	1.0%	0.5%
M1	0.6%	0.9%
M2	1.1%	0.4%
M3	1.2%	0.8%

False starts

In table 2, the percentages of false starts are given. It can be seen that the younger speakers produce more false starts than the older ones in the picture descriptions. The male and female speakers between 20 and 60 years of age produce relatively few false starts. The percentages range from 0.4 to 2.1.

Verbally deleted words later in the utterance

In table 3 the percentages of verbally deleted words later in the utterance are given. It can be observed that the older male speakers produce most verbally deleted words (2.9%) in the monologues and that the young male speakers produce most verbally deleted words in the picture descriptions (2.2%).

Table 2: Percentage of false starts for Female (F) and Male (M) speakers for the three age categories (1=<20, 2= between 20 and 60, and 3=>60 years of age), for monologues and picture descriptions.

	Monologues	Picture description
F1	0.6%	2.1%
F2	0.5%	0.6%
F3	0.9%	0.8%
M1	1.3%	1.3%
M2	0.4%	0.4%
M3	0.8%	0.8%

Table 3: Percentage of verbally deleted words later in the utterance for Female (F) and Male (M) speakers for the three age categories (1=<20, 2= between 20 and 60, and 3=>60 years of age), for monologues and picture descriptions.

	Monologues	Picture description
F1	1.1%	1.3%
F2	1.1%	1.6%
F3	1.4%	1.6%
M1	2.4%	2.2%
M2	1.3%	0.9%
M3	2.9%	1.8%

DISCUSSION AND CONCLUSION

Here we will concentrate on the disfluencies in spontaneous and semi-spontaneous speech, since it turned out that disfluencies in the read text were almost not present. This is due to the fact that the sentences in the texts were short and simple. We all know that spontaneous speech is not fluent: speakers produce many hesitational sounds, mispronunciations, and verbal deletions. As far as we know, the number of these disfluencies have never been addressed on the basis of a large number of speakers of different age groups.

Our counts show that hesitational sounds occur on average about once every 20 words. Verbally deleted words (repetitions, false starts, and deleted words later in the utterance taken together) occur on average 3 times in every hundred words. The group of male and female speakers between the ages 20 and 60 years produced fewer verbally deleted words than the younger and older group. Especially in the picture descriptions, the younger speakers produced relatively many verbal deletions.

It must be remarked here, that most verbal deletions were repaired *implicitly*. There were very few instances of explicitly repaired deletions, like <dog> <I> <mean> cat. Furthermore, it must be noted that the disfluencies mentioned above were actually repaired; there are only few instances of disfluencies which were *not* repaired by the speaker.

Most verbal deletions occurred after a word was finished (77% of all verbally deleted words in the monologues and picture descriptions together). Verbal deletions after word fragments occurred less frequently. In the corpus, we observed very few instances of silent pauses within words (43 times). In most of these cases these pauses occur between the two parts of a compound. In addition, hesitational sounds almost never occurred within words. From this, and the fact that most verbally deleted words have been completed by the speakers, we may conclude that words are preferably articulated as a whole.

REFERENCES

[1] Den Os, E.A. (1994), "Transliteration of the Dutch Speech Styles Corpus", *Proceedings of the Institute of Phonetic Sciences, University of Amsterdam*, 18, 87-94.

[2] Hirschmann, L. (1992), "Multi-site data collection for spoken language corpus", *Proceedings ICSLP'92*, 2, 903-906.

[3] Godfrey, J.J., Holliman, E.C. & McDaniel, J. (1992) "Switchboard: Telephone Speech corpus for Research and Development", *Proceedings ICASSP*, 1, I-517 - I-520.

SOME PHONETIC CHARACTERISTICS OF IAAI

Ian Maddieson and Victoria B. Anderson
University of California, Los Angeles, USA

ABSTRACT

Iaai is an Austronesian language with a relatively large vowel inventory as well as some less-common contrasts among consonants. This paper presents the first detailed phonetic description of Iaai, paying particular attention to the formant structure and the lip positions of the vowels, and the articulation and acoustic characteristics of the releases of coronal consonants.

IAAI PHONETICS

Iaai [ja:i], one of the twenty-five or so indigenous languages of New Caledonia, a French "overseas territory" in the South Pacific, is spoken by about two thousand people on Ouvéa, the northernmost of the Loyalty Islands. Its grammar and lexicon have been described by Ozanne-Rivierre [1, 2]. However, there are no studies which have focussed on the phonetics of the language, and in particular no published instrumental phonetic studies.

A number of aspects of this language are of particular phonetic interest. For an Austronesian language, Iaai has a relatively large vowel inventory consisting not only of ten different vowel qualities, but also a phonemic length distinction. In distinguishing these vowels, large differences in lip rounding and spreading are used and these are independent of the front-back distinction. Moreover, there are interesting limitations on the distribution of certain vowels according to the consonant context. The consonant inventory is also quite extensive. The language has three coronal places of articulation, dental, retroflex and pre-palatal, for stops and nasals. In the stops, these three places appear to be acoustically differentiated along lines which differ from most other languages of the world which make use of such distinctions. The Iaai consonant inventory also contains voiced and voiceless sonorants which have phonemic status. What follows will present a general phonetic survey of the language, with emphasis laid on these various aspects of particular interest.

Detailed studies of some of these aspects will be presented, based on analysis of audio and video recordings of five speakers (three female and two male), and palatography for four of them. To characterize the vowels, formant measurements, durations and intrinsic pitch data were obtained from audio recordings. Using videotape, measurements were also made of lip aperture area, the height and width of this aperture, the distance between the outside corners of the lips, and the amount of side contact between them.

IAAI VOWELS

A standard chart of Iaai vowels is given in Figure 1. The vowels can be broadly divided into three height sets, three high vowels, four mid vowels and three low vowels. Front high vowels clearly contrast in rounding, as in /ţii/ 'tea' and /yy/ 'quarrel' (n.) (cf. /uu/ 'fall' (v.)). Mid back vowels /ɤ, o/ also contrast in rounding but the rounding contrast in mid front vowels is not functionally robust since /ø/ occurs in only a very few words. This vowel is always followed by a velar consonant and almost always preceded by a labial. The low vowels /a/ and /æ/ are largely in complementary distribution, with /æ/ restricted to occurrence after the labial consonants /b, m̥, m, p, f, ɥ, ɸ/ and the vowel /y/. This set of sounds also conditions a fronted [œ]-like variant of the lower mid back rounded vowel /ɔ/. Figure 2 shows the first two formants of both long and short variants of 9 of these vowels for the three female speakers. /ø/ is omitted as data is too sparse.

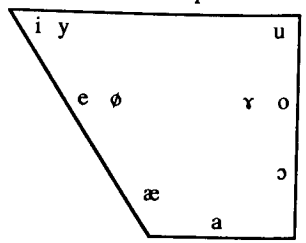


Figure 1. Chart of Iaai vowels.

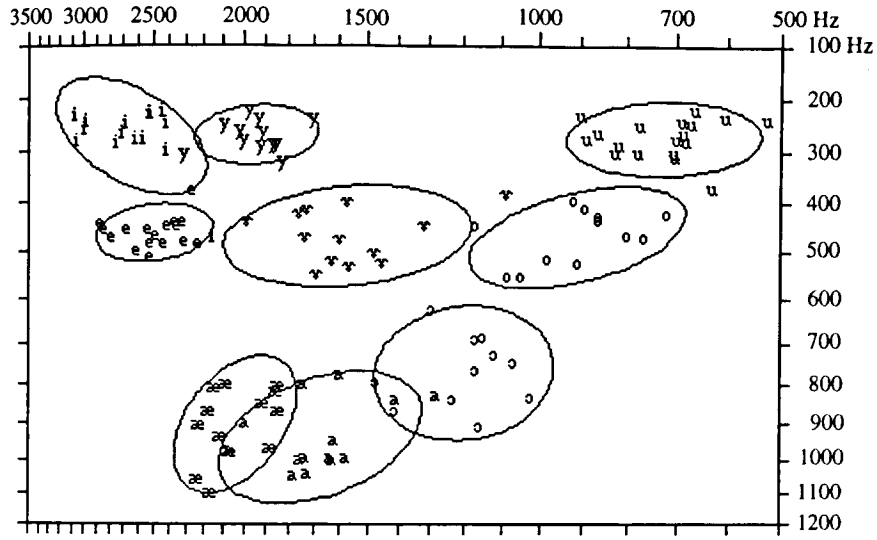


Figure 2. Formants of Iaai vowels (long and short combined) from three female speakers.

There are no consistent effects of length on the formant values; long vowels do not consistently have a higher or lower F1 or F2, nor are they more peripheral or more central in the acoustic vowel space than their short counterparts. Table 1 shows that the ratio of short to long vowel durations approaches 2.0.

Table 1. Long and short vowel durations.

	Short	Long	Difference
Women (9 vowels)	116.1	211.1	95.0
Men (8 vowels)	89.4	175.1	85.6

Lip position

The time-coded videotape was viewed frame-by-frame and the frame in which the lips reached the culminating position of the gesture for the particular vowel determined. This frame was digitized and a number of distances between lip points were measured in the transverse plane. These distances were *Lip Height* (distance between the lower surface of the upper lip and the upper surface of the lower lip at the center); *Lip Width* (the vertical distance between the point of contact of the lips at the left and the point of contact on the right); *Lip Corner* distance (the horizontal distance between the corners of the lip, i.e. the lateral margins of the vermilion border); *Side Contact* distance (the horizontal distance over which the lips are in contact between the corners and the aperture, i.e. *Corner* distance minus *Width*). Also *Lip Area* (the area of the visible opening enclosed by the lips) was measured. Since the video shows a frontal view, there is no quantitative data on lip protrusion. However, some qualitative idea of the degree of protrusion can be obtained. Useful information on the position of the tongue can also be obtained for those vowels with a more open jaw position.

Lip measurement results for three speakers, one male and two female, are shown in Table 2. To normalize across the speakers, all measurements were converted to standardized scores (with a within-speaker mean of zero and standard deviation of 1) before means were calculated and statistical tests performed. This transformation of the data, roughly, sets the value of a neutral lip position to zero. The values in Table 2 are the means of the within-speaker standardized scores. One can see, for example, that the area of the aperture between the lips in pronouncing /ɤ/ is very close to the mean lip area, /y/ and /u/ have the smallest area, indicated by the large negative number, and /a/ the largest.

Table 2. Mean normalized lip measurement values for ten Iaai vowels, from 3 speakers.

Word	*Vowel*	*Lip Height*	*Lip Width*	*Lip Corners*	*Lip Sides*	*Lip Area*
ṭii	i	0.53	0.70	0.89	0.31	0.69
yy	y	-1.41	-1.15	-0.63	0.70	-1.14
eet	e	0.19	1.26	1.27	-0.08	0.51
møøk	ø	-0.45	-0.75	-1.10	0.12	-0.78
vææt	æ	1.22	0.87	0.79	-0.62	1.12
aat	a	1.38	1.03	0.45	-1.18	1.51
θɔɔn	ɔ	0.61	-0.05	-0.61	-0.71	0.14
oṭ	o	-0.75	-1.01	-0.84	0.55	-0.92
ɤṭ	ɤ	-0.15	0.40	0.70	0.12	-0.01
ṭuṭu ~ kaluu	u	-1.17	-1.29	-0.92	0.79	-1.12

The vertical distance between the lips (*Lip Height*) is least for rounded non-low vowels, and greatest for unrounded low vowels. Although a three-way classification of vowels by height (high, mid, low) predicts a significant amount of this variance, the four mid vowels show quite substantial differences, with rounded mid vowels having higher lip position than unrounded ones. Most strikingly, the unrounded high vowel /i/ is more open than /e/ at the lips. The difference between /i/ and /e/ is made by raising the jaw independently of the lips, as the frames in Figure 1 show.

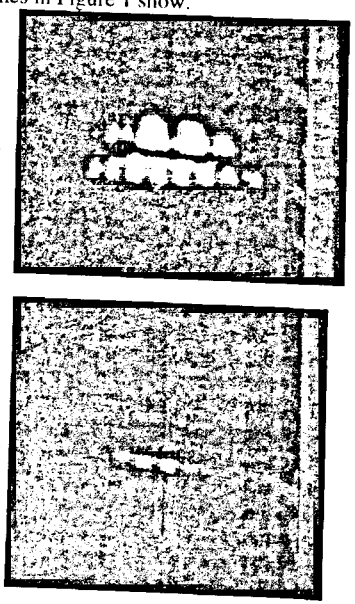


Figure 3. Video frames illustrating lip and jaw position in /i/ and /e/.

Lip Height does not divide rounded from unrounded vowels but *Lip Width* does. The distance between the outside corners of the lips (*Lip Corners*), however, is the best of our measures at effecting such a separation. All five rounded vowels have negative values of at least -0.61; all the unrounded vowels have positive values of 0.45 or greater. Moreover, the *Lip Corner* distance relates well only to the classification of vowels by rounding and not to classification by height as well. This measure seems the best index of lip protrusion when measurements are only taken in a flat plane, transverse to the body. Drawing the corners of the lips closer together is a consequence of protruding them.

Contact at the sides of the lips (*Lip Sides*) was measured following the suggestion of Goldstein [3] that "rounded vowels must be produced with contact along the sides". Although rounded high vowels have the greatest amount of side contact this measure does not separate the Iaai vowels into rounded and unrounded classes, and little of the variance in the *Lip Sides* measure can be predicted from the classification of vowels by rounding ($F(1, 29) = 2.87$, $p = .1011$).

Naturally enough, both *Lip Height* and *Width* measurements are very highly correlated with *Lip Area* (.95 and .92 respectively). Since /y/ and /u/ have the smallest height and width, they have the smallest area of lip opening, having almost identical mean normalized values. In this respect Iaai differs from a number of other languages with a similar pair of vowels, such as French, Swedish, Cantonese and Finnish, where the lip area for /y/ is considerably larger than that for /u/, and is actually comparable to that for /i/ [4]. Iaai also has a larger than expected area for /i/. *Lip Area* broadly separates rounded from unrounded vowels, with vowel height ranking vowels within those groups.

Formants and lip measures

Normalized F1 correlates most highly with *Lip Height*. Acoustic theory predicts a relationship between vowel height and F1; the more open a vowel, is the higher the F1 frequency. Despite the *Lip Height*./vowel height discrepancy with /i/ and /e/, lip height generally goes with openness. F2 and F3 both correlate most highly with the *Lip Corner* measure (.62 and .59 respectively). This measure is associated with rounding and is hypothesized to be related to lip protrusion; low values indicate protruded lips. Since increasing the effective vocal tract length by protruding the lips lowers the frequency of these higher formants, the correlation is attributed to this component of their variation. As these formants are also very sensitive to the location of constrictions inside the oral cavity, the strength of the correlations with this lip measure are quite striking.

IAAI CONSONANTS

The extensive consonant inventory of Iaai includes three coronal places of articulation, voiced and voiceless nasals and lateral and central approximants. Given the strong constraints operating between vowel qualities and labial consonants, it may be the case that all labial consonants include a secondary articulation of palatalization or labialization.

Palatograms of the three coronal series of stops were made of four speakers. Linguagrams were also obtained from one male speaker. For this speaker, the dental in /aṭ/ 'person' has a relatively large contact area entirely covering the upper front teeth and the alveolar ridge. The linguagram confirms that the contact is laminal, or more precisely apico-laminal [5], and includes a considerable extension of the contact laterally back toward the molar teeth on both sides. The post-alveolar ('retroflex') stop in /ḍa/ 'blood' involves a much narrower band of contact entirely behind the teeth toward the back part of the alveolar region. The linguagram shows this contact is strictly apical, with only the narrow anterior-facing surface of the tongue tip and a small area on the upper surface of the tip involved. There is markedly less lateral contact behind the front closure than for the dental, indicating that the mid part of the tongue is lower in this articulation. For the pre-palatal stop in /ca/ 'leg, foot' there is a broad contact area from the back of the alveolar ridge to a point about at the location of the second molars. The linguagram shows that the contact is strictly laminal, with no contact on the frontmost part of the tongue (about the first 1 cm).

The palatograms of the other speakers, who are of a slightly younger generation, tended to show a less clear articulatory distinction in the location of the contacts on the palate and in the tongue contact area. However, all speakers maintain a three-way acoustic distinction. Dental place is characterized by a lack of frication of the release, and if voiceless and prevocalic, by a very short voice onset time. Both the palatal and retroflex places are characterized by a noisy and sustained release. The noise of the release seems to be concentrated in the area of the third formant resonance for the palatal and the fourth formant for the retroflex. Other languages with dental and post-alveolar stops seem more frequently to have a more fricated or noisier release for the dental rather than for the post-alveolar [5], [6]. The Iaai facts show that this is a language-specific property, not a universal.

REFERENCES

[1] Ozanne-Rivierre, F. (1976), *Le Iaai.* SELAF, Paris.

[2] Ozanne-Rivierre, F. (1984), *Dictionnaire Iaai-Français.* SELAF, Paris.

[3] Goldstein, L. (1991), Lip rounding as side contact. *Proc. 12th Int. Congress of Phonetic Sciences, Aix-en-Provence, 1*: 97-101.

[4] Linker, W. (1982), *Articulatory and Acoustic Correlates of Labial Activity in Vowels: A Cross-Linguistic Study.* (= UCLA Working Papers in Phonetics 56.)

[5] Dart, S. (1991), *Articulatory and Acoustic Properties of Apical and Laminal Articulations* (= UCLA Working Papers in Phonetics 79).

[6] Anderson, V. B. and Maddieson, I. (1994), "Acoustic characteristics of Tiwi coronal stops." *UCLA Working Papers in Phonetics* 87: 131-162.

THE ROLE OF SIMILARITY IN PHONOLOGY: EXPLAINING OCP-PLACE

Stefan Frisch, Michael Broe, and Janet Pierrehumbert
Northwestern University, Evanston, IL, USA

ABSTRACT

This paper introduces an improved similarity model to account for cooccurrence restrictions in the verbal roots of Arabic, extending the results of [1]. A new quantitative measure of OCP effects and a new similarity metric are presented based on information theory. Similarity is computed by applying an entropic formulation of the cognitive "basic level" to a hierarchical representation of natural classes as in [2].

INTRODUCTION

The canonical Arabic verb form is a sequence of three consonants. Vowels provided by other morphemes are interleaved with the consonants to produce surface forms. There are strong, gradient cooccurrence restrictions among the consonants in the root. In particular, roots containing more than one consonant from one of the following classes are highly underrepresented [3].

(1) a. Labials = {b, f, m}
b. Coronal Sonorants = {l, r, n}
c. Coronal Obstruents = {t, d, T, D, θ, ð, s, z, S, Z, ʃ}
e. Dorsals = {g, k, q, χ, ʁ}
f. Gutturals = {χ, ʁ, h, ʕ, h, ʔ}

There are also gradient effects within the major classes. For example, /l/ and /r/ have a stronger restriction than /l/ and /n/.

The traditional account of the OCP effects relies on categorical cooccurrence rules. The traditional model encounters a number of problems. It cannot properly account for the gradience of OCP phenomena, both among adjacent consonants and over distance [1]. In addition, it is a negative constraint, and thus does not account for the patterns of overrepresentation presented below.

THE SIMILARITY ACCOUNT

According to [1], the degree to which the OCP is violated by two homorganic consonants in a root is a function of the perceived similarity of those two consonants. In addition, intervening consonants are interference and thus reduce the perceived similarity of more distant consonants. In this way, the gradient nature of the OCP is captured. The OCP is strongest in the case of adjacent identical consonants, which have a high degree of perceived similarity. It is weaker for identical consonants at a distance and for non-identical consonants. It is weakest for non-identical consonants which are non-adjacent. Given that the similarity account can capture the gradience of the OCP effects, it is a more empirically adequate account.

The challenge for the similarity account is to determine a function that supplies the best fit between the similarity gradient over the consonant inventory and the observed cooccurrence restrictions. In [1], similarity was computed for each consonant pair by the ratio of shared to shared plus nonshared features. Contrastive underspecification was used to capture gradient effects across classes. For small classes, like the labials and coronal sonorants, very few features are needed to differentiate sounds, which increases the value of the similarity function. For large classes like the coronals, there are many features needed to differentiate them, which reduces the value of the similarity function between members within the class.

However, the original similarity account did not capture all of the OCP effects [1]. The model failed to capture the strength of the restriction between /χ/ and /ʁ/ and the other dorsals. Also, the division within the coronal sonorants, where /l/ and /r/ form a subclass in contrast to /n/ was not captured. In addition, the use of contrastive underspecification is undesirable. It is responsible for the failure to differentiate the subclasses within the coronal sonorants. Recent work shows that the phenomena which were originally taken to support contrastive underspecification can be given a more satisfactory reanalysis in terms of privative features and licensing [5]. Finally, it is undesirable on formal grounds: contrastive underspecification is inherently derivational and logically intractable [2].

In the remainder of the paper, we first present additional evidence that the cooccurrence restrictions in Arabic are based on similarity, and not on categorical rules. We then present a new approach to the similarity function which more adequately models the data.

COMPUTING OCP EFFECTS

We have studied the Arabic root cooccurrence constraints using notions from information theory. Any consonant can be characterized by examining the quantitative extent of its cooccurrence with each of the other consonants in the system. This set of values can be represented as a cooccurrence vector, the elements of which are normalized to indicate overrepresentation or underrepresentation. The vectors of two consonants can then be compared, revealing the degree to which their cooccurrence profiles match across the entire consonant inventory. The match is quantified using information theoretic interdependence. Two consonants will have a high degree of interdependence if the cooccurrence restrictions of one consonant are predictable from the cooccurrence restrictions of the other. This can occur in two distinct ways: either the two consonants can have identical restrictions, or they can have complementary restrictions. If the restrictions are identical, the consonants pattern the same way with respect to OCP effects. If the restrictions are complementary, the consonants have opposite patterns with respect to OCP effects. When the interdependence of two consonants is low, there is no relation between the distributions of the consonants.

The computation of interdependence is based on the entropy of the pattern of cooccurrence restrictions in the system. Entropy is a measure of uncertainty of the outcome of an event. Entropy is

$$H(x) = H(p_1, \ldots, p_n) = -\sum_i p_i \log_2(p_i)$$

where p_i is the probability of x having outcome i. If all outcomes are equiprobable, then there is high uncertainty and the entropy is large. Less equiprobable outcomes result in lower entropy, as the outcome is relatively more predictable.

For a single consonant, we are interested in the uncertainty between two possible outcomes: overrepresentation or underrepresentation with respect to other consonants. For a pair of consonants there are four possible outcomes. They may be: both underrepresented, one underrepresented and the other not (for each), or they may both be overrepresented with respect to other consonants. As the correlation between cooccurrence vectors increases, the uncertainty of the joint outcome goes down. So interdependence can be expressed as:

$$J(x,y) = H(x) + H(y) - H(x,y)$$

Interdependence quantifies the degree to which entropy is shared by both consonants, an entropic measure of correlation of information. Table 1 is a sample calculation on a simplified data set, employing discrete over and underrepresentation.

Table 1: Computing interdependence of t and s. All outcomes equiprobable.

	b	d	z	ʃ
t	Over	Under	Under	Over
s	Over	Under	Under	Under

H("t") = H(0.5, 0.5) = 1
H("s") = H(0.25, 0.75) = 0.81
H("t,s") = H(0.25, 0.5, 0.25) = 1.5
J("t,s") = 1 + 0.81 - 1.5 = 0.31

Table 2 shows the interdependence computed over the entire Arabic system, based on the degree of over and underrepresentation between consonant pairs. Interdependence is normally unsigned, but a sign has been added for clarity. Positive values indicate shared cooccurrence restrictions, negative values indicate complementary restrictions.

Gray shading in table 2 indicates interdependence of at least 0.03. All of the major classes with cooccurrence restrictions have interdependence at or above this level. In addition, a cooccurrence restriction between /w/ and the labials is revealed.

The interdependence measure also reveals a pattern of overrepresentation, indicated by boxes in table 2. Labials are consistently overrepresented with the coronal obstruents; the glides /w/ and /y/ are overrepresented with the coronal obstruents; and the coronal sonorants are overrepresented with the dorsals and gutturals.

We claim these patterns of overrepresentation are another reflex of similarity. The coronal sonorants share place of ar-

Table 2: Interdependence of cooccurrence restrictions among the Arabic consonants.

	bfm	td	TD	θð	sz	SZ	ʃ	kg	q	χʁ	hʕ	h?	lr	n	wy
bfm	0.60	-0.09	-0.10	-0.03	-0.11	-0.05	-0.04	-0.01	-0.02	-0.01	0.00	-0.03	-0.04	0.00	0.13
td	-0.09	0.14	0.12	0.07	0.15	0.08	0.04	-0.02	0.00	-0.02	-0.03	0.00	-0.02	-0.01	-0.06
TD	-0.10	0.12	0.31	0.08	0.14	0.15	0.06	0.01	-0.02	0.00	-0.03	0.01	-0.04	-0.03	-0.08
θð	-0.03	0.07	0.08	0.09	0.12	0.05	0.03	0.00	0.00	-0.01	-0.03	0.00	-0.01	0.01	-0.05
sz	-0.11	0.15	0.14	0.12	0.37	0.09	0.09	-0.01	0.01	-0.04	-0.07	0.00	0.00	0.00	-0.09
SZ	-0.05	0.08	0.15	0.05	0.09	0.15	0.05	0.00	-0.01	-0.01	-0.01	0.01	-0.02	-0.01	-0.05
ʃ	-0.04	0.04	0.06	0.03	0.09	0.05	0.06	0.00	0.00	-0.02	-0.02	0.01	0.00	0.00	-0.04
kg	-0.01	-0.02	0.01	0.00	-0.01	0.00	0.00	0.48	0.08	0.05	-0.01	-0.08	-0.26	-0.17	-0.01
q	-0.02	0.00	-0.02	0.00	0.01	-0.01	0.00	0.08	0.16	0.12	-0.07	0.03	-0.13	-0.09	-0.02
χʁ	-0.01	-0.02	0.00	-0.01	-0.04	-0.01	-0.02	0.05	0.12	0.37	0.20	0.13	-0.10	-0.12	-0.01
hʕ	0.00	-0.03	-0.03	-0.03	-0.07	-0.01	-0.02	-0.01	-0.07	0.20	0.38	0.16	-0.07	-0.04	0.00
h?	-0.03	0.00	0.01	0.00	0.00	0.01	0.01	-0.08	0.03	0.13	0.16	0.25	-0.06	-0.05	-0.03
lr	-0.04	-0.02	-0.04	-0.01	0.00	-0.02	0.00	-0.26	-0.13	-0.10	-0.07	-0.06	1.00	0.36	0.01
n	0.00	-0.01	-0.03	0.01	0.00	-0.01	0.00	-0.17	-0.09	-0.12	-0.04	-0.05	0.36	0.30	-0.01
wy	0.13	-0.06	-0.08	-0.05	-0.09	-0.05	-0.04	-0.01	-0.02	-0.01	0.00	-0.03	0.01	-0.01	0.13

ticulation with the coronal obstruents, and thus are more similar to them than to the dorsals and gutturals. The glides /w/ and /y/ involve a dorsal articulation, and thus are more similar to the dorsals and gutturals than the coronals. Finally, the glide /w/ has a labial articulation, which gives some OCP effect between the glides and the labials.

The patterns of overrepresentation are especially significant, showing that the cooccurrence data cannot be accounted for solely in terms of a negative constraint. If the OCP effects were only based on a cooccurrence restriction, the unrestricted cases should be uniformly overrepresented. Instead, pairs of similar consonants are restricted, and highly dissimilar ones are more likely to cooccur. Thus, the patterns of overrepresentation provide additional evidence for the similarity account, extending its application from underrepresentation to overrepresentation. We now turn to the new similarity model, which provides an improved fit for the data.

A NEW APPROACH

The problems with the previous similarity account [1] can be remedied by adopting a heirarchical approach to natural classes and feature specification [2]. In the approach developed there, the system of contrasts in a language is a structural relation among the natural classes of the language. The natural classes, which form a partially ordered set, can be represented as a "tangled" hierarchy, or lattice.

In this model, the classification of the phonemes of a language into natural classes is very similar to other cognitive classification systems. The natural kinds of mammal, dog, and German Shepherd are related to one another in the same way that the classes coronal, coronal stop, and /t/ are. Research in cognitive science has shown that within such hierarchies, mid-level categories, like dog, are privileged with respect to superordinate or subordinate ones [6]. For example, these so called "basic level" categories are the first to be acquired by children and are more readily accessible (reflected in faster reaction times).

We claim that there is also a cognitive basic level in phonological systems, and this basic level is the most important one in determining OCP effects. We propose that the correct feature specification of the Arabic consonant system has the categories in (1) as basic level categories.

COMPUTING SIMILARITY

The function we use to compute similarity differs in two ways from [1]. First, rather than computing similarity based on individual features, we propose to compute similarity based on natural classes. Second, we use a weighting scheme to capture the primacy of basic level categories in perceived similarity. Natural classes at the basic level are weighted

Table 3: Computation of similarity of the Arabic consonants.

	bfm	td	TD	θð	sz	SZ	ʃ	kg	q	χʁ	hʕ	h?	lr	n	wy
bfm	0.49	0.04	0.04	0.03	0.03	0.04	0.04	0.04	0.03	0.05	0.05	0.05	0.07	0.10	0.19
td	0.04	0.68	0.33	0.28	0.27	0.13	0.25	0.34	0.35	0.06	0.01	0.01	0.05	0.05	0.03
TD	0.04	0.33	0.69	0.13	0.12	0.30	0.12	0.17	0.17	0.01	0.08	0.08	0.07	0.07	0.04
θð	0.03	0.28	0.13	0.68	0.62	0.25	0.59	0.14	0.14	0.17	0.05	0.05	0.05	0.05	0.03
sz	0.03	0.27	0.12	0.62	0.69	0.24	0.63	0.13	0.14	0.16	0.05	0.05	0.05	0.05	0.03
SZ	0.04	0.13	0.30	0.25	0.24	0.68	0.24	0.05	0.06	0.06	0.22	0.22	0.07	0.07	0.04
ʃ	0.04	0.28	0.12	0.59	0.63	0.24	1.00	0.14	0.22	0.17	0.05	0.05	0.02	0.02	0.01
kg	0.04	0.34	0.17	0.14	0.13	0.05	0.14	0.68	0.66	0.18	0.06	0.06	0.01	0.01	0.07
q	0.03	0.35	0.17	0.14	0.14	0.06	0.22	0.66	1.00	0.19	0.06	0.06	0.00	0.00	0.03
χʁ	0.05	0.06	0.01	0.17	0.16	0.06	0.17	0.18	0.19	0.67	0.23	0.23	0.01	0.01	0.09
hʕ	0.05	0.01	0.08	0.05	0.05	0.22	0.05	0.06	0.06	0.23	0.68	0.62	0.01	0.01	0.09
h?	0.05	0.01	0.08	0.05	0.05	0.22	0.05	0.06	0.06	0.23	0.62	0.68	0.01	0.01	0.09
lr	0.07	0.05	0.07	0.05	0.05	0.07	0.02	0.01	0.00	0.01	0.01	0.01	1.00	0.76	0.34
n	0.10	0.05	0.07	0.05	0.05	0.07	0.02	0.01	0.00	0.01	0.01	0.01	0.76	1.00	0.35
wy	0.19	0.03	0.04	0.03	0.03	0.04	0.01	0.07	0.03	0.09	0.09	0.09	0.34	0.35	0.70

higher than those which are above or below it. The weighting function is also entropic and based on the optimization of information balance inside and outside of the category [7].

Table 3 shows the results of one similarity computation. Shading indicates homorganic consonant pairs with similarity greater than 0.1. All of the major classes are modeled at this level of similarity. The correct patterning of /χ/ and /ʁ/ with both the gutturals and the velars is captured, and the subclassification of the coronal sonorants is also obtained. In addition, the significant overrepresentation shown in table 2 is accounted for by pairs with very low similarity. The boxed regions in table 3 show similarity below 0.05.

Weighting replaces the use of contrastive underspecification; the higher weighting of basic level categories compensates for the additional noncontrastive features that might increase the perceived differences within categories. We are still exploring the proper combination of feature assignments and weighting functions in order to find the best fit to the data. The empirical advantage of this approach obtains regardless of the particular function used.

CONCLUSION

OCP-Place is a phonological reflex of a cognitive universal: similarity. The pattern of cooccurrence restrictions across the lexicon of Arabic reflects both cooccurrence restrictions between similar consonants and an overrepresentation of highly dissimilar consonants. Perceived similarity between two consonants is a function of the natural classes in which those consonants are found which are weighted based on their proximity to the basic level.

ACKNOWLEDGMENT

This work was supported in part by NSF Grant No. BNS-9022484.

REFERENCES

[1] Pierrehumbert, J. (1992), "Dissimilarity in the Arabic verbal roots", *Proceedings of NELS 23*, Amherst: GSLA.

[2] Broe, M. (1993), *Specification theory: the treatment of redundancy in generative phonology*, Ph.D. dissertation, Edinburgh.

[3] McCarthy, J.J. (1994), "Guttural phonology", in *Papers in laboratory phonology III*. Cambridge: Cambridge University Press.

[5] Steriade, Donca. (1994), "Underspecification and markedness", in *Handbook of Phonology*, J. Goldsmith, ed. Oxford: Basil Blackwell.

[6] Rosch, E. et al. (1976), "Basic objects in natural categories", *Cognitive Psychology* 8: 382-439.

[7] Broe, M. (1995), "An entropic measure of category utility", Ms., Northwestern University.

Syllabification of Intervocalic Consonants in Dutch: Single Consonants or Geminates?

*J. Verhoeven, S. Gillis * and G. De Schutter*
University of Antwerp - U.I.A., Antwerp, Belgium

ABSTRACT

Single intervocalic stops in Dutch are analysed phonologically as geminates when preceded by a short vowel, and as single consonants when preceded by a long vowel. We investigate the phonetic correlates of this phonological distinction. Measurements of consonant durations show no significant difference relative to the preceding vowel, and hence, the underlying phonological distinction between geminates and single consonants appears to be neutralised.

1. INTRODUCTION

Single intervocalic consonants in Dutch are considered to be tautosyllabic with the following vowel or ambisyllabic depending on the nature of the preceding vowel. After a lax vowel the ambisyllabicity condition holds ($VC_1.C_1V$), while a consonant following a tense vowel is syllabified with the following vowel ($V.C_1V$). This analysis is generally accepted in the phonological literature (see Kager's [1] state-of-the-art overview). It is argued that the syllable in Dutch is minimally and maximally bimoraic which means that a long vowel, a diphthong and a sequence of a short vowel plus a consonant are legal syllables. It also follows from this *bimoraic constraint* that an unchecked short vowel cannot be syllable final. In other words, short vowels are restricted to preconsonantal positions in which 'close contact' exists between the vowel and the following consonant. Hence, a single intervocalic consonant after a short vowel is analysed as ambisyllabic (see Figure 1 bottom) while after a long vowel the consonant is syllabified in the following syllable (see Figure 1 top).

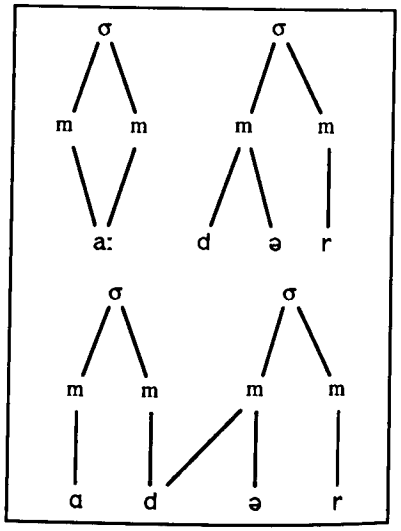


Figure 1. Syllable structure of 'ader' /a:dər/ and 'adder' /ɑdər/ (σ = syllable, m = mora).

There is considerable disagreement

concerning the phonetic correlates of the phonological distinction between geminates (as in /ɑdər/) and single consonants (as in /a:dər/). On the one hand, measurements of consonant duration in Dutch words by Nooteboom [2] reveals that ambisyllabic consonants following a short stressed vowel are significantly longer than tautosyllabic consonants following a stressed long vowel. On the other hand, Jongman & Sereno [3] and Kuijpers [4] found no durational differences of the intervocalic consonants in this environment.

Nooteboom's findings clearly agree with the phonological analysis in the sense that the underlying phonological distinction between geminate and single consonants is reflected at the phonetic level. Failure to detect phonetic differences between those consonants indicates (contextual) neutralisation, i.e., the identical realisation of distinct phonological segments.

In this paper we revisit this controversy. We investigate the duration of intervocalic stops in disyllabic words as a function of the quantity of the preceding vowel.

2. METHOD

A controlled production experiment was carried out in which informants produced 15 minimal disyllabic word pairs. These pairs of existing Dutch lexemes were chose in such a way that all the oral and nasal stops were represented. They contained long - short vowel pairs that have minimal spectral differences in the first syllable. The rhyme of the second syllable consisted of /ə/ followed by /r/ or /l/. Due to difficulties in finding suitable word pairs, 3 pairs ended in a vowel. Words ending in /ən/ were avoided because the final consonant is often deleted.

The subjects, 5 male and 5 female native speakers of Dutch produced each word in a standard carrier sentence. The word was presented on a computer monitor in a large font in a particular colour. 6 different colours were used, and the subject had to insert the word and the colour in which it appeared on the screen in the Dutch equivalent of the sentence 'The colour of ... is ... '. Each subject completed the task three times and during each run, the subjects were presented with a different randomisation of the target word list mixed with 40 distracter words.

Subjects' deliveries were recorded by means of a Sennheiser Microphone MKE 66 and a Sony Digital Audio Tape-recorder TCD-D3. Recordings were digitised (Fs = 16.000 Hz, Fc = 8.000 Hz) on a an Apple Quadra 700 by means of a Digidesign Sound Designer II signal processing card and Audiomedia software. Resulting Audiofiles were further processed in Signalize™.

Measurements were made of the duration of the intervocalic stop. In order to establish the duration of a segment, the waveform was used for voiceless stops only. The total duration of the stop was measured as the silence during the occlusion and the release burst.

Measurements of voiced segments were based on a time-aligned wideband (125 Hz) spectrogram.

3. RESULTS

In the first instance, the durations of the intervocalic oral and nasal stops were compared by means of an ANOVA. The results of this analysis are summarised in Figure 2. This analysis turns out to be highly significant ($F(11, 853) = 86.7828$, $p < .0001$). Subsequently, a Student's t-test was carried out between each relevant pair of geminate vs. single consonant, e.g. /b - bb/, /t-tt/ etc. None of these comparisons turns out to be significant.

Next, intervocalic consonant length was compared in terms of place of articulation by means of an ANOVA. This analysis suggests a significant relationship between place of articulation and stop duration ($F(2, 862) = 38.4813$, $p < .0001$). In addition, pairwise comparisons were made between the stop durations for the different places of articulation by means of a Student's t-test. Each of these comparisons turns out to be significant : labial-alveolar ($t = 4.1233$, d.f. = 803, $p < 0.001$), labial-velar ($t = 10.5903$, d.f. = 415, $p < 0.001$), alveolar-velar ($t = 5.8854$, d.f. = 506, $p < 0.001$).

4. DISCUSSION

The aim of this investigation was to investigate whether significant durational differences could be found between underlying geminate and single intervocalic stops. The results presented above clearly indicate that no significant differences exist between stop segments that are regarded distinct in a phonological perspective. These results are in agreement Jongman & Sereno [3] and Kuijpers [4]. This absence of a phonetic difference suggests that we are dealing here with an instance of (contextual) neutralisation.

The results of this investigation indicate that no significant durational differences were found between these two types of segments. However, the analysis of stop durations in terms of their respective place of articulation reveals a striking difference in that velar stops are longer than alveolars, which in turn are longer than labials. The relevance of this finding are discussed in Verhoeven, Gillis & De Schutter [6].

These findings are important from a developmental perspective. Gillis & De Schutter [5] found that children's intuitive syllabifications do not reflect the phonological distinction between geminates (after short vowels) and single consonants (after long vowels). At least there was no trace of that distinction in 5-year-olds' syllabifications. However 8-year-olds do syllabify a word like 'adder' as 'ad.der' in more than 50% of the cases. It was hypothesised that the older children's familiarity with the spelling conventions of the language (esp. the splitting of words: 'adder' is split graphemically as 'ad-der') was instrumental in their intuitive syllabifications. The younger

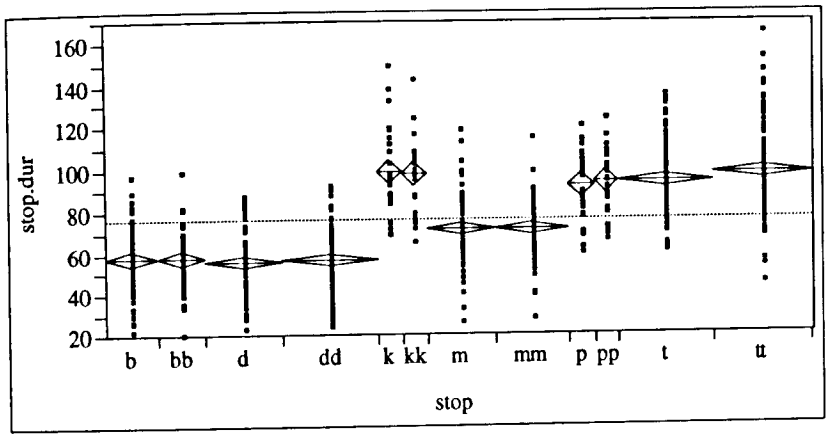


Figure 2. Comparisons of consonant durations for geminate and single consonants.

children who are not familiar with the written code have no basis for that syllabification. The present findings strengthen this hypothesis in the sense that the acoustic signal does not support the distinction between geminates and single consonants.

ACKNOWLEDGEMENT

This research was supported by the Fund for Joint Basic Research (grant 2.0101.94) of the Belgian National Science Foundation.

REFERENCES

[1] Kager, R. (1989), *A metrical theory of stress and destressing in English and Dutch*. Dordrecht: Foris.

[2] Nooteboom, S. (1972), *Production and perception of vowel duration*. Eindhoven: Philips Research Reports Supplements No. 5.

[3] Jongman, A. & Sereno, J. (1991), "On vowel quantity and post-vocalic consonant duration in Dutch.", *Proceedings of the XIIthe International Congress of Phonetic Sciences*, Vol. 2, pp. 294-297.

[4] Kuijpers, C. (1993), *Temporal coordination in speech development*. Unpublished Dissertation, University of Amsterdam.

[5] Gillis, S. & De Schutter, G. (1995), "Intuitive syllabification: Universals and language specific constraints." Submitted.

[6] Verhoeven, J., Gillis, S. & De Schutter, G. (1995). "Intervocalic geminate and single stops in Dutch syllable structure". Submitted.

* Research Fellow of the Belgian National Science Foundation.

THE CENTER OR EDGE: HOW ARE CONSONANT CLUSTERS ORGANIZED WITH RESPECT TO THE VOWEL?

Douglas N. Honorof† and Catherine P. Browman**
*†Yale University and *Haskins Laboratories, New Haven, Connecticut, U.S.A.*

ABSTRACT

Stable inter-gestural timing patterns were sought for phonotactically-permissible (CC)CVX and XVC(CC) accented monosyllables in American English. Movement evidence for four speakers confirmed the hypotheses of Browman and Goldstein [1] that a pre-vocalic consonant or cluster is organized with respect to a 'tautosyllabic' nuclear vowel by its center (*i.e.*, C-center), but a post-vocalic consonant (or sequence of consonants) by its (first) left edge.

INTRODUCTION

Having examined x-ray microbeam data for one speaker of American English, Browman and Goldstein found evidence for the *C-center* (defined below) [1]. Specifically, they argued that, judging by patterns of articulatory stability, the C-center of a *pre-vocalic* consonant or consonant cluster is more tightly coordinated with the vowel gesture that corresponds to a following acoustic vowel than is either the *left edge* (henceforth, *LE*) of the first pre-vocalic consonant plateau or the *right edge* (henceforth, *RE*) of the last one. However, at least for the monosyllabic target words in their data set, they suggested that it is the LE of the first *post-vocalic* consonant rather than the C-center of the whole sequence of consonants (or the C-center of just the coda consonants) that is most tightly coordinated with the vowel gesture, regardless of whether that vowel and consonant are separated by a word boundary. (See Fig. 1). They did report, however, that those vowel-to-LE measures are even more stable when there is no intervening word boundary (but see [2] for apparent counter-evidence.)

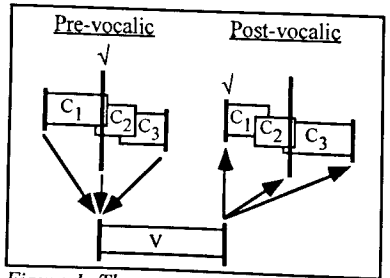


Figure 1: The potential phasing relations considered in the present experiment are indicated schematically with bold vertical lines and arrows. Those phasing relations that are argued for in [1] are indicated with check marks.

One implication of this picture of organization is that increasing the number of consonant gestures in a coda should not reduce the acoustic duration of a 'tautosyllabic' vowel. This implication is at odds with the notion of 'compensatory shortening'. (See [3].)

The experimental results that we report here allow us to address these articulatory and acoustic issues.

METHODS

Design and Stimuli

The present design systematically varies one-, two-, and three-consonant pre-vocalic and post-vocalic consonant 'clusters' in accented, monosyllabic (real and nonsense) English target words. The utterances were designed to disallow rightward re-syllabification on phonotactic grounds. (See Table 1.)

Data Collection Technique and Procedure

All data were collected at the University of Wisconsin x-ray microbeam facility [4]. The utterances were presented to the subjects on a video screen in quasi-random order. The microbeam system then tracked the Cartesian coordinates of gold pellets (2.5-3.0 mm in diameter) affixed to the mid-line surfaces of the subject's articulators as he or she read each utterance aloud at least five times, while acoustic data were simultaneously recorded. Before analyzing the articulatory data, we compensated for any head movement added to articulator movement by using position data

Table 1: List of utterances. Capital letters indicate accent.

Frame: 'I read [past] _____ again.'

	Position of Target Cs	
Target Cs	*Pre-vocalic*	*Post-vocalic*
[s]	cuff SAYED	CUSS fade
[p]	cuff PAID	CUP fade
[ps]	————	CUPS fade
[sp]	cuff SPAYED	CUSP fade
[sps]	————	CUSPS fade
[l]	cuff LAID	————
[lp]	————	CULP fade
[lps]	————	CULPS fade
[pl]	cuff PLAYED	————
[spl]	cuff SPLAYED	————

gathered on reference pellets affixed to the nose and upper incisor. The data were automatically rotated to the occlusal plane.

Subjects

Four college students participated in the present study. All were natives of Wisconsin between the ages of 18 and 20, three female, one male. All were of normal speech and hearing ability, with the exception of Subject STR2 who had a 30 dB notch at 8 kHz in the right ear only, which we do not believe to have affected his performance on the task.

Measurement Procedure

The sampling rates for the x-ray data differed from pellet to pellet (40-160 Hz), and sometimes for a single pellet from utterance to utterance. Therefore, when smoothing articulatory data, we set the number of points in our (software) triangular filters according to channel-specific sampling rates so that window sizes were always brought as close as possible to 25 msec. The x-ray data were then re-framed by interpolation to 200 Hz (T=5 msec/frame).

In all cases we followed Browman and Goldstein in defining consonant and vowel gestures on the oral tier only [1] (see also [5]). Wherever possible, we labeled the relevant articulatory trajectory for each target consonant gesture by automatic algorithm, first finding the relevant extremum, then marking as the LE the first frame whose displacement amplitude fell within a spatial noise level (SNL) equal to five percent of the mean of the speaker's mean range of displacement across all displacement trajectories analyzed for the present experiment (including those from utterances later excluded from analysis). We marked the right edge (RE) as the last frame whose amplitude fell within that same SNL.

Lip Aperture

In order to label 'p' in a principled way, we computed a trajectory that we call *lip aperture* (henceforth, *LA*) by subtracting vertical displacement of the lower lip pellet from that of the upper lip pellet (positioned on the lower and upper vermilion borders, respectively). Thus, a minima or *valley* in LA presumably corresponds to attainment of mid-line closure of the lips.

We had hoped to measure 'p' as the LE and RE of a basin around that valley. Indeed, in general this is what we did (though for one token we were only able to do so by reducing the SNL by .1 mm) However, in many cases, edges of contiguous 'p#f' and 'f#p' sequences were 'blurred' in LA, perhaps due to conflicting demands being made on the lower lip by the two closure gestures. Although the failure of the articulators to return consistently to their 'neutral' positions between these two contiguous gestures presents no theoretical difficulties, in cases where labial gestures were contiguous, we were forced to label 'p' by automatically picking the 'p' edge that was not contiguous with 'f', and then calculating the contiguous edge in terms of an utterance-type-specific ratio of the known edge to the relevant anchor point. These ratios were calculated on the basis of various averages of the ratios for non-contiguous edges to relevant anchor points for the tokens whose 'f'-contiguous labels were found automatically ('f#sp', 'f#spl', 'sps#f', 'ps#f' and 'lps#f').

Tongue Tip Constriction Degree

Due to the non-zero slope of the hard palate in the region where alveolar consonants are articulated, peaks in mid-line vertical displacement of the tongue tip pellet (positioned 7 to 9 mm back of the actual tip) do not always co-occur in time with the actual moment of tightest constriction as measured for that pellet. Therefore we measured the relative

amplitude of tongue tip displacement for 's' in a trajectory that we call (inverted) *tongue tip constriction degree* (henceforth, *TTCD),* which is simply the x and y coordinates of the tongue tip pellet rotated to the slope of a relevant segment of a mid-line pellet tracing of the palate by the formula:

$$TTCD = -TTX * SIN\theta + TTY * COS\theta$$

where θ is the slope of the palate segment found by linear regression.

We also labeled 'l' in TTCD. Our justification for doing so is found in Sproat and Fujimura's suggestion that the tongue tip gesture for 'l' may be consonantal, but the tongue body gesture, vocalic [6]. Because the present research concerns the timing of consonant gestures with respect to vowel gestures, we ignored the movement of the tongue body for 'l'.

Right Anchor Point

We labeled the trans-vocalic *right anchor point* (henceforth, *RAP, i.e.*, the attainment of target for post-vocalic 'd'—see [1]) by automatic algorithm, choosing the first positive-to-negative zero crossing in the region of interest of the first derivative of smoothed TTCD (SNL=10% of the subject's mean velocity range for that pellet across utterances). Our algorithm required two consecutive frames to be within the zero region, and none without, before the edges of the zero region were declared.

Left Anchor Point

We used a trans-vocalic *left anchor point* (henceforth *LAP*) similar to that used in [1]. That is, we identified absolute peak vertical displacement of a pellet affixed to the tongue dorsum (58 to 64 mm back of tongue tip). This measure is effectively equivalent to the C-center of the singleton onset ('k'). The slope of the soft palate remains nearly horizontal in the region where tongue dorsum constrictions are articulated for all four subjects, so we found it unnecessary to rotate the tongue dorsum data before labeling the LAP.

C-centers

Again, following [1], we computed the C-center as the mean of the temporal midpoints of the plateaus (or basins) surrounding the spatial peaks (or valleys) associated with the consonant gestures in a cluster. However, we were not always able to distinguish the neighboring edges of 's' and 'l' in TTCD when both occurred in one cluster, *i.e.*, 'spl' and 'lps'. Nor were we always able to distinguish the RE of the first 's' from the LE of the second 's' in 'sps' clusters. Therefore, in such cases, we chose to label only the LE of the first TTCD gesture and the RE of the second. We then counted the resulting 'plateau' as a single gesture for the purposes of computing the C-center. (However, we chose not to analyze the post-vocalic 'sp(C)' and 'lp(C)' utterances of Subject SCH2, who appeared to have had difficulty producing many of them under the experimental conditions.)

For an 'f#p' or 'p#f' contiguous utterance, we computed the C-center with reference to the normalized 'p' edge as discussed above.

Acoustic Vowel Duration

We also measured the acoustic duration of [ʌ] before post-vocalic target consonants. To this end we segmented the waveform by placing two labels without particular reference to articulatory labels, one at the start of aspiration following 'k' and another at the instant of gross spectral change corresponding to the first following post-vocalic consonant. However, we were not always able to identify discrete acoustic boundaries between the vowel and post-vocalic 'l'.

RESULTS AND DISCUSSION

For the articulatory data, separate ANOVA were run for each subject for each position (pre-vocalic and post-vocalic) on one factor with three levels: RE, LE, and C-center.

For the pre-vocalic consonants, in each case the measure for each of the three levels was subtracted from the RAP, as in [1]. For all subjects, the pre-vocalic C-center organization was more stable (*i.e.*, had a smaller standard deviation from the group mean with a lower Levene's p-score) than either of the other measures, as was found in [1]. By subject, $F(2,87)=7.81$, Levene's $p<.01$; $F(2,87)=4.08$, Levene's $p<.03$; $F(2,81)=9.67$, Levene's $p<.01$; $F(2,84)=10.09$, Levene's $p<.01$.

For the post-vocalic consonants, the measure for each of the three levels had the measure for the LAP subtracted from it, as in [1]. This time, for all four subjects, the local LE organization proved more stable than the C-center or RE, again as the analysis of data from the single subject in [1] suggested. By subject, $F(2,99)=27.03$, Levene's $p<.0001$; $F(2,42)=6.04$, Levene's $p<.01$; $F(2,93)=17.56$, Levene's $p<.0001$; $F(2,102)=19.41$, Levene's $p<.0001$.

The acoustic vowel duration for [ʌ] in the accented syllable did not show a significant or consistent trend toward increasing or decreasing with cluster complexity; Levene's p-values ranged across speakers from >.17 to >.89. While this finding would not be easily explained by 'compensatory shortening' (see [3]), it is expected given the results of our articulatory analyses in which the LE was most stable.

We interpret the pre- and post-vocalic results as strongly supporting the suggestion in [1] that there is a difference in pre- and post- vocalic organization (in American English monosyllabic words[1]), at least for labial consonant gestures and consonant sequences involving labial gestures. Nevertheless we refrain from drawing rigid conclusions until our findings can be confirmed for other constriction locations, and until comparable results obtained from point-source tracking and linguapalatal devices have been scrutinized, which we hope to do in the future.

ACKNOWLEDGEMENTS

We would like to thank Dani Byrd for her helpful comments. The present study was supported by NIH Grant HD-01994 and NIH Grant DC-00121 to Haskins Laboratories.

REFERENCES

[1] Browman, C. P. & Goldstein, L. (1988), "Some notes on syllable structure in articulatory phonology", *Phonetica,* vol. 45, pp. 140-155.

[2] Byrd, D. (in press), "C-centers revisited", *Phonetica.*

[3] Munhall, K., Fowler, C., Hawkins, S. & Saltzman, E. (1992), "'Compensatory shortening' in monosyllables of spoken English", *Journal of Phonetics,* vol. 20, pp. 225-239.

[4] Nadler, R. D., Abbs, J. H. & Fujimura, O. (1987), "Speech movement research using the new x-ray microbeam system", *Proceedings of the XIth International Congress of Phonetic Sciences,* vol. 1, Tallinn, Estonia: Academy of Sciences of the Estonian S.S.R. Institute of Language and Literature, pp. 221-224.

[5] Fujimura, O. & Lovins, J. B. (1978), "Syllables as concatenative phonetic units", in Bell, A. & J. B. Hooper (Eds.): *Syllables and Segments,* New York: North-Holland Pub. Co., pp. 107-120.

[6] Sproat, R. & Fujimura, O. (1993), "Allophonic variation in English /l/ and its implications for phonetic implementation", *Journal of Phonetics,* vol. 21, pp. 291-311.

[1] For a discussion of how these claims relate to P-centers, weight units and moraic structure, phonetic affixes and extrasyllabicity, compensatory lengthening, allophonic variation, and issues of universality, see [1].

MANDIBLE AS SYLLABLE ORGANIZER

N. Rhardissse and C. Abry
Institut de la Communication Parlée, INPG/Université Stendhal
BP 25 F-38040 Grenoble Cedex 9

ABSTRACT

A correlation between adherence of consonants to vowel nucleus and mandibular height has been proposed by Lindblom and colleagues [1-4]. To support this proposal [i,a] and [s,l] in VC(:)V items were recorded for 2 French and 2 Moroccan subjects. Results indicate a mandibular order increasing from low to high as follows [a]<[l]≤[i]<[s], and a perfect overall correlation between normalized *relative* coarticulability (σ/m) and mean height of the jaw.

1. INTRODUCTION

In his proposals arguing for economy of speech gestures, Lindblom [1] drew a hypothesis to explain the formation of complex syllables from VCV behaviour. Consonant propensity to cluster in syllables could depend on their jaw height that determines coarticulatory compatibility: what we call *coarticulability*. This relationship corresponds to the intuition that consonants are more assimilated by vowels than the reverse. At the same time it explains both propensity for consonants to settle more or less further apart from the vowel within the syllable (like *s* in *straight*) and their propensity to coarticulate maximally (*relatively*) enough when the same segments occur close to the vowel (in *sane*).

Results from Swedish [1] were reinterpreted by [2]. On the basis of English data, she pointed out that vowels *and some consonants* adopt jaw height to accommodate other consonants, typically [s] (to support an aerodynamic rationale for this behaviour, see [5]). Finally data from Swedish and English were examined [3] : jaw height measurements, depicted in percentage of maximum opening relative to clench (for absolute values, see [4]), are displayed on Fig. 1 for [f,b,t,d,s,n,l,r,k,h] realized in [a-a], [e-e] and [i-i] contexts. An overall correlation (r=.80) is clearly visible between height and coarticulability ranks. The latter is expressed by the coefficient of variability (σ/m), which compensates better for the fact that overall absolute variance for high segments like [s] is smaller than for low segments, say [a] (the «very high and invariant [*sic*]» jaw position, claimed for [s,ʃ] in Palestinian Arabic and French by [5] in token-to-token measurements, is not contradictory, since σ/m is not taken into account). In other words σ/m captures the overall accommodation of the consonant to the vowel in the opening scale.

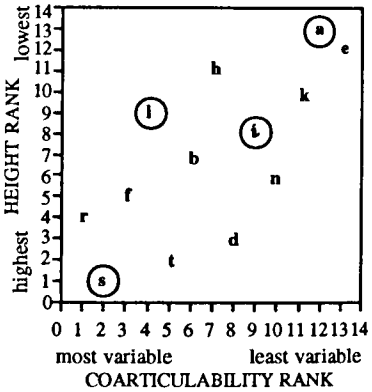


Fig.1– Ranks of mean height (m) and coarticulability (σ/m) for consonants and vowels (from % jaw opening, see text). English and Swedish combined (adapted from [3]). Test segments chosen hereafter are within circles.

In this investigation we will use Moroccan Arabic and French data to support Lindblom's hypothesis, portrayed in the frame of the [a]–[s] height scale, with a particular emphasis on the «meeting-point» of the nearest consonant to the highest vowel. Hence the test segments we chose according to previous results were [i,a] for extreme vowel magnitude in height, and [s,l] for the consonants, [l] being the closest consonant to [i] (following [4], opening values are in mm: [l]=5.50, [b]=4.86 and [i]=5.29). Arguments for comparing [l] and [i] or [j] heights can be taken from their proximity in the different proposals of sonority scales by phonologists [6, 7] and from the frequent confusion of [l] and [j] in language acquisition [8,9]. As concerns coarticulability, we will not consider the most coarticulable «guttural» [5] consonants [ʔ,h,ħ,ʁ,χ] and [q,k]. A recent study of jaw movements in 6 degrees of freedom [10] reports an interesting case of *pure translation* in a [li] syllable, with no rotation component as for [s] (and [ʃ,r,t,k,p,f]). But [l] is clearly among the consonants that resist to jaw lowering in the [a] context. Hence we will show that the correlation between coarticulability and height holds where it is not trivial, i.e. for consonants who display some coarticulation resistance of their own to vowel opening.

2. METHOD

To show up [l] and [s] jaw heights, in contact with high [i] and low [a], and their mutual coarticulation resistance, three factors were manipulated: consonantal gemination (simple *vs.* double consonants), rate (conversational *vs.* fast) and context ([i-a],[a-i],[a-a],[i-i], for consonants,[-l-],[-ll-],[-s-],[-ss-], for vowels). Arabic stimuli (9 words, 7 logatemes) were inserted in the carrier interrogative sentences [a:l —] (he says). French sequences are preceded by first name «Al» and contained significant combinations to obtain comparable sequences. Arabic stimuli were preceded by a little glottalization. Note that Moroccan Arabic gemination is tautomorphemic and French heteromorphemic.

Two native speakers of Fez Arabic (sisters A and N, present author) and two French (C woman and F man) recorded sequences in an anechoic room, in random order, producing 12 repetitions by item at two speaking rates (conversational and fast).

The tracking system was a mandibular kinesiograph (Myotronics K5AR) with a magnet fixed to the lower incisors, moving in the linear portion of the kinesiograph [11]. Vertical displacement was recorded on an FM tape. Analyses are based on 9 (out of 12) correctly produced utterances.

Jaw vertical position and audio signals were digitized in stereo at 8 kHz with Audiomedia (Macintosh). Then edited and measured using Signalyze™. Mandibular signal was undersampled at 500 Hz. Numerical signal values were converted in mm (owing to the calibration carried out by bite blocks when recording). Jaw height was measured for each segment at about its acoustic centre, relative to the minimum minimorum mandibular opening of each subject. A total of 3,456 values were thus obtained.

3. RESULTS

We will examine here only a global description for each subject, based on mean values and standard deviation to portray his/her height and coarticulability scales, without regrouping segments on the basis of the results of ANOVA for main effects (gemination, rate, context). This in order to compare our results with the overall correlation given in Fig. 1 by [3]. For all subjects mean values and standard deviations were normalized which gives us the general trend for Arabic and French. Our results will be annotated progressively in relation to English and Swedish ones.

3.1. Height and coarticulability: individual strategies

	a	l	i	s
A	7.17 *0.150*	4.62 *0.227*	3.74 *0.172*	2.50 *0.262*
N	11.96 *0.159*	9.32 *0.122*	7.51 *0.177*	5.05 *0.188*
C	10.33 *0.198*	9.18 *0.136*	6.73 *0.308*	2.45 *0.475*
F	8.95 *0.231*	6.18 *0.251*	5.86 *0.233*	3.56 *0.333*

Table 1– Mean jaw opening (mm) and *coefficient of variability* (σ/m) for test consonants and vowels. Arabic (A and N) and French (C and F) subjects.

Table 1 shows that mean jaw opening range is clearly subject-dependent (from 4.67 mm for A to 7.88 mm for C), but that on the mandibular height scale all subjects get the same ranking increasing from low to high: [a]<[l]<[i]<[s]. Thus this order is the same for Arabic, French and Swedish. Separate results taken from [3] show that velarized English [ɬ] is slightly higher than [i] (i.e. in mm: [a]=9.25, [ɬ]=4.53, [i]=5.33, [s]=2.50). In summary, an ordering [a]<[l]≤[i]<[s] corresponds to the proximity of [l] et [i], in Arabic (for A more than for N) and in French (only for F).

Coarticulability displays two different behaviours, which are not language-dependent since N (Arabic) and C (French) display the same ranking *vs.* A and F (Fig. 2), who show the same pattern as English and Swedish combined (Fig. 1). In fact the overall correlation found by [3] is fairly well reproduced, since segments are set not farther from one rank off the positive diagonal (r=0.80). [s] has the most stable coarticulability rank *vs.* [l], which changes most, [i] and [a] being in between.

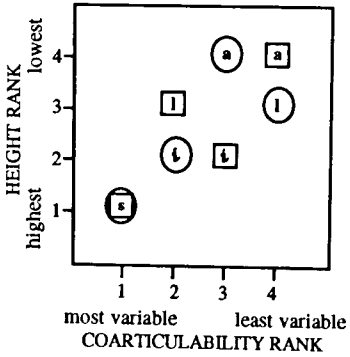


Fig.2– Height and coarticulability ranks for subjects A and F (squares) and N and C (circles).

3.2. Height and coarticulability: general trend

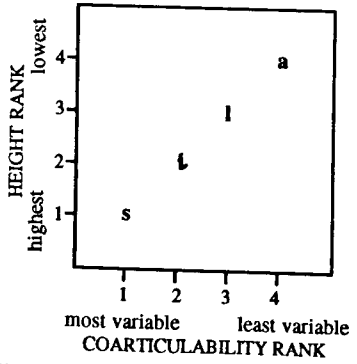


Fig.3– Ranks of normalized (see text) height and coarticulability for the four test segments. Arabic and French combined.

To get the general behaviour of the four test segments [a,l,i,s], we normalized mean height for each speaker with the following formula: $m_{norm}\% = 100(m - m_{min})/(m_{max} - m_{min})$. σ/m is normalized in two steps : $\sigma_{norm}\% = \sigma/(m_{max} - m_{min})$; then $\sigma_{norm}\%/m_{norm}\%$. (Note that the convention of an origin normalized at 0% for [s] gives corresponding coefficients of variability tending towards ∞).
The means $m_{norm}\%$ and $\sigma_{norm}\%/m_{norm}\%$ across the four subjects are displayed on Fig. 3, showing a perfect correlation between normalized height and coarticulability for both languages combined. Our results support Lindblom's hypothesis, for whom scales of mandibular position and variability can be used as indexes to account for the economy of the syllable.

4. DISCUSSION

Jaw height scale, [a]<[l]≤[i]<[s], resists to language and speaker difference – and, just to mention results to be published, resists also to metrical ([l] *vs.* [ll]; [s] *vs.* [ss]) and speech rate changes.
The correlation between mandibular height and coarticulability scales, that globally holds for speakers of languages as dissimilar as Arabic, English, French and Swedish can meet optimization principles. In motor control for large and small targets (like Fitt's law). But also in the articulatory-to-acoustics mapping [12], knowing that acoustics is fairly well related to log-area ratios [13], a sensitivity in which the mandible takes part, synergetically, even if it is not the end effector of the constriction [5].
More generally, such an approach could renew the question of the constituency of the syllable, provided that it is reconsidered at the control level for the production of the opening-closing modulation of the basic cycle of speech, for which the mandible is the carrier articulator.
In this frame, questions addressed by phonologists concerning the sonority scales and syllable structure could be reformulated. So the highest mandibular position for [s], makes [s]+plosive clusters not so weird. Processed both by Germanic metric [14] and French pig Latin (*verlan* [15]), they are too recurrent to be confined to borrowings from elite languages [16]. Thus, in mandibular height terms they don't need extrasyllabicity [7] or contextual processing by harmony phonology [17].

As concerns the proximity of [l] and [i] or [j], one can ask whether there is some empirical rationale to such a refined scale as the one proposed for Meknes Arabic [18]: [a]<[j]<[r]<[ʃ]<[l,n,w]<[d,s]<[t]; knowing that [l] and [i] are so closed in the mandibular scale (not to speak of other sonority proposals from De Brosses [16] to [6], and not to mention acquisition data again [8,9]). More necessary: the traditional paucity in degrees in the vowel sonority scale has been repaired up to a point following [19] in Berber.
Of course one could state, with Ohala that: "'Sonority' [does] not exist" [16]; and of course our interest for proposals [1] that consider the mandible as a *syllable organizer* accounts for the modulation principle in speech. But before giving up phonological constructs, more knowledge has to be gathered on the dynamic behaviour of the jaw, at least in our test segments, especially the nearest ones, [i] and [l], following emerging pioneering work on the control of *the* carrier articulator [10].

Acknowledgement: To T. Lallouache, R. Kies and A. Arnal, for the time they took to assess our signals ; to C. Smith and C. Stoel-Gammon, for their relevant references; to A. Diana for English.

5. REFERENCES

[1] Lindblom, B. (1983), "Economy of speech gestures", In P. F. MacNeilage (Ed.), *The production of speech*, (pp. 217-245), N.-Y.-Heidelberg: Springer Verlag.
[2] Keating, P. (1983), "Comments on the jaw and syllable structure", *J. Phonetics*, vol. 11, pp. 401-406.
[3] Keating, P., Lindblom, B., Lubker, J., & Kreiman, J. (1990), "Jaw position in English and Swedish VCVs", *U.C.L.A. W.P.*, vol. 74, pp. 77-95.
[4] Keating, P., Lindblom, B. Lubker, J., & Kreiman, J. (1994), "Variability in jaw height for segments in English and Swedish VCVs", *J. Phonetics*, vol. 22, pp. 407-422.
[5] Lee, S., Beckman, M., & Jackson, M. (1994), "Jaw targets for strident fricatives", *I.C.S.L.P.*, pp. 37-40.
[6] Van Coetsem, F. (1979), "The features 'vocalic' and 'syllabic'", in I. Rauch and G. F. Carr (Eds.), *Linguistic Method: Essays in Honor of Herbert Penzl*, (pp. 547-556), The Hague: Mouton.
[7] Clements, G. (1989), "The role of the sonority cycle in core syllabification", in J. Kingston, & M. E. Beckman (Eds.), *Between the Grammar and Physics of Speech*, Papers in Laboratory Phonology 1, (pp. 283-333), Cambridge: Cambridge University Press.
[8] Lappin, K. (1981), "Analyse de variation : les liquides chez des enfants bilingues", *Recherches linguistiques à Montréal*, vol. 17, pp. 75-87.
[9] Studdert-Kennedy, M., & Goodell, W. (1992), "Gestures, features and segments in early child speech", *Haskins SR*, vol. 111/112, pp. 89-102.
[10] Bateson, E., & Ostry, D. (in press), "An analysis of the dimensionality of jaw movement in speech", *J. Phonetics*.
[11] Kies, R. (1992), *Étude des signaux mandibulaires. Plan de phase dans les occlusives*, Rapport Interne I.C.P., Grenoble.
[12] Stevens, K. (1989), "On the quantal nature of speech", *J. Phonetics*, vol. 17, pp. 3-45.
[13] Fant, G. (1960), *Acoustic theory of speech production*, The Hague: Mouton.
[14] Kuryłowicz, J. (1971), "A problem of Germanic alliteration". In M. Barahmer, S. Helsztyński and J. Kryżanowski (Eds.), *Studies in language and literature in Honour of Margaret Schlauch*, New York : Russell & Russell.
[15] Azra, J.-L., & Cheneau, V. (in press), "Jeux de langage et théorie phonologique. Verlan et structure syllabique du français", *Journal of French Language Studies*.
[16] Ohala, J. (1992), "Alternatives to the sonority hierarchy for explaining segmental sequential constraints", *Papers from the Parasession on the Syllable*. Chicago: Chicago Linguistic Society, pp. 319-338.
[17] Laks, B. (in press), "A connectionist account of French syllabification", *Lingua*.
[18] Heath, J. (1987), *Ablaut and ambiguity : phonology of a Moroccan Arabic dialect*, New York: State University of New York Press.
[19] Dell, F., & Elmedlaoui, M. (1985), "Syllabic consonants and syllabification in Imdl'wn Tashlhiyt Berber", *Journal of African Languages and Linguistics*, vol. 7, pp. 105-130.

SYLLABLE-LESS PHONOLOGY

Katarzyna Dziubalska-Kołaczyk
Adam Mickiewicz University, Poznań, Poland

ABSTRACT

The basic assumptions of a syllable-less model of phonology developed by the author [4] are presented. The beats-and-binding model, as it is called, is a natural functional model constructed within the framework of Natural Phonology [1,7] and natural polycentristic theory of language [2].

INTRODUCTION

The following exposition will be limited to the presentation of the preferences and principles of the beats-and-binding model which account for the four levels of phonological structure, i.e. level 0: the level of rhythmical preferences, level 1: the level of underlying phonological binding preferences, level 2: the level of phonotactic preferences and level 3: the level of articulatory preferences.

The unit "syllable" has been used in phonology to account for the processes thought to be conditioned by "syllable boundaries", "syllable weight", and segments sequences occurring within the "syllable". In the present framework, segments phonotactics constitutes a necessary consequence of the operation of the binding preferences, in the sense of counteracting the latter by keeping appropriate sonority distances between segments. Linguistic timing relationships between beats account for what used to be called "syllable weight". "Syllable boundaries" do not, in principle, constitute a phonological issue: speakers of a language are able to produce pauses in between beats as, among others, a side-effect of the operation of the binding preferences. The notion of so-called "syllable contact" (cf. [8]) is also untenable: patterns of consonants are a corollary of binding consonants to beats.

PREFERENCES AND PRINCIPLES

Level 0

A preference for isochrony and for the rhythmic structuring of a sequence in general is rooted in universal principles of human perceptual and motor behaviour. Rhythm can be broadly defined as "the structure of a sequence" consisting of not necessarily linguistic units. Humans possess strong motor-perceptual biases, which on the one hand constrain their production (in rate and pattern) and on the other impose structure on auditory sequences, even if the structure is physically not there. In speech, rhythm facilitates communication and intelligibility.

(1) The primary rhythm units are feet and their constituents - rhythmical beats, similarly as in music. There is a universal preference for two beats per foot: the former beat is preferably strong, the latter - weak, i.e. they constitute a trochee (a metrically falling accent).

(2) A beat (henceforth notated as "B") is realized by a phoneme which is traditionally referred to as a "syllable nucleus"; preferentially, it is a vowel (notated as "V"); secondarily, a consonant may acquire the function of a beat. A vowel is a better candidate for a beat due to its saliency potential based on its high sonority value and articulatory openness. Therefore, those among consonants which possess the latter two features to a higher extent qualify better for a beat than others.

(3) In accordance with the semiotic principle of figure and ground (cf. [3]), a hiatus between two beats is avoided by means of inserting a non-beat (henceforth notated as "n") in between, i.e. a consonant (notated as "C"). Only in this way do the figures, i.e. beats (B), receive a necessary ground, i.e. non-beats (n), in the form of consonants.

So, thanks to the preferences 1 to 3, speech flow consists of beats and non-beats, which are phonetically realized by perceptually and articulatorily contrasting sounds - vowels and consonants respectively. This is the most general structural level of phonology, the level of rhythmical preferences (=level 0).

The universal perceptual preferences operate at two levels:
Level 1: the level of underlying phonological binding preferences between beats (B) and non-beats (n), and
Level 2: the level of phonotactic preferences predicting the preferred actual sonority distances between and among vowels (V) and consonants (C), necessary for the origin and maintenance of consonant clusters.

Level 1

(4) Beats (B) and non-beats (n) in a sequence are joined by means of bindings in a binary fashion, i.e., e.g. in a sequence {BnB} there are maximally two bindings, i.e. a Bn-binding (a non-beat is bound to the preceding beat) and a nB-binding (a non-beat is bound to the following beat), i.e. {Bn + nB}. A beat, however, may potentially stay alone while a non-beat must be bound to a beat.

So, non-beats actively work against beat hiatus. The latter is a sequence of two beats, with NO binding between them. If a {B + B} sequence is not "broken" by a non-beat, it either (a) reduces to one beat {B}, represented phonemically/segmentally by a short vowel, which involves a change in the structure on Level 1 or (b) remains underlyingly a two-beat unbounded sequence i.e. {B+B}, represented phonemically either by a diphthong or a long vowel, which leaves the structure on the level of bindings (level 1) unaffected.

Two neighbouring beats (by default unbounded) without any "trace" of a non-beat in between them (i.e. no gliding, no preglottalization) and without any morphological boundary separating them, count as a long beat on level 0, i.e. on the level of universal rhythmical preferences. Counting of beats on this level corresponds to what is usually interpreted as speakers' intuitions about the number of "syllables". For example, a trochee on level 0 counts two, while on level 1 it often consists of three beats.

"Heaviness" in a beat-and-binding model is expressed by means of a number of

beats AND bindings (and beats and bindings count equal) contained within a binary foot, from a beat to another one or to a phonological word boundary. For example, in a {Bn} cluster (/VC/) there is one beat and one binding, i.e. count 2, and in a {BB} cluster (/V:/, /VV/) there are two beats, i.e. also count 2. Another possibility is to have an intervening morphological boundary between two beats in a hiatus.

Since bindings are perceptually based, binding preferences (i.e. how bindings preferably arise and combine) belong to the universal perceptual level of phonology. The latter consists of two levels; binding preferences occupy level 1, i.e. the level of underlying phonological binding preferences between beats (B) and non-beats (n).

(5) The two bindings differ in strength: the {nB} binding, i.e. the binding of a non-beat to the following beat (preferentially realized by a /CV/ sequence), is always stronger than the {Bn} binding, i.e. the binding of a non-beat to the preceding beat (preferentially realized by a /VC/ sequence).

A subjective perceptual measure of contrast between a beat and a nonbeat is constituted by sonority. At the level of phonological bindings beats are uniformly more sonorous than nonbeats. In objective terms, it is the degree of modulation in several acoustic parameters (amplitude, periodicity, spectral shape, F0; cf. [6]) that decides about a {nB}-binding being uniformly stronger than a {Bn}-binding. As Ohala (1990) notices, larger modulations have more survival value than lesser ones and therefore will persist in the languages.

Level 2

(6) Actual distances between segments in terms of sonority become relevant only at the level of phonotactic preferences (level 2). At this level sonority becomes a relative measure of distances between (and among) consonants and vowels, the values of which decide about the fate of segments in a phonotactic sequence.

The universal preferences consist in the strength-by-distance relations between segments measured in distance among the six positions on the sonority scale (e.g. *la* - distance of two positions, *st* - distance of one, *ka* - distance of five, etc.; so, e.g., ka > la > st).

Level 3

(7) Two main functions of phonology: to serve clarity of perception and ease of articulation are reflected in perceptual, hearer-friendly preferences, on the one hand, and in articulatory, speaker-friendly preferences, on the other. Another level of structure, called level 3, will be reserved exactly for the speaker-friendly preferences for articulatorily easy phonotactic sequences. While contrast is an underlying principle on the perceptual levels (cf. a figure-and-ground principle), similarity reigns on the articulatory level (cf. the proximity law).

The Principle of Balance

(8) Conflicts among universal preferences, and especially those between hearer-friendly and speaker-friendly preferences, are mediated by the major tendency for balance (cf. [5]), which is realized on a language-specific level. Conflict solutions are implemented language-specifically to establish language-specific or typological relationships between bindings, phonotactic preferences and articulatory preferences.

In the present framework, the effectiveness and optimality of the balanced solutions are emphasized, which is clearly possible within a functional approach to phonology advocated by the natural framework.

REFERENCES

[1] Donegan, Patricia and David Stampe. 1979. The study of Natural Phonology. In D. Dinnsen (ed.), *Current Approaches to Phonological Theory*. Bloomington: Indiana University Press. 126-173.

[2] Dressler, W.U. 1984. Explaining Natural Phonology. *Phonology Yearbook* 1. 29-50.

[3] Dressler, Wolfgang U. 1985. *Morphonology: the Dynamics of Derivation*. Ann Arbor: Karoma.

[4] Dziubalska-Kołaczyk, Katarzyna. 1994. *Phonology Without the Syllable. A Study in the Natural Framework*. Habilitationsschrift.

[5] Maddieson, Ian. in press. Universals of segment sequences: a cross-linguistic lexical survey. In Dressler, Wolfgang U., Prinzhorn, Martin and Rennison, John. (eds.) *Phonologica 1992*. Torino: Rosenberg& Sellier.

[6] Ohala, John J. 1990b. The phonetics and phonology of aspects of assimilation. In Kingston, J. and M. Beckman (eds.). *Papers in Laboratory Phonology I*. Cambridge: CUP. 258-275.

[7] Stampe, David. 1979. *A Dissertation on Natural Phonology*. New York: Garland.

[8] Vennemann, Theo. 1988. *Preference Laws for Syllable Structure and the Explanation of Sound Change*. Berlin: Mouton.

EFFECT OF ACCENT AND SEGMENTAL CONTEXTS ON THE REALIZATION OF VOWEL DEVOICING IN JAPANESE

Y. Nagano-Madsen
Department of Oriental Languages, University of Gothenburg,
Västra Hamngatan 3, 41117 Göteborg, Sweden.

ABSTRACT

Realization of vowel devoicing in Japanese was examined acoustically by varying accentual and segmental factors on nonsense words systematically. Effect of accent was found to be quite small, affecting only those vowels which occurred before [ɕ, ç, f]. There was a tendency for a vowel to resist devoicing when it was adjacent to [ç] and [f]. The phonetics and phonology of vowel devoicing in Japanese is discussed.

INTRODUCTION

There has been growing interest in the phenomenon of vowel devoicing in recent phonetic literature. These studies indicate that the phonetic conditions in which vowels are devoiced across several languages are extremely similar.

Vowel devoicing has hitherto been most extensively studied for Japanese. A standard description is that the high vowels /i/ and /u/ are devoiced between the two voiceless consonants, or between a voiceless consonant and a pause. It is known that devoicing is influenced by both segmental and accentual factors. However, it was not until recently that the relationship between the two factors was considered [1, 2]. While these two studies have concentrated on counting the frequency/rates of devoiced vowels in a dictionary and large database respectively, the present study focuses on an acoustic analysis, using a well-controlled material in which segmental and accentual factors are varied systematically. Potential spectral, durational, and intensity changes across these two variants have not been explored previously.

MATERIAL AND ANALYSIS

The corpus consisted of nonsense words having three CV syllables, where the segments in the first CVC were varied systematically. All possible combinations of devoiced vowels were included (cf. Table 1). The nonsense words such as *pipaka, pitaka, pikaka,* were embedded in a carrier sentence "Korewa ___ desu " (this is ___).

The entire corpus consisted of 110 words and it was read twice by a female speaker of Standard Japanese for each accent category at a comfortable speaking rate. Total of 440 samples were obtained. All speech material was digitized at 20kHz sampling rates. Duration and intensity (dB) were measured using the spectrograph and energy calculation commands of the CSL software package installed on a PC. Intensity was measured at a point approximately one third from the end of the frication phase.

RESULTS

Realization of devoicing

Results of the realization of vowel devoicing in different segmental contexts are shown in Table 1 with accent information. It is shown that there were three types of realizations: (1) those which were devoiced regardless of accent category, (2) those which were not devoiced, and (3) those which were not devoiced when accented. The vowels in the categories (2) and (3) all belong to the syllables which were followed by the postvocalic [ɕ, ç, f]. Those vowels which were not devoiced regardless of accent occurred more often with [ç] or [f] as the adjacent consonant.

Acoustic manifestation

In the present material, the vowel in question appeared in the initial syllable of CVCVCV, which was embedded in a carrier sentence. When the vowel was devoiced after a stop consonant, its presence was usually detectable in the much longer duration of aspiration (or frication noise), and the difference in quality (/i/ or /u/) was shown as different spectral patterns (cf. Figure 1). Syllables such as /pi, pu, ki, ku/ all had this pattern with slight reduction in duration and intensity when unaccented. Differences in quality between [ɕi̥] and [ɕu̥] with devoiced vowels, is more difficult to identify spectrographically.

*Table 1. Effect of accent and segmental contexts on the realization of vowel devoicing. *=was devoiced, X=was not devoiced, A=was not devoiced when accented. Two tokens for each syllable.*

	post-vocalic C									
syllable	p	t	k	tɕ	ts	s	ɕ	h	ç	f
pi	*	*	*	*	*	*	*	*	A	A
pu	*	*	*	*	*	*	A	*	A	X
tɕi	*	*	*	*	*	*	X	*	X	X
tsu	*	*	*	*	*	*	A/*	*	X	X
ki	*	*	*	*	*	*	A	*	A	X
ku	*	*	*	*	*	*	A	*	A	A
su	*	*	*	*	*	*	A	*	A	A
ɕi	*	*	*	*	*	*	A	*	X	X
çi	*	*	*	*	*	*	A	*	A	X
ɕu	*	*	*	*	*	*	A	*	X	A
fu	*	*	*	*	*	*	A/*	*	X	X

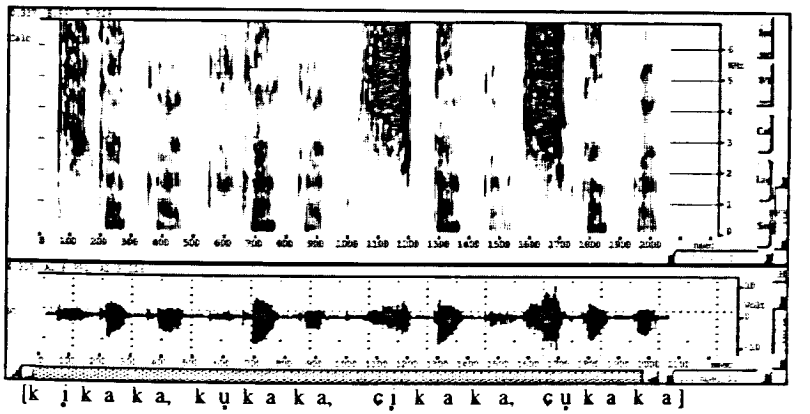


Figure 1. Sample spectrogram showing /kikaka, kukaka, ɕikaka, ɕukaka/ where /i/ and /u/ are devoiced.

Figure 2 displays sample tokens for [ki̥kaka, ku̥kaka, ɕi̥kaka, ɕu̥kaka].

The frication phase in accented syllables always had longer duration and the difference was found to be significant. Figure 2 presents the duration of frication phase for each syllable type except for [ç] since the recordings for this sound were unfortunately of poor quality.

As for intensity, the original material did not render reliable measurements of intensity partly because the recordings was divided into two sessions with a pause, partly because the speaker decreased overall intensity in the second session, and partly because the distance from the microphone was not fixed. These conditions made it difficult to do meaningful intensity measurements since the list was randomized. Therefore, the same speaker was called for an additional recording for some selected materials, this time using a headset microphone (cf. Figure 3). Like duration, the frication phase in accented syllables always had higher intensity, though the difference was not found to be significant by t-test for the pairs [pip, pits, ɕik, ɕitɕ].

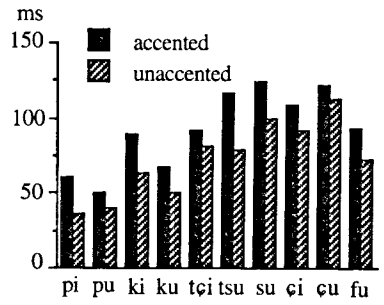


Figure 2. Duration of frication phase for each syllable type with or without accent.

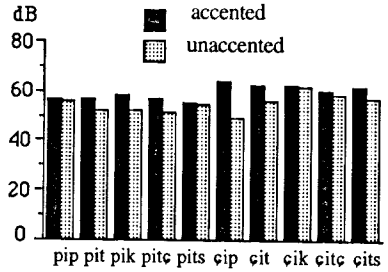


Figure 3. Comparisons of intensity of the frication phases of [pi] and [ɕi] before stops and affricates with or without accent.

DISCUSSION

Influence of accent

The influence of accent in the realization of vowel devoicing in Japanese has been known for a long time (cf. Sakurai [3] for a detailed description). By and large, it has been described that a potentially devoiceable vowel tends to resist devoicing once it is accented. Influence of accent, however, can appear in a different way by shifting accent to either the preceding or the following vowel. Many works using word in isolation form agree that accent indeed has a strong influence in preventing devoicing (cf. [2] for reference). However, recent works using longer words [4] and sentence database [2] found no overriding influence of accent in preventing devoicing. However, due to the restricted and uncontrolled curpus, these studies did not reveal much clearity as to the relationship between accentuation and segmental contexts.

In the present material of controlled corpus, the influence of accent on vowels in different segmental contexts appeared very regularly. Effect of accent in preventing devoicing was found only for those which were followed by [ɕ, ç, f]. The phenomenon appears partly phonological and partly phonetic because all these consonants are the allophonic variants before /i/ and /u/ (see discussion in next session), and because the three consonants do not behave exactly the same.

Influence of segmental context

The influence of prevocalic and postvocalic consonants has been discussed much in the literature and it has been indicated that there is a division between stops and fricatives. The division into stops and fricatives is also supported from physiological studies of glottal opening [8, 9].

Most studies agree that the phonetic nature of the postvocalic consonant is the most decisive factor in determining vowel devoicing, claiming that devoicing rate is higher when a vowel is followed by a stop than a fricative [5, 6]. The results of the present work agree with those of the previous studies in this respect except that the difference among fricatives became clearer in the present study. [s] and [h] behave besically the same as stops and affricates as postvocalic consonant.

The influence of prevocalic consonant is implicit in Kawakami's division of devoiced vowels in Japanese into two types: he says when [ki, pi, ku, pu, ɕu, tɕu] occur before a voiceless consonant, they will have devoiced vowels. On the other hand syllables [ɕi, tɕi, çi, su, tsu, fu] before a voiceless consonant usually do not even possess a devoiced vowel [7]. It is notable that the former group has, except for [(t)ɕu], stop consonants while the latter has fricatives.

The results of the present investigation indicate, however, that it is not adequate to consider the influence of adjacent segment in terms of segment alone. Instead, the nature of entire syllable should be considered. In the present study, those vowels which either did not get devoiced at all or affected by accent all occurred when they were followed by [ɕ, ç, f]. It should be remembered that [ɕ, ç, f] are the allophonic variants of /s/ and /h/ in Japanese: [ɕ] for /s/ before /i/, and [ç] and [f] for /h/ before /i/ and /u/ respectively. Furthermore, in the present material, these syllables were followed by [k], making the vowels in these syllables potentially devoiceable. In other words, in these words including [ɕ, ç, f], there were two succesive high vowels in a devoiceable environment. In only one word, namely in [ɕiɕika], both high vowels were devoiced regardless of accent. In all other words of this category, the first high vowel was either not devoiced at all or was affected by accent. It is interesting to note Kondo's results which report that devoicing of two successive syllables are avoided [4].

One interesting observation in this line of explanation, is that [tɕ] and [ts] do not behave similarly to [ɕ, ç, f] even though they too are the allophones (of /t/) before /i/ and /u/.

It may be at this point that the differentce between stop and fricative plays a role. In the acoustic signal, when a vowel is integrated in the frication phase or deleted all together, what is left is a frication phase. When a stop or affricate follows, there will be an occlusion after this phase, while if it were a fricative, another frication follows immediately. In the present material, except for [ɕiɕika], where both [i]s were devoiced, no two sequence of frication phases were observed when the first syllable was accented.

With few exceptions [10], the phenomenon of vowel devoicing has been discussed solely in terms of phonetics. However, the foregoing discussion indicates that the role played by phonology is considerable in the realization of vowel devoicing in Japanese. The problem we are concerned with is not the nature of prevocalic or postvocalic consonant per se. Rather, it is the question of which syllable including the vowel preceeds or follows, and how these syllables are structured within the enitre phonological system, where the allophonic variants of some consonants are typically found before high vowels.

ACKNOWLEDGEMENT

The present research is supported partially by a grant from Swedish Research Council in Humanities and Social Sciences.

REFERENCE

[1] Kawai, H., N. Higuchi, T.Simizu & S.Yamamoto. (1993), "A study of a devoicing rule for speech synthesis of Japanese (in Japanese)", *Proceedings of Acoustic Society of Japan 1993 Oct.*, 243-244.

[2] Nagano-Madsen, Y. (1994), "Vowel devoicing rates in Japanese from a sentence corpus", *Working Papers* 42, 117-127, Department of Linguistics and Phonetics, Lund University.

[3] Sakurai, S. (1985), "Kyootsuu-go no hatsuon de chuui subeki kotogara" Appendix to NHK (ed). *NHK Accent Dictionary*. Tokyo.

[4] Kondo, M. (1993), "The effect of blocking factors and constraints on consecutive vowel devoicing in Standard Japanese", Poster presented at Labphon 4, Oxford, 11-14 August 1993.

[5] Takeda, K. & H. Kuwabara. (1987), "Analysis and prediction of devocalizing phenomena (in Japanese)", *Proceedings of Acoustic Society of Japan 1987 Oct.*, 105-106.

[6] Yoshida, N. & Y. Sagisaka. (1990), "Factor analysis of vowel devoicing in Japanese (in Japanese)", *ATR Technical Report* TR-1-0159. ATR (Interpreting Telephony Research Laboratories),

[7] Kawakami, S. (1977), *Nihongo Onsei Gaisetsu*. Tokyo: Ohusha.

[8] Yoshioka, H., A. Löfqvist, H. Hirose & R.Collier. (1986), "How voiceless sound sequences are organized in terms of glottal opening gestures", *Annual Bulletin of the Research Institute of Logopedics and Phoniatrics* 20, 55-67, Tokyo University.

[9] Simada, Z.B., S. Horiguchi, S. Niimi, & H. Hirose. (1991), "Devoicing of Japanese /u/: an electromyographic study. *Proceedings of the XIIth International Congress of Phonetic Sciences*, 54-57. Aix-en Provence, France.

[10] Maekawa, K. (1989), "Boin no museika", In Miyoko Sugito (ed.), *Nihon-go no Onsei-Onin* 1, 135-153. Tokyo: Meiji Shoin.

SPATIAL PROPERTIES OF SPEECH MOVEMENTS

Vincent L. Gracco and Anders Löfqvist
Haskins Laboratories, 270 Crown Street, New Haven, CT USA

ABSTRACT

If there is one characteristic of speech that has plagued speech production theorists for years, it is variability. Acoustic correlates of a given phoneme and by inference vocal tract configurations exhibit variability arising from a number of sources. The present experiment was designed to examine the degree of spatial variation among a limited set of phonetic segments. Results suggest that variability in vocal tract positioning may be sound dependent reflecting different degrees of perception/production stability.

INTRODUCTION

An issue of theoretical importance in speech production is to determine the precision at which articulatory actions are being controlled. One characteristic of speech, however, that has plagued the development of a realistic understanding of control precision is variability. Variation may arise from many sources and it's understanding is crucial to the development of a realistic perspective on speech motor control. A general perspective can be obtained from consideration of the structure and function of the human nervous system as an information processing device. As pointed out by von Nuemann [1] the nervous system is an analog device that is ideally suited for reliable operation not precision. In this context it can be suggested that articulatory performance is good enough without incurring excessive "costs" [2] with the degree of precision inherently dependent on the listener's ability to extract meaning from the speech code. Alternatively, variation may be related to certain articulatory/acoustic relations such as those reflected in Stevens quantal theory [3,4] which implicitly assumes that certain consonants should exhibit more or less articulatory variability as a function of the proximity of so-called primary articulators to sensitive regions of the vocal tract in which small changes in articulation have large acoustic consequences.

Evaluation of spatial precision assumes that speech movements may in fact have spatial targets associated with them. The concept of spatial targets for speech was suggested by Lashley [5] in discussing space coordinate systems for controlling serial movements such as those for speech. In spite of the intuitiveness of speech motion planning in a spatial reference frame, the notion of spatial targets for speech have received little attention. One reason for the limited attention, again, appears to be related to the presence of variability in the observable signal in which variable vocal tract shapes yield acceptable acoustic signals [6] However, as noted above, this limitation may be related more to an unrealistic expectation regarding the precision of the articulatory target for speech. A recent perspective has been offered by Guenther [7] in which spatial targets for speech are viewed as regions rather than points (convex hulls) in orosensory space and conceptually is more attractive than target points. While the available data is limited it is difficult to imagine that speech is not planned to some extent in a spatial coordinate frame since inappropriately placed articulators will produce seriously compromised sounds. The purpose of the present experiment was to examine the spatial variation of a few of the phonetic segments of the language to determine how speech movement control varies as a function of phonetic identity.

PROCEDURE

The experimental group consisted of four subjects (two males, two females). Movements of the lips, jaw, and four points on the surface of the tongue were obtained using an electromagnetic transduction device [8]. The tongue receivers were placed approximately 1 cm behind the tongue tip and spaced approximately 1 cm apart. Data were hardware low pass filtered (200 Hz) and sampled at 625 Hz (12 bit resolution). Following the digitization, the voltages were digitally smoothed (25 points with a 3 dB point at 18 Hz) and the voltages were converted to positions in the midsagittal plane of the device. All data were rotated to the subjects' occlusal plane.

Subjects repeated a number of words embedded in the carrier phrase "Say _____ again." ten times. In order to examine the spatial variation associated with different phonetic segments, the two dimensional positions of the four tongue receivers were obtained at the time of zero (or minimum) speed associated with the target acoustic segment [9].

RESULTS

Shown in Figure 1 is the average tongue shape estimated from a cubic spline interpolation of the four average receiver locations for one subject for the three phones /s/, /r/, and /ae/ along with one standard deviation bars. For these comparisons the variability reflects the variation associated with repeating each word ten times. The words represented are /s/-"sack", /r/-"rack", and /ae/-"had". In considering the spatial variation there are two points of note: the degree of variation is quite different for the three segments and the different tongue regions display different degrees of variability.

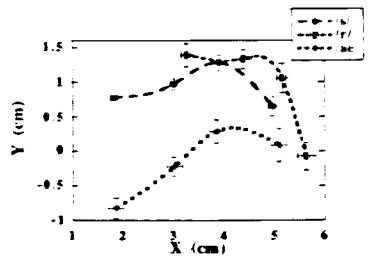


Figure 1. Derived tongue shape, from a cubic spline interpolation, during the steady-state or quasi steady-state behavior during the three phonetic segments /s/, /r/, and /ae/. The X dimension represents the subjects occlusal plane. Error bars reflect one standard deviation.

To examine the spatial variation in a more systematic manner the standard deviations in the spatial positions of the tongue receivers were obtained at the minimum speed associated with each phonetic segment of interest. The standard deviations in each spatial dimension were added providing an estimate of the average variation for each receiver location. For these comparisons the phone of interest was examined when it was in the initial position of the word. The words examined were: for /s/—"sack" and "sag"; for /r/—"rack" and "rag"; for /l/—"latter" and"ladder"; for /n/—"need" and "neat". The standard deviation in the X and Y dimensions were added for each of the 20 repetitions and are plotted as a function of the four phones reported here.

Figures 2 and 3 present the combined standard deviation in the X and Y dimensions for /s/ and /r/ (Figure 2) and /l/ and /n/ (Figure 3). As shown in Figure 1 there is a general trend for the variation at all positions to be greater for /r/ than for /s/ with a trend for the tongue front to show the smallest deviation compared to the tongue rear. Figure 3 shows the variation in tongue receiver positions for /l/ and /n/. The trend for these phones is for /l/ to show more spatial variation than /n/ for all receiver positions.

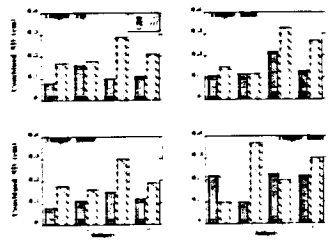


Figure 2. Combined standard deviations for /s/ and /r/ associated with each spatial dimension for each of the four tongue receivers for each of the four subjects.

While it might be concluded from these data that different phones differ in the degree of control precision required it should be noted that the results become more complicated when considering the same phones in different syllable positions. Shown in the next three figures are the estimated tongue shape for a single subject when the phones /l/, /r/, and /r/ were produced in different word/syllable positions. Figure 4 shows the tongue shape for /l/ in the words "ladder/latter" and "medal/metal". Figure 5 presents a similar comparison for /r/ when produced

word initial "need" and syllable medial "and". Interestingly, while /l/ demonstrated more spatial variation than /n/ in a similar context, the estimated tongue shape for /l/ is much more consistent across contexts than is /n/.

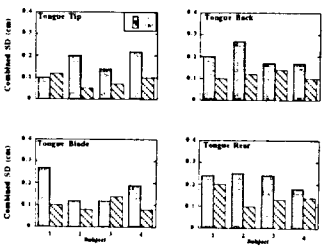


Figure 3. Combined standard deviations for /l/ and /n/ associated with each spatial dimension for each of the four tongue receivers for each of the four subjects.

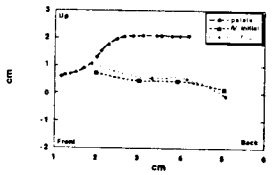


Figure 4. Estimated tongue shape for /l/ in two different syllable positions. Each receiver position reflects the average of 20 tokens (10 for each word). The top trace is an outline of the subjects hard palate.

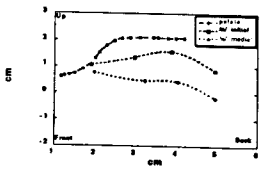


Figure 5. Estimated tongue shape for /n/ in two different syllable positions. Each receiver position reflects the average of 20 tokens (10 for each word). The top trace is an outline of the subjects hard palate.

DISCUSSION

Informed explanation of articulator variability rests on a number of assumptions regarding the control of speech movements; the degree of control precision and the goals for speech. At a conceptual level speech movements can be understood as goal-directed [10,11] and reflecting a level of control consist with obtaining changes in vocal tract configurations rather than movements of individual articulators [12,13]. The present results are consist with nervous system control operating on ensembles of articulators with differential degrees of precision depending on the context in which the variation is observed.

An example of the apparent looseness in the precision of articulatory control can also be found in recent simulation and synthesis results reported by Gay, Boe, & Perrier [14]. Parametric manipulation of vocal tract cross sectional area and constriction location was used to determine the acoustic and perceptual boundaries of certain isolated vowels. It was shown that the formants for each of the vowels were most sensitive to changes in cross sectional area compared to constriction location. Vowel perception, however, was insensitive to both manipulations. The results from Gay et al. [14] were somewhat at odds with the notion of the quantal characteristics of speech [3,4] suggesting rather that quantal regions may not necessarily be avoided because of the tolerance of the perceptual system. From these results it was concluded that the speech production mechanism has "...considerable latitude..." in specifying the articulatory targets. Limited kinematic data reported by Perkell and colleagues [15,16] is consistent with a relaxed degree of articulatory control.

In summary, the present report, while limited in scope, suggests that the specification of control precision can be thought of as an inherent property of each of the speech production units (phonetic segments) of the language. Moreover, the degree of variability may be systematically related to and ultimately reflect the perceptual tolerance of the language.

ACKNOWLEDGMENT

This work was supported by Grants DC-00121 and DC-00594 from the National Institute on Deafness and Other Communication Disorders, and in part by Esprit-BR Project 6975 - Speech Maps through Grant P55 from the Swedish National Board for Industrial and Technical Development.

REFERENCES

[1] von Nuemann, J. (1958). *The computer and the brain.* New haven: Yale University Press.

[2] Nelson, W. L. (1983). Physical principles for economies of skilled movements. *Biological Cybernetics*, 46, 135-147.

[3] Stevens, K. N. (1972). The quantal nature of speech: Evidence from articulatory-acoustic data. In E. E. David & P. B. Denes (Eds.), *Human communication: A unified view* (pp. 51-66). New York: McGraw-Hill.

[4] Stevens, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics*, 17, 3-45.

[5] Lashley, K. S. (1951). The problem of serial order in behavior. In L. A. Jeffries (Ed.), *Cerebral mechanisms in behavior* (pp. 112-136). New York: John Wiley.

[6] Ladefoged, P., DeClark, J., Lindau, M., & Papcun, G. (1972). An auditory-motor theory of speech production. *UCLA Working Papers in Phonetics*, 22, 48-75.

[7] 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.*

[8] Perkell, J., Cohen, M., Svirsky, M., Matthies, M., Garabieta, I., & Jackson, M. (1992), "Electro-magnetic midsagittal articulometer (EMMA) systems for transducing speech articulatory movements", *J. Acoust. Soc. Am.*, vol. 92, pp. 3078-3096

[9] Löfqvist, A. & Gracco, V. (1994), "Tongue body kinematics in velar stop production: Influences of consonant voicing and vowel context", *Phonetica*, vol. 51, pp. 52-67.

[10] Gracco, V. L., & Abbs, J. H. (1986). Variant and invariant characteristics of speech movements. *Experimental Brain Research*, 65, 156-166.

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

[12] Gracco, V. L. (1991). Sensorimotor mechanisms in speech motor control. In, H. Peters, W. Hultsijn, & C. W. Starkweather (Eds.), *Speech motor control and stuttering.* (pp. 58-78). North Holland: Elsevier.

[13] Gracco, V. & Löfqvist, A. (1994), "Speech motor coordination and control: Evidence from lip, jaw, and larynx interactions", *Journal of Neuroscience*, vol. 14, pp. 6585-6597.

[14] Gay, T., Boe, L.-J. & Perrier, P. (1992). Acoustical and perceptual effects of changes in vocal tract constrictions for vowels. *Journal of the Acoustical Society of America*, 92, 1301-1309.

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

[16] Perkell, J. S., & Cohen, M. (1989). An indirect test of the quantal nature of speech in the production of the vowels /i/, /a/ and /u/. *Journal of Phonetics*, 17, 123-133.

ARTICULATORY KINEMATICS IN STOP CONSONANTS

Anders Löfqvist and Vincent L. Gracco
Haskins Laboratories, New Haven, CT, USA

ABSTRACT

This paper examines tongue movement kinematics in stop consonant production with particular emphasis on variations due to vowel context and voicing.

INTRODUCTION

The aim of this work is to examine the nature and extent of articulatory variability in stop consonant production as a function of vocalic context and stop consonant voicing. Such an examination is useful for understanding the control of speech movements, since it can reveal the nature of the variability and how it is structured by different sources of influence.

PROCEDURE

The movement data were recorded using a three-coil transmitter system described in [1]. Receivers were placed on the upper and lower lips, the lower incisors, and at four positions on the tongue. For the sake of convenience, the tongue receivers will be referred to by their locations as tongue tip, tongue blade, tongue body, and tongue rear, cf. Figure 4. In addition, receivers placed on the bridge of the nose and on the upper incisors were used for correction of head movements. Two receivers attached to a plate were used to record the occlusal plane by having the subject bite on the plate during recording. All data were subsequently corrected for head movements, and then rotated and translated to bring the occlusal plane into coincidence with the x axis.

The linguistic material consisted of VCV sequences with all possible combinations of the vowels /i, a, u/ and the stop consonants /p, t, k, b, d, g/. The sequences were placed in the carrier phrase "Say ... again" with sentence stress occurring on the second vowel of the VCV sequence. Ten tokens of each sequence were recorded at self-selected speaking rates and intensity levels.

The articulatory movement signals (induced voltages from the receiver coils) were sampled at 625 Hz after low-pass-filtering at 200 Hz. The speech signal was pre-emphasized, low-pass filtered at 9.5 kHz and sampled at 20 kHz. The resolution for all signals was 12 bits. After voltage-to-distance conversion, the movement signals were low-pass filtered using a 25-point triangular window with a 3 dB cutoff at 18 Hz.

The tangential velocity of each receiver was calculated and used for velocity measurements and also for locating points in time for making position measurements. That is, movement onsets and offsets were taken as points of minimum (usually non-zero) tangential velocity. Movement displacement was calculated as the path traversed by a receiver between movement onset and offset. See [2] for a discussion of issues in the processing of two-dimensional movement signals.

RESULTS AND DISCUSSION

In this paper, we shall present results of tongue movements for two of the four subjects recorded. We shall first discuss the closing movement, then the articulatory configuration during the stop closure, and finally the release movement.

Stop consonant voicing has been shown to influence articulatory kinematics, but the data have mostly been limited to lip and jaw movements and are somewhat conflicting. We have shown [3] that the raising movement towards closure for a velar stop consonant was reliably faster, larger, and longer for a voiced than for a voiceless stop in a similar vowel context. The larger movement displacement was due to a lower position at movement onset for the voiced stop, as illustrated in Figure 1. For alveolar stops, these differences were not as robust, however. This is illustrated in Figure 2 which plots peak tangential velocity of the tongue tip receiver for the closing movement.

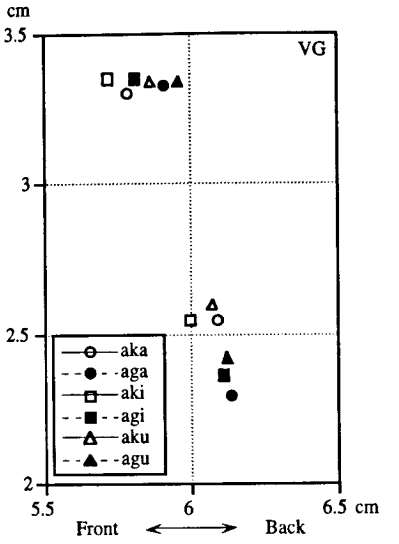


Figure 1. Average positions at onset and offset of tongue body receiver raising movement towards consonantal closure.. Onset positions in lower right, offset positions in upper left.

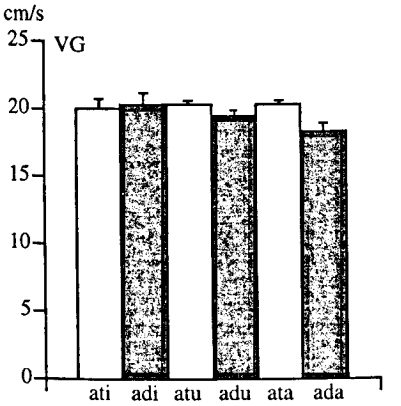


Figure 2. Peak tangential velocity of closing movement for alveolar stops (mean and standard error of the mean).

Statistical analysis showed no significant effect of vowel or voicing. The explanation for this is provided in Figure 3 plotting tongue tip receiver positions at onset and offset of the raising movement towards consonantal closure for the same data set. In contrast to the data shown in Figure 1, there is no clear difference in the onset positions in Figure 3 between voiced and voiceless stops.

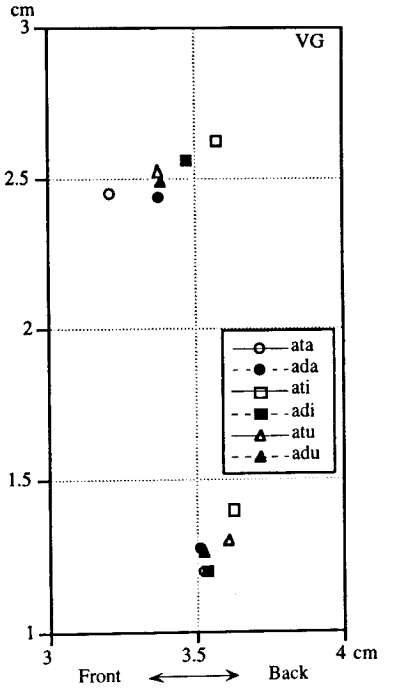


Figure 3. Average positions at onset and offset of tongue tip receiver raising movement towards consonantal closure. Onset positions in lower right, offset positions in upper left.

The vowel context has been shown to affect the articulatory configuration during stop closure, in particular for velar stops [4, 5, 6, 7].

Figure 4 shows average tongue receiver positions in six VCV sequences with velar stops and identical vowels before and after the consonant. The positions have been identified from minimum tangential velocity of each tongue receiver during consonantal closure. Cubic splines have been fitted to the data to get an estimate of the tongue

shape. The influence of the vowel on the consonantal closure is clearly evident from the different horizontal positions of the signals. The whole tongue is shifted horizontally depending on the vowel, and the order from front to back is /i/, /a/, and /u/.

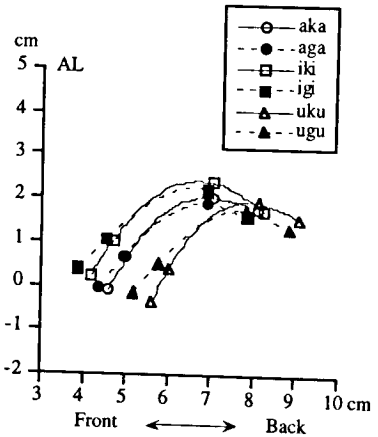


Figure 4. Average tongue receiver positions for velar consonants located at point of minimum tangential velocity during consonantal closure for each receiver. Cubic splines have been fitted to the data.

Figure 5 shows average tongue receiver positions for sequences with velar stops and an identical first vowel.

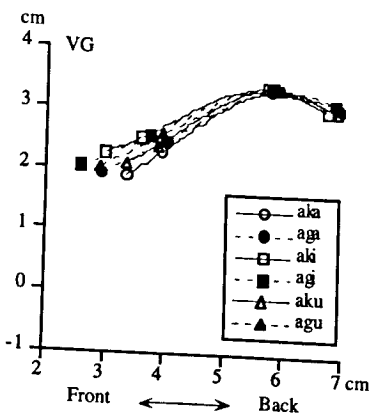


Figure 5. Average tongue receiver positions for velar consonants.

Here, the tongue body receiver shows a similar, although smaller, horizontal variation depending on the second vowel. The positions of the tongue tip and tongue blade receivers show larger variability, most likely because they are less directly involved in making the velar closure.

Figure 6 presents a similar plot for alveolar consonants in sequences where the first vowel is identical. Here, the tongue body and tongue rear receivers show more variation as a function of vowel context than those on the tongue tip and tongue blade. Again, most likely because the anterior part of the tongue is making the closure.

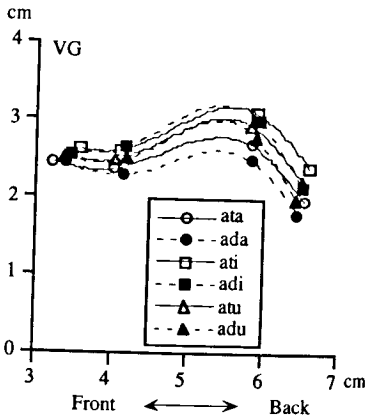


Figure 6. Average tongue receiver positions for alveolar consonants.

The release movement from the consonantal closure to the following vowel was heavily influenced by the quality of the vowel. Figure 7 plots peak tangential velocity of the tongue body receiver for the release movement. The increasing order of velocity is /i/, /u/, /a/, which corresponds to the displacement of the movement. While the vowel effect was robust, there was no consistent influence of stop consonant voicing on the release movement across subjects.

The point of minimum tangential velocity during consonantal closure offers an instant in time that can be used for measuring receiver positions. It is not necessarily the case, however, that such a minimum can be found for a given receiver, in particular for receivers on

those parts of the tongue that are not directly involved in making the closure. We should also note that at this point the tangential velocity is usually not zero. Tongue movements for velar stops usually follow curved paths, and there is thus continuous movement during the stop closure. cf. [6, 8]. This is also evident from the fact that the vertical and horizontal velocity profiles do not show any period of zero velocity. Thus, the goal in velar stop production should properly be seen as the making of a closure and not as a spatial position of the articulators.

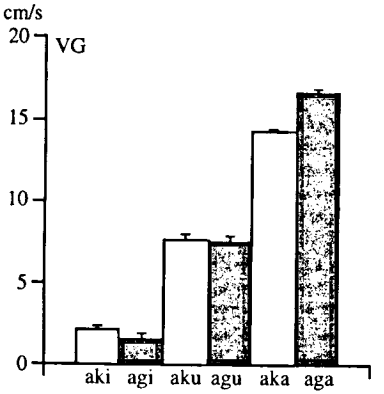


Figure 7. Peak tangential velocity of tongue body receiver for release movement for velar stops (mean and standard error of the mean)

In summary, the present results exemplify how articulatory movements in stop consonant production vary as a function of context. As we have argued elsewhere [9], such variability can be seen as the result of dynamic processes that operate on speech motor programs to scale them according to phonetic context.

ACKNOWLEDGMENT

This work was supported by Grants DC-00121 and DC-00594 from the National Institute on Deafness and Other Communication Disorders, and in part by Esprit-BR Project 6975 - Speech Maps through Grant P55 from the Swedish National Board for Industrial and Technical Development.

REFERENCES

[1] Perkell, J., Cohen, M., Svirsky, M., Matthies, M., Garabieta, I., & Jackson, M. (1992), "Electro-magnetic midsagittal articulometer (EMMA) systems for transducing speech articulatory movements", *J. Acoust. Soc. Am.*, vol. 92, pp. 3078-3096 .

[2] Löfqvist, A., Gracco, V., & Nye, P. (1993), "Recording speech movements using magnetometry: One laboratory's experience", in *Proceedings of the ACCOR Workshop on Electromagnetic Articulography in Phonetic Research.* Forschungsberichte (Institut für Phonetik und Sprachliche Kommunikation, Universität München) vol. 31, pp. 143-162.

[3] Löfqvist, A. & Gracco, V. (1994), "Tongue body kinematics in velar stop production: Influences of consonant voicing and vowel context", *Phonetica,* vol. 51, pp. 52-67.

[4] Öhman. S. (1967), "Numerical model of coarticulation", *J. Acoust. Soc. Am.,* vol. 41, pp. 310-320.

[5] Houde, R. (1968), "A study of tongue body motion during selected speech sounds", *SCRL Monograph* No. 2.

[6] Perkell, J. (1969), *Physiology of speech production,* Cambridge: MIT Press.

[7] Gay, T. (1977), "Articulatory movements in VCV sequences", *J. Acoust. Soc. Am.,* vol. 63, pp. 183-193.

[8] Mooshammer, C., Hoole, P., & Kühnert, B. (in press). "On loops", *Journal of Phonetics.*

[9] Gracco, V. & Löfqvist, A. (1994), "Speech motor coordination and control: Evidence from lip, jaw, and larynx interactions", *J. Neuroscience,* vol. 14, pp. 6585-6597.

AN ARTICULATORY STUDY OF LIQUID APPROXIMANTS IN AMERICAN ENGLISH

*Shrikanth Narayanan, Abeer Alwan, and Kate Haker**
Department of Electrical Engineering, UCLA, Los Angeles, CA 90024
**Cedars-Sinai Medical Center, Los Angeles, CA 90048*

ABSTRACT

Articulatory patterns of the liquid approximants in American English are analyzed through MRI and EPG. MR images of the vocal tract during sustained productions of /l/ (both dark and light allophones) and /r/ (word-initial, syllabic, bunched, and retroflexed) by 4 subjects are used for quantitative and qualitative analyses of the 3D vocal tract geometry and tongue shapes. EPG contact profiles are used for studying inter- and intra-speaker variabilities in linguapalatal contact patterns.

INTRODUCTION

Articulatory patterns of the liquid sounds /l, r/ in American English are analyzed through magnetic resonance imaging (MRI) and electropalatography (EPG). MR images of the vocal tract during sustained productions of the lateral sound /l/ (both dark [ɫ] and light [l] allophones) and the rhotic approximant /r/ (in word-initial and syllabic positions, and the bunched and retroflexed allophones) are used for measuring vocal-tract lengths, area functions, cavity volumes, and for the analysis of the 3D vocal tract and tongue shapes. EPG contact profiles are primarily used as a source of converging evidence for the results of the MRI study and for studying inter- and intra-speaker variabilities in linguapalatal contact patterns.

TECHNIQUES

The sounds were produced in a neutral vowel context by 4 phonetically-trained subjects (2 males: MI, SC and 2 females: AK, PK). MR images were collected using a GE 1.5 Tesla SIGNA machine under a fast SPGR protocol with 3 mm image slice thickness and no interscan spacing in the coronal, axial, and sagittal planes. The subjects, in supine position, sustained each sound for about 13-16 s enabling four to five image slices to be recorded (about 3.2 s/image/plane). Analysis techniques are similar to those described in [1]. EPG data were collected using a Kay palatometer that employs an acrylic pseudopalate, custom-fit for each subject, with 96 sensing electrodes. The sweep rate of the system was 1.7 ms and the sampling period, 10 ms. Eight repetitions of each sound were recorded with the subjects in both supine and normal sitting positions.

RESULTS

Lateral approximants: MR images for both [l] (as in the word 'led') and [ɫ] (as in 'bell') indicate that the *mid-sagittal* tongue shapes can be different across subjects. Common characteristics, however, are revealed in cross-sectional and 3D tongue shapes, area functions, and linguapalatal contact profiles. These sounds are characterized by a lingual occlusion or, just a constriction as observed in the [ɫ] of one subject. The occlusion location is 1.0-1.5 cm away from the lip opening and the occlusion length, 0.5-1.0 cm in the alveolar region with relatively small openings (ranging in area between 0.1-0.8 cm^2) around both sides of the occlusion. The side '*lateral channels*' begin appearing from where the alveolar occlusion/constriction is seen and continue posteriorly until the lingua-velar contact is established (i.e. lingual contact with the roof of the oropharynx in the velar region which is about 5-6 cm from the lip opening). The right and left channels appear to be, in general, unequal and their areas start increasing behind the alveolar occlusion (due to inward lateral compression of the tongue body) and start decreasing again as the lingua-velar contact is approached. Area functions calculated along the midline of the vocal tract are shown in Fig. 1. Note that the lateral areas in the region of medial occlusion are not shown in this figure.

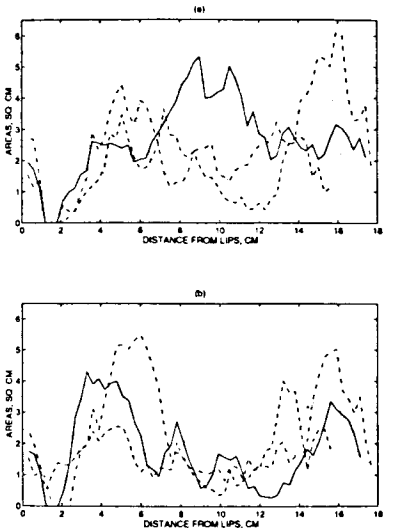


Figure 1: *Area functions, in cm^2, for different subjects: solid (AK), dashed (PK), dot-dashed (MI), dotted (SC). Top panel: [l], bottom panel: [ɫ]. The abscissa for the area functions are distances (in cm) from the lip opening.*

The linguapalatal contact associated with the alveolar occlusion extends laterally into the palatal region, the degree of which varies between the light and dark variants with [l] exhibiting less lateral contacts. The articulations of subjects MI and SC were apical while those of subjects AK and PK were laminal. The extent of the lateral linguapalatal contact also appears to depend on the apical or laminal nature of the articulation, with the laminal articulations exhibiting a more extended contact than the apical ones.

The cross-sectional tongue shapes immediately behind the medial linguoalveolar occlusion appeared either flat or slightly 'concave' (particularly until the disappearance of the lateral linguapalatal contacts). The concave shape was mainly due to a 'grooving' tendency along the mid-sagittal line. Analysis of the 3D tongue shapes, however, revealed that the general tongue body shape behind the occlusion tends to be convex. A sample 3D tongue shape is shown in Fig. 2 for [ɫ] of subject SC. The 3D tongue shapes indicate that the posterior tongue body shows a tendency towards an *inward lateral compression* which is directed towards the mid-sagittal plane. This enables the creation of lateral flow channels in the space between the curved sides of the tongue body and the teeth. The anterior medial grooving observed in the laterals, which is less prominent than that observed in alveolar sibilants such as /s/ [1], is attributed to the inward compression of the posterior tongue body. Unlike alveolar fricatives, the grooving does not continue through the posterior tongue region as a concave surface, suggesting that it is not a key component of a medial airflow channel. Hence, the modification of the tongue body contour, in terms of surface flattening and/or grooving, observed in some portions of the tongue surface for some subjects, is *not* a primary articulatory characteristic satisfying an aerodynamic constraint, but merely represent secondary effects of the linguapalatal bracing and the lateral contraction of the posterior tongue body.

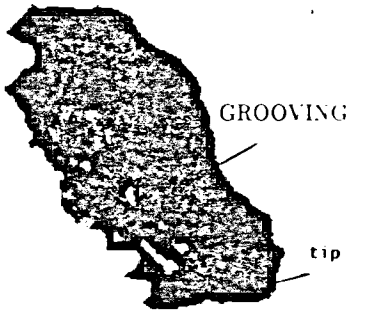


Figure 2: *Lateral perspective of the 3D tongue shape: [ɫ] (subject SC).*

The merging of the lateral channels with the central opening along the palatal region results in crescent-shaped cross-sections and relatively large area values (Fig. 1). The extent of the lateral flow component in the palatal region behind the linguoalveo-

lar occlusion is limited by the extent of the linguapalatal contact behind the occlusion: [l] typically reveals more lateral contact than [ɬ], thus explaining the smaller area function values in the palatal region consistently observed in [l] when compared to [ɬ]. The back region areas for [l] show significant inter-subject variability: the areas of the upper pharyngeal/uvular region are much smaller for subjects MI and SC due to a slightly raised and retracted posterior tongue body, perhaps a result of their apical articulation. In the case of [ɬ]s, on the other hand, all subjects reveal decreased areas in the upper pharyngeal/uvular regions due to a significant retraction of the tongue root and/or raising of the posterior tongue body. In addition, the effect of this upper pharyngeal 'constriction' is found to extend either as far as the velar/uvular region or through the entire lower pharyngeal region depending on the particular part of the tongue body actively involved in the constriction formation: either the upper-part of the tongue root (together the posterior/dorsal tongue body) or the entire tongue root. These results indicate that velarization, which is typically associated with /ɬ/, is not a consistent characteristic across speakers although decreased pharyngeal areas, when compared to those of /l/, is a consistent feature for all subjects.

The linguapalatal contact profiles from the EPG data were consistent with the observations of the MRI data. In addition, the differences between the articulations in the supine and upright positions, in general, were not significant. Left-right asymmetry in relative tongue positions and linguapalatal contacts were found only for subjects AK and PK; subject PK exhibited consistent asymmetry in the postpalatal/velar region with greater right-side linguapalatal contact while that of subject AK was not consistent.

Rhotic approximants: During the production of the American English /r/, the vocal tract appears to be characterized by three cavities due to the presence of two distinct supraglottal constrictions. The primary constriction occurs in the buccal cavity and the secondary constriction, in the pharyngeal cavity. A sample 3D vocal tract of /r/, by subject SC, and the 3D tongue shape associated with it are shown in Figs. 3 and 4 , respectively. The buccal constriction may arise anywhere in the palatal region: the more forward ones are typically due to a raised anterior tongue and the posterior ones, due to a raised dorsum. For our subjects, the buccal constriction begins 2.4-4.8 cm away from the lips and extends over 1.5-2 cm with minimum areas anywhere in the range of 0.25-0.7 cm^2. The secondary constriction occurs typically in the mid-pharyngeal region due to an advanced tongue root ('pharyngealization'). Analyses indicate that a more anterior buccal constriction is associated with a more superior pharyngeal constriction. A large volume anterior to the buccal constriction arises due a tongue body that is drawn inwards. The anterior tongue body is characterized by convex cross-sections.

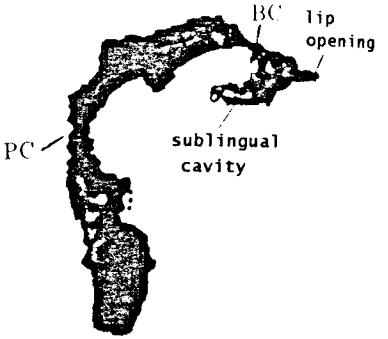


Figure 3: *Lateral perspective of the 3D vocal tract: 'word-initial' /r/ (subject SC).* BC *refers to the buccal constriction and* PC *refers to the pharyngeal constriction.*

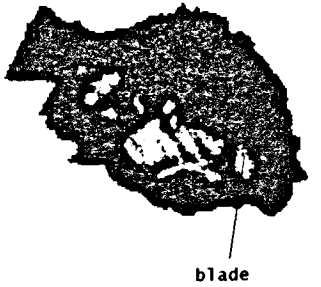


Figure 4: *Lateral perspective of the 3D tongue shape: 'word-initial' /r/ (subject SC).*

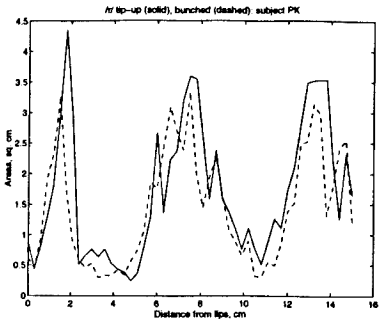


Figure 5: *Area functions for subject PK's /r/: 'retroflex' (solid), 'bunched' (dashed). The axes are similar to Fig. 1.*

Similarly, a large volume posterior to the buccal constriction (and superior to the pharyngeal constriction) is created by a significantly lowered posterior tongue body that exhibits a prominent concave shaping. The change in the cross-sections, from the convex anterior shapes to the more concave posterior shapes, appears to be more abrupt for the buccal constrictions that are at a more posterior location in the front region, resulting in more abrupt changes in the area functions. Variabilities in the details of the relative cavity sizes and their locations, which largely depend on the individual subject's articulation patterns and oral morphology, are expected to introduce variabilities in the corresponding acoustic patterns.

Subjects AK and MI produced /r/s as they would appear in 'word-initial' and 'syllabic' positions while subject PK produced deliberately /r/s with the tongue tip curled up ('retroflex' /r/) in one case and the dorsum bunched ('bunched' /r/) in the other (both varieties occur in PK's speech). Subject SC produced only the 'word-initial' version. Comparison of the bunched and retroflex /r/s produced by subject PK revealed that, in spite of the raised tongue tip in the latter case, the primary buccal constriction is attributed to a raised dorsum in both cases, and a three-cavity vocal tract description still holds. Area functions calculated along the midline for the bunched and retroflexed /r/ for subject PK are shown in Fig. 5. The minimum constriction is found at a more posterior location for the retroflex /r/ and the areas in the constriction region tend to be slightly larger than the bunched /r/ due to a relatively flatter tongue cross-sectional surface. The areas anterior to the buccal constriction tend to be larger for the retroflex /r/; the anterior cavity volume was 4.5 cm^3 for the bunched /r/ and 6.1 cm^3, for retroflexed /r/. The areas behind the buccal constriction were similar for both the bunched and retroflexed cases (Fig. 5). EPG contact profiles, which are restricted to the lateral postpalatal regions, were similar for both cases. Asymmetry was found in subject PK's palato-velar region with a more right-side favored linguapalatal contact observed in both the MRI and EPG data.

For the other subjects, the general tongue body shapes and area functions appeared very similar for the /r/s in both word-initial and syllabic positions although syllabic /r/s tend to show larger areas in the cavity between the buccal and pharyngeal constrictions. The buccal constriction in the /r/s of subjects AK and MI, in both word-initial and syllabic cases, were produced with a raised dorsum resulting in a tongue body shape resembling a canonical bunched /r/. The /r/ of subject SC, on the other hand, was produced with a raised anterior tongue body, rather than a raised tongue tip, resulting in a more anterior, and shorter, buccal constriction when compared to those seen in the other subjects.

As in the case of the lateral approximants, the EPG contact profiles revealed no systematic differences between the articulations in supine and upright positions.

ACKNOWLEDGEMENTS: This work was supported in part by NSF.

REFERENCES

[1] S. Narayanan, A. Alwan, and K. Haker. "An MRI study of Fricative Consonants," *ICSLP '94 Proc.* pp. 627-630. A full-version of this paper will appear in *JASA*, July, 1995.

QUANTITY CONTRAST IN SWEEDISH KINEMATIC AND ACOUSTIC PATTERNS

Rudolph Sock and Anders Löfqvist

Institut de Phonétique de Strasbourg, 22 rue Descartes - 67084 Strasbourg, France and Haskins Laboratories, 270 Crown Street, New Haven-CT 06511 U.S.A.

ABSTRACT

The aim of this investigation is to find out if vowel quantity contrasts for Sweedish emerge in the acoustic VC and CV domains and also on the movement level, with increased speaking rate. Results obtained on both the acoustic and movement levels, suggest that quantity contrast is not only portrayed in the domains examined, but that it is also maintained across speech rate.

INTRODUCTION

Sweedish, is a language, that uses quantity for grammatical and lexical distinctions. It has a contrast between two lengths, and this contrast occurs only in lexically stressed syllables, where the vowel is either reffered to as being short or long. If the vowel is long, it is either word final or followed by a short consonant, VVC; if the vowel is short, it is followed by a long consonant, VCC [1]. It should, however, be noted that while quantity contrast in Sweedish may also be accompanied by additional correlates such as vowel quality and diphthongization, the importance of these factors is not easy to pinpoint [2].

The present investigation attempts to show how vowel quantity contrasts emerge, not only on the acoustic level but also, on the articulator movement level, specified by jaw, lower lip, tongue tip and tongue body vertical displacements. Moreover, the CV span will be examined to see how the complementary distribution of the quantities (*i.e.* short-long *vs.* long-short) contributes to the distinctivity of contrasts [2]. The effect of voicing on contrasts is also examined. Speech rate will be varied as a perturbing factor of the linguistic system; thus it would be possible to evaluate the robustness of the linguistic contrasts.

METHOD

One Sweedish subject produced CVCV words and non words that were embedded in a carrier sentence. Utterances were produced fourteen times at two speech rates, normal (conversational) and at a self-selected fast speaking rate, with all possible combinations of the vowels /i, a, u/ *vs.* /i:, a:, u:/ and the stop consonants /p, t, k/ *vs.* /b, d, g/. Thus there were 24 conditions in all: 2 speech rates (normal and fast) x 2 vowel lengths (short and long) x 3 consonant-types (bilabials, dentals and velars) x 2 consonant categories (voiced and unvoiced). Results reported here will only focus on the vowel /a/ context.

The movement data were recorded using the Haskins Laboratories three-coil transmitter system [3]; [4]. Receivers were attached to the upper lip, lower lip, jaw, tongue tip, tongue body, and tongue rear (locations of the tongue receivers are reffered to approximately). Additional receivers, placed on the bridge of the nose and on the upper incisors, were used for correction of head movements. Care was taken during each reveiver placement to insure that it was positioned at the midline with its long axis perpendicular to the sagittal plane[4].

To obtain instantaneous velocity, the first derivative of the position signals was calculated using a 3-point central difference algorithm. Velocity signals were smoothed using the same triangular window.

Based on articulatory events in the velocity signal, two cycles were identified in the movement of each articulator (jaw, lower lip, tongue tip and tongue body). These velocity cycles were determined, as the interval between successive negative or positive peaks associated, respectively, with the lowering or raising movement in the production of a vowel and a consonant. An oral opening phase, associated with the production of the vowel, was defined within the oral opening cycle, and an oral closing phase, associated with the production of the consonant, was defined within the oral closing cycle. Two acoustic cycles were also defined: one as the recurrence of the onset of a clear formant structure (*i.e.* in the VC domain), corresponds to the vocalic cycle; the other as the offset of a clear formant structure (*i.e.* in the CV domain), corresponds to the consonantal cycle. Acoustic phases were specified within the appropriate acoustic cycle, as the interval that presents a stable formant structure for the vocalic phase, and as the obstruent portion, for the consonantal phase. It is hypothesized that oral opening and vocalic phases would reveal quantity contrasts, while oral closing and consonantal phases would highlight concomitant consonantal differences.

RESULTS

Data processing is based on the percentage of time taken by each phase in its cycle. ANOVAs were performed on measured intervals as dependent variables and grouping factors *Quantity*, *Voicing*, speaking *Rate* and *Place* of articulation.

Quantity contrasts

Acoustic relative timing

In the Vocalic Cycle corresponding to the VC domain, and at a normal speaking rate, quantity differences for the unvoiced category emerge distinctly along the vocalic phase ($p<0.01$), while cycle or syllabic durations are comparable for VCCs *vs.* VVCs. These quantity differences are maintained with speech rate increase ($p<0.01$), due to the relative timing stability of the linguistic classes at around 58% of the cycle for long vowels and at around 35% for their shorter counterparts (see Figure 1, left panels). This finding is true also for the voiced category [2], although there is a general tendency for linguistic classes to converge as speaking rate increases; however, all classes remain distinct due to the combined effect of vocalic and syllabic differences ($p<0.01$).

Movement relative timing

In the articulator lowering peak velocity cycle — generally associated with the acoustic VC domain — and at a normal speaking rate, quantity differences for the unvoiced category also emerge distinctly along the oral opening phases ($p<0.01$); movement cycles still show comparable values for VCCs *vs.* VVCs. In their vertical lowering displacements, jaw, lower lip, tongue tip and tongue body opening movements for vowel production show clear-cut phasing patterns. Patterns obtained for jaw lowering, in the bilabial context, however, are more distinct than those observed for lower lip lowering, thus indicating that the jaw plays a more critical role in portrayiing vowel quantity contrasts. Figure 1(right panels) shows how short and long vowels emerge and differ as to the opening phases of the articulators; these differences are further maintained in fast speech largely by relative intraclass stability. These results are also valid for the voiced category.

Closure durations

Acoustic relative timing

In the consonant cycle corresponding to the CV domain, and at a normal speaking rate, closure differences for the unvoiced category emerge distinctly along the consonantal phase ($p<0.01$): the closure duration for the VCC

category for /p, t, k/ take up around 65% of the CV domain, while that of the VVC category occupy, on an average, 45% of the cycle for the three consonant types. When speaking rate is increased, these consonant closure differences are maintained for /p, t/ ($p<0.01$) but tend to dissapear for /k/. The latter remark also applies to the voiced category, although closure differences become more critical in fast speech, especially for the lingual consonants.

Movement relative timing

In the articulator raising peak velocity cycle — generally associated with the acoustic CV domain — and at a normal speaking rate, closure differences for both the unvoiced and voiced categories do not emerge distinctly along the oral closing phases. The tendency, however, is for the VCCs to take up a higher closing phase percentage of the oral closing cycle.

Voicing contrasts

On the acoustic level, there is a systematic vowel phase difference, in the VC domain, between the voiced and unvoiced categories for both short and long vowels: the vowel, proportionally, is longer in the voiced context than in the unvoiced context at both speaking rates ($p<0.01$). In the CV domain, closure differences are less evident: unvoiced consonants tend, however, to take up a higher phase percentage of the closure cycle than their voiced counterparts.

On the movement level, the oral opening phase, associated with vowel production, does not show clear-cut patterns, even though the tendency is for vowels in the voiced context to have a longer opening phase than in the unvoiced context [4], especially for /p, t/. The oral closing phase, associated with consonant closure, does not show coherent pattern differences between the two categories.

In summary, these preliminary results show that, the effect of rate is significant as all cycles (syllables) are compressed with increase of speaking rate. However, linguistic contrasts are maintained by the relative stability of vowel (and associated oral opening) phases. The complementary consonantal differences also contribute to maintaining the linguistic contrast, although such differences are less resistant to speech rate, as revealed by the tendency for consonantal phases (and associated oral closing phases) to converge when the task becomes difficult. Systematic relative stability may be accounted for, to some extent, by biomechanical and aerodynamic constraints. However, acoustic and kinematic maintained differences that correspond to the different linguistic tasks are presumably constrained by the perceptual requirements of the linguistic code [5].

ACKNOWLEDGMENTS

This research was supported by NIH Grant DC-00865, from the National Institute of Deafness and Other Communication Disorders, the Fyssen Foundation, Esprit-BR Project 6975 — Speech Maps, and Grant P-55 from the Sweedish National Board for Industrial and Technical Development.

REFERENCES

[1] Elert C.C. (1964) Phonological studies of quantity in Sweedish (*Almqvist & Wiksell, Uppsala*).
[2] Engstrand O. Krull D. (1994) Durational correlates of quantity in Sweedish, Finnish and Estonian: cross-language evidence for a theory of adaptive dispersion. *Phonetica 51, 80-91.*
[3] Perkell J.S. Cohen M.H. Svirsky M.A. Matthies M.L. Garabieta I. Jackson M.T. (1992) Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. *J.A.S.A., 3078-3096.*
[4] Löfqvist A. Gracco V. (1994) Tongue body kinematics in velar stop production: influence of consonant voicing and vowel context. *Phonetica 51, 52-67.*
[5] Abry C. Orliaguet J-P. Sock R. (1990), Patterns of speech phasing. Their robustness in the production of a timed linguistic task: single *vs.* double (abutted) consonants in French. *European Bulletin of Cognitive Psychology 10,* 269-288.

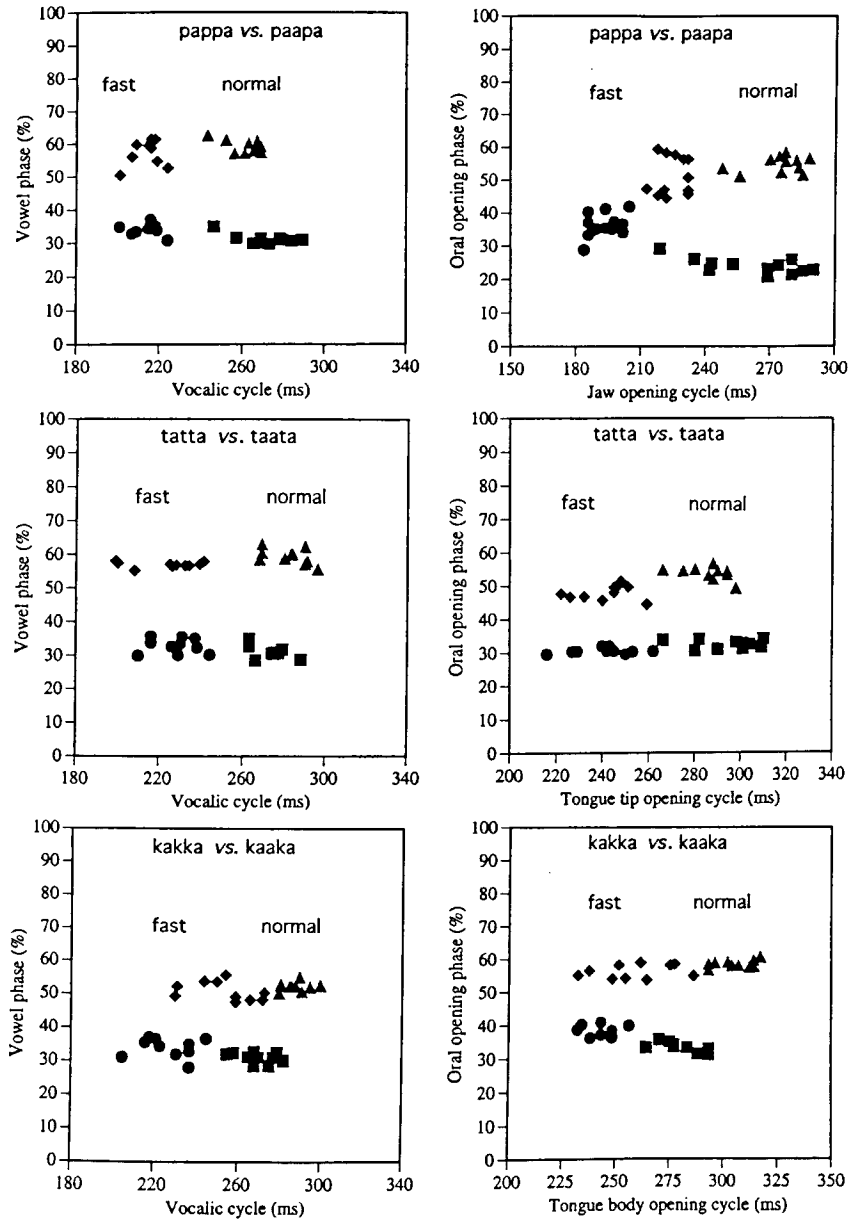


Figure 1. Scatterplots of vocalic patterns for VVCs in normal and fast speech (▲, ◆) and for VCCs in normal and fast speech (■, ●) on the acoustic level (left panels) and on the movement level (right panels). See text for details.

Perceptual analysis of compensatory strategies in the production of the French rounded [u] perturbed by a lip-tube.

SAVARIAUX Christophe, PERRIER Pascal & SCHWARTZ Jean-Luc
Institut de la Communication Parlée, URA CNRS 368, INPG & Université Stendhal
46, avenue Félix Viallet, 38031 GRENOBLE Cedex 01, France
E-mail : {savariaux, perrier, schwartz}@icp.grenet.fr

ABSTRACT

In order to analyse the compensation strategies develop by 11 speakers to achieve, in spite of the lip perturbation, the perceptual goal we proposed two perceptual tests. Thus, the vowel [u] produced both in normal and perturbed conditions are perceptually evaluated. The results allow us to determine a [u]-region in the F1/(F2-F0) plane within which the produced sound is identified as a good quality [u].

INTRODUCTION

In a previous work we proposed an experiment the aim of which was to test the respective roles of the articulatory and acoustic levels in the control of vowel production. This experiment involved a lip perturbation impeding the usual articulatory strategy for the production of a rounded vowel [1]. To study how speakers are able to achieve their speech goal, in spite of the perturbation, a 20-mm diameter lip-tube was inserted between their lips while they produced the French rounded vowel [u]. Acoustic and X-ray articulatory data were recorded for eleven French native speakers for both normal and lip-tube conditions. Articulatory measurement in lip-tube conditions were made first immediately after the insertion of the tube (PF condition) and second after an "adaptation" procedure of 20 trials (PL condition).

The analysis of the results based on the F1/F2 comparison [1], showed that one out of eleven speakers (speaker OD) was able to acoustically compensate for the perturbation. For this aim, he moved his tongue backwards into the pharyngeal cavity. Remaining speakers showed a large variability in their compensation ability: four speakers (YP, ML, LJ, ML) presented no noticeable articulatory change and six others (MP, BC, CH, GA, JM, JY) displayed variable extents of tongue backward movements within the palatal region. Speaker OD's compensation abilities show that there is no physiological limitation to move the tongue backwards. Moreover, articulatory reactions, when they exist, were directed in the "right" direction, namely towards a compensation in the acoustic space. Thus speech production seems to be guided by auditory requirements.

However, for the majority of speakers the compensation is not completely achieved and the shape of the tongue remains close to what it is in the normal condition. One interpretation would be that speakers experienced difficulties in producing unusual articulatory configurations and hence resorted to the usual one. Another explanation could be found at the perceptual level. Indeed, the perceptual goal is probably not sufficiently characterised in the F1/F2 plane; in spite of an insufficient compensation in the F1/F2 plane, speakers could have reached their perceptual goal.

To refine the analysis of our data, we propose in this paper the results of two perceptual tests which have been achieved on both normal and perturbed utterances for all speakers.

THE IDENTIFICATION TEST

Method and procedure

The aim of the first test was to categorise the vowels produced in perturbed condition. These vowels were presented within a set of seven vowels delimiting the maximal vowel space of each speaker: the three extreme vowels [i, a, u] and four vowels acoustically located close to vowel [u], i.e. [o, œ, y, ø]. Each vowel was recorded in three conditions: one normal condition (N) and two perturbed conditions with a lip-tube (PL and P2). Thus, 21 stimuli are available for each speaker. All stimuli were presented to 17 French adult listeners.

The procedure adopted for this test was as follows: in a sound-treated room, listeners listened to a stimulus by means of a headphone and then had to choose, without any time constraint, an answer on a monitor. The choice was made among seven possibilities [i, a, u, o, œ, y, ø]. Two seconds after the selection, another stimulus was sent to the headphone.

Results

Only results concerning the identification of [u] in N and PL conditions are presented here. For condition N and for all speakers, maximum identification (100%) was essentially obtained. The results for the PL condition are not so clear: seven speakers obtained a score near to maximum identification; 100% for 6 of them (CH, GA, LR, MP, OD, YP) and 94% (1 error among 17) for one (BC). For the remaining speakers, identification varies from 12% (speaker JM) to 0% for speaker JY (6% for speakers ML and LJ).

Interpretation

In condition N vowel [u] is clearly identified. We shall consider that an identification score smaller than 94% (more than one mistake) corresponds to a sensitive decrease of the vowel quality. Thus in the PL condition, 7 speakers have achieved, in spite of the perturbation, the required perceptual goal for vowel [u]. This somehow contradicts the analysis we proposed on the basis of acoustic data. We can observe an increase as high as 60% in F1 (speaker CH) or 46% in F2 (speaker LR) without change in vowel identification. However, when the F1 and F2 values become respectively higher than 400 Hz and 1100 Hz, the identification is no more correct. A large increase of F1 leads to a change of category from [u] to [o] (speaker LJ); simultaneous increases of F1 and F2 lead to confuse perturbed [u] with [œ] (speakers JY and ML).

The JM case clearly shows that F1 and F2 are not sufficient to understand in detail identification scores: the score was only 12% correct in spite of a formant pattern comparable to the one for speaker LR (F1 = 343 Hz, F2 = 851 Hz versus F1 = 344 Hz, F2 = 876 Hz) who obtained a maximum identification score.

The high perceptual scores obtained for 7 speakers could incite, in a first analysis, to minimise the conclusion made from the study of F1/F2 pattern: seven speakers and not only one, appear to be able to roughly compensate for the lip perturbation. However, a finer-grain analysis seems necessary if one looks at the speakers' strategy within the adaptation procedure. Indeed, the large formant pattern variability observed from trial to trial suggest that speakers, in spite of a correct identification score, were looking for a suitable articulatory configuration likely to produce a "better" quality [u].

THE EVALUATION QUALITY TEST

Method and procedure

To test such a hypothesis, we realised a second perceptual test the aim of which was to evaluate the vowel quality produced in the lip-tube condition for each speaker. For this objective, listeners had to rate the quality of the [u] on a scale from 1 to 7 (1 = very bad [u], 7 = very good [u]). The corpus was made of the three [u] of each speaker produced in conditions N, PL, P2 so a total of 33 stimuli have been used. Eighteen listeners participated to this second perceptual test, among whom sixteen had participated to the first one. The same procedure as in the first test was used. For this test, listeners randomly listened to each stimulus five times.

Results

First of all, the analysis made on the [u] produced in condition N shows that all the average rate are higher than 5 except for one speaker (JM; 4,09). Starting from this, we consider that a stimulus having an average equal or higher than 5 will be considered as a sound with the desired perceptual quality.

The average rating between listeners for all stimuli in the N and PL conditions are plotted in figure 1 in the F1/(F2-F0) plane. All frequencies are in Bark [2, 3], and a normalisation by F0 is adopted for F2 in reference to Traunmüller's studies [4].

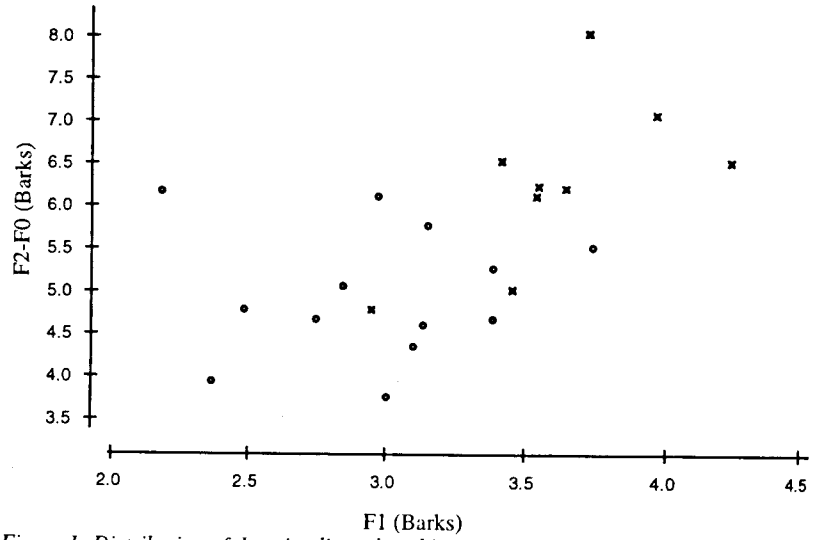


Figure 1: Distribution of the stimuli produced in the N and PL conditions for the eleven speakers. The stimuli having an average rate greater than 5 are represented by circle and those having an average rate smaller are represented by cross.

From this kind of representation, it is possible to propose a [u]-region within which a vowel has the required perceptual quality: almost all stimuli having an average rating greater than 5 are located in a region where F1 is smaller than 4 Barks and F2-F0 is smaller than 6 barks. The remaining well-rated stimuli correspond to F2-F0 slightly greater than 6 barks and F1 smaller than 3 barks. Note that stimuli produced in the PL condition by speakers CH, MP and GA are in this [u]-region. This tends to demonstrate that these 3 speakers have achieved the compensation in the perturbed condition. Note also that in this [u]-region 2 stimuli (JM in condition N and OD in condition PL) have an average rating less than 5. Of particular interest is speaker OD's production in the PL case: he produced a backward movement of the tongue leading to a complete compensation in the acoustic plane (attested by the position in the [u]-region); however, the perceptual effect is not sufficient (rate: 3.71).

Interpretation

These results shed light on the different strategies which are possible to compensate for the perturbation. Speaker CH did not produce a large reorganisation of the usual articulatory configuration for vowel [u]. Hence, his formant F2 remained fairly high, but this was compensated by a high F0 value leading to a correct F2-F0 value.

Speakers GA and in a lesser extent MP moved the tongue backwards inducing a decrease of F2-F0. This decrease was reinforced by a F0 increase though not sufficiently, but a low F1 value allowed to reach the desired perceptual quality.

As mentioned above, speakers JM and OD presented singular characteristics. In the case of speaker JM in the normal condition, the low [u] quality in spite of correct F1 and F2-F0 values show that one must consider all the spectrum to understand the perceptual effect of a sound. However, we shall concentrate further on the perturbed condition.

The case of speaker OD is more interesting. His F1/F2-F0 patterns in N and PL conditions are quasi-identical, but the perceptual evaluation falls down from 6.67 for condition N to 3.71 for condition PL. The F3 value provides the only significant difference between both conditions. This difference is then likely to play a decisive role in the perceptual evaluation. This hypothesis was verified with a simple perceptual test, in which the stimuli N and PL of this speaker were evaluated with and without a low-pass filtering at 1500 Hz. Results showed that in the PL condition the perception was better when the stimulus was low-pass filtered.

But if the perceptual product was insufficient, why did this speaker produce such a backward movement leading to a strong change of the tongue shape? The answer is given by the results of a last perceptual test, the aim of which was to understand speaker OD's behaviour through trials during the adaptation procedure. Results showed that at the beginning of the adaptation procedure, the [u] was confused with [œ] whereas at the end, the confusion was with [o]. Thus, the extent of the backward movement of the tongue during the adaptation procedure had a perceptually relevant effect: it allowed to reach, in the PL condition, a clear categorisation of the perturbed [u] as a velar vowel whereas before adaptation, confusion was possible between a central and a velar vowel.

CONCLUSION

First, this perceptual study confirms the importance of the perceptual goal in speech production. Speakers seem to have a clear representation of this goal and act in general in the right way to reach this goal in spite of the perturbation.

Moreover, this study proposes some interesting data useful for the understanding of this perceptual goal. It appears that in the F1/(F2-F0) plane, we can propose a [u]-region in the acoustic space within which the produced sound is identified as a "good" [u].

Finally, the study shows that in such a complex task where the perturbation imposes a complete reorganisation of the articulatory gestures, one can find: (1) subjects who do not need to compensate thanks to a correct initial configuration (speaker CH); (2) subjects who need to compensate and appear to know enough about the articulatory-perceptual relationships to be able to perceptually compensate up to a certain extent (speaker OD); and (3) subjects who should compensate and appear not to be able to do so (for example, speaker YP).

ACKNOWLEDGEMENT

The authors are sincerely grateful to all speakers and listeners for their contribution to these experiments.

This work is supported by the Esprit Basic Research Project n° 6975, Speech Maps.

REFERENCES

[1] Savariaux C., Perrier P. and Orliaguet J.-P. (in press), "Compensation strategies for the perturbation of the rounded vowel [u] using a lip-tube: A study of the control space in speech production", *Journal of the Acoustical Society of America.*

[2] Schroeder M. R., Atal B. S. and Hall J. L. (1979), "Objective measure of certain speech signal degradations based on masking properties of human auditory perception", in B.E.F. Linblom and S.E.G Ohman (Eds.), *Frontiers of Speech Communication Research*, (pp. 217-229). London: Academic Press.

[3] Zwicker E. and Feldtkeller R. (1981), *Psychoacoustique. L'oreille récepteur de l'information*, traduction française par C. Sorin, Paris: Masson et CNET-ENST.

[4] Traunmüller H. (1981), "Perceptual dimension of openness in vowels", *Journal of the Acoustical Society of America*, Vol. 69, pp. 1465-1475.

RECONSTRUCTION OF BASE FORMS IN PERCEPTION OF CASUAL SPEECH

Linda Shockey, Linguistic Science
Anthony Watkins, Psychology
University of Reading, UK

ABSTRACT

Perception of normal, relaxed speech involves relating phonologically reduced forms to their mental representations (assuming that lexical storage does not simply involve making a list of all pronunciations of all base forms). Many researchers into how the lexicon is accessed assume that word recognition normally occurs within the phonological boundary of the word being processed and that therfore it is only in exceptional cases that a decision about the identity of a word is postponed. The research reported here suggests that such delayed recognition may be a very commonly-used strategy for understanding of conversational speech forms.

THE EFFECT OF CASUAL SPEECH PHONOLOGY ON PRONUNCIATION

It has been demonstrated [4,6] that there is phonetic reduction in words which have once been focal but have since passed to a lower information status: the first time a word is used, its articulation is more precise and the resulting acoustic signal more distinct than in subse- quent tokens of the same word. By 'phonetic' we mean that the effect can be described in terms of of vocal tract inertia and ease of articulation: since the topic is known, it is not necessary to make the effort to achieve a maximal pronunciation after the first token. We expect the same to happen in all languages, though there may be differences of degree.

Phonetic effects are not the only ones which one finds in relaxed, connected speech: there are also language-specific reductions which occur in predictable environments and which appear to be controlled by cognitive mechanisms rather than by physical ones. These we term phonological reductions because they seem to be part of the linguistic plan of a particular language. While they may not make a change in meaning, they contribute to acceptable relaxed pronunciation. They help to make a native speaker sound native. Among these in English are effects such as changing /t/ to [ʔ] before another consonant in syllable-final position, as in "hatbox" pronounced [hæʔbɒks].

Casual speech processes cause changes in everyday conversational speech which make some of the forms found quite different from their dictionary representation or "citation form." They can, for example, cause ambiguities: the distinction between /n/ and /m/ is often not observed before bilabial consonants. This means that "screen play" and "scream play" are often not pronounced differently.

More extreme differences are possible: the word "handbag" is often pronounced "hambag" The /d/ is deleted or suppressed, and the /n/ which remains changes to match the following /b/, as in the example above. The word "can't" is often pronounced [kɒʔ], without the [n], and with the final /t/ changed to a glottal stop

Phonological effects are common in casual speech, but some models of speech perception (e.g. [2]) assume fully-specified input which is processed in a linear order: there are no segments absent from the signal (though overlap of gestures can occur), nor are there any segments which are not present in the phonemic inventory of the language, but which appear as the result of phonological processes, such as the nasalised [ɒ] in [kɒ̃ʔ].

Some researchers [6,7] have begun to explore the changes which will have to be made in lexical access models in order to accommodate phonological variation, and this paper is a contribution to that explorationl.

Sequential lexical access

It is believed that "Listeners generally recognise words before hearing them completely," [9] . A special case is made for homonyms, which have to be disambiguated by following information [1].

But personal experience tells us that it is quite possible to revise our notion of what was heard based on subsequent information, especially when we are listening under unfelicitous conditions, e.g. to a foreign language with which we are only adequately familiar, to our own language in a noisy environment or even to gated sentences. Experimental work by Grosjean [3] and Bard et al [1] supports this intuition.

We hope to do a series of studies aimed at finding out how ambiguities caused by phonological reduction are resolved by listeners, and how, in general, reduced forms are related to the fully-articulated forms which (presumably) constitute entries in the mental lexicon. We assume that the scope of material used to unravel these reductions varies with the degree of reduction: as the phonetic information becomes less dependable, more semantic information is called for. We also suspect that subjects vary a great deal in the extent to which they depend on one or the other of these sources.

THE PILOT EXPERIMENT

The following sentence was produced by the experimenter and recorded digitally:

The screen play didn't resemble the book at all.

The word "screen" was pronounced with a final **m**, as is normally done in this environment in unself-conscious speech.

The sentence was then presented to 10 subjects using a gating technique. There were 33 equal gates, beginning in the middle of the vowel of "screen." All of the subjects originally judged the first word in the sentence to be "scream." When the segment (**p**) which is the motivation for the /n/ to /m/ change was presented, three of the subjects reversed their judgement from /m/ to /n/. Even though this is subsequent to the end of the word, it is easily explained by extremely local factors: the notion of 'underspecification' [7] could explain this result. Eight subjects had changed [m] to /n/ by the end of the word "play," but two withheld judgement on the /m/ or /n/ decision until clearer, non-phonological, information was available from other sources.

The word "didn't" was even more reduced, (to [dĩn]) and here there was even greater evidence of late recognition: six opted for "didn't" at various stages in the word "resemble," and two subjects couldn't commit themselves until after "book" was recognised.

It is thus clear that here the identity of the lexical item was not resolved before its end. Further, it seems that it was not resolved purely through phonological knowledge, though an implicit knowledge of the sorts of reductions permissible in English was called for.

A similar experiment using the sentence:

The scream play was part of primal therapy

showed that many subjects perform an [m] to /n/ transformation on this case also, even though it is inappropriate, and that this transformation must be reversed by subsequent semantic input for the sentence to be understood. This suggests both that the influence of the phonetic conditioning factor is very strong and that reversal of a decision based on subsequent information must be a robust part of our linguistic competence. As Warren [10] has it, "...succesful speech perception cannot proceed as a Markovian process, with perception occurring first on lower and then higher levels of organisation. Processing of this nature does not benefit fully from the redundancy of the message and does not permit the correction of mistakes."

This pilot experiment will be followed by others which explore whether other conversational speech processes (such as tapping, palatalisation, and devoicing) are perceived similarly and the consequences for models of word recognition.

Bibliography

[1] Bard, E.G., Shillcock, R., 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, pp. 395 - 408

[2] Cole, R. and Jakimik, V., (1979) "A Model of Speech Perception" in R. Cole (Ed.), *Perception and Production of Fluent Speech*, Erlbaum

[3] Grosjean, François, (1985) "The Recognition of Words After Their Acoustic Offset: Evidence and Implications," *Perception and Psychophysics* 38(4), pp. 299 - 310

[4] Fowler, C.A. and Housum, J, (1987), "Talkers' Signalling of "New and "Old" Words in Speech.....," *Journal of Memory and Language* 26, pp. 489 - 504

[5] Frauenfelder, U.H. and Tyler, L.K., (1987) "The Process of Spoken Word Recognition: an Introduction," *Cognition* 25, pp. 1 - 20

[6] Frauenfelder, U.H. and Lahiri, A., (1989), "Understanding Words and Word Recognition: Does Phonology Help?" Chapter 10 in W. Marslen-Wilson (ed), *Lexical Representation and Process*, MIT Press

[7] Lahiri and Marslen-Wilson, (1991), "The Mental Representation of Lexical Form: A Phonological Approach to the Recognition Lexicon" *Cognition* 38: pp. 245 - 294

[8] Lieberman, P., (1963) "Some Effects of Semantic and Grammatical Context on the Production and Perception of Speech," *Language and Speech* 6, pp. 172 - 187

[9] Marslen-Wilson, W.D., (1987) "Functional Parallelism in Spoken Word Recognition," *Cognition 25*, 1987, 71 - 102

[10] Warren, Richard, (1982), *Auditory Perception, A New Synthesis*, Pergamon

EVIDENCE FOR DIRECT LEXICAL ACCESS FROM RESPONSE TIME EXPERIMENTS

Kari Suomi
Deparment of Finnish, Saami and Logopedics, University of Oulu, Finland

ABSTRACT

Monitoring experiments are reported that compared response times (RTs) to three target types in Finnish: whole words, word-final syllables, and word-final phonemes. Care was taken to ensure that the targets could not be responded to on the basis of only partial analysis of the stimuli. Throughout, whole words were detected faster than coterminous syllables and phonemes, suggesting that words are not recognised through intermediate phonemic or syllabic representations.

INTRODUCTION

Both the idea of lexical access through compulsory intermediate phonemic and/or syllabic representations and the idea of direct access postulating some other kind of sound representation to mediate between the sensory input and the lexicon have support in the literature on spoken word recognition (see e.g. [1], [2]). Measurement of RTs to monitoring targets, a paradigm in which shorter RTs are interpreted to indicate earlier on-line processing, is a tool that, potentially at least, could be profitably used to inquire whether lexical access is direct or not. To my knowledge, however, the paradigm has never been used to explicitly address this question. The virtual lack of reported comparisons of RTs to words and their phonological constituents in the last two decades seems to be due, in part, to a preconception among some researchers who use this paradigm to study lexical access that there is no alternative to access involving intermediate phonological units and representations. The conclusions of McNeill & Lindig [3] seem to have been influential in shaping the preconception, these authors claiming that that minimum RT in target monitoring experiments occurs whenever the linguistic level of the target and the search list is the same, and since the level where the target and the list match is entirely determined by the experimental design, it is no possible that monitoring experiments can reveal the perceptual units of speech. However, these widely cited conclusions are not warranted by the authors' experiments that have been shown to suffer from a number of methodological weaknesses that render the results highly unreliable. Thus, due to the way in which the stimuli were constructed, subjects were able to base their responses on only the initial portions of the target-matching stretches of the stimuli (see [4]), they were in fact urged to do so, and the experiments included conditions involving what the authors call *downward search* in which subjects were required e.g. to detect target sentences and target words in search lists consisting of syllables and phonemes (sic!) (see [5], which also contains a more detailed account of the present experiments).

In brief, there is no evidence from target monitoring studies that would force the conclusion that phonemes and/or syllables must be identified before a word is accessed and recognised. Prompted by a desire to test the DAPHO model [6] that postulates one version of direct access, the present experiments were designed to measure RTs to words, syllables and phonemes under as comparable bottom-up conditions as possible. Each target-bearing or target-constituting stimulus word contained all three target types. E.g., RTs were measured to each of the targets "PALKKI", "KI" and "I" in the stimulus word *palkki*. In a given stimulus word, the three target types were all coterminous, and thus the time course of how subjects were exposed to the distinguishing auditory information in the stimulus was exactly the same for each target type. And since RTs were always measured from exactly the same temporal location in a given experimental word for each target type, any systematic differences observed in RTs to these targets must be due to differences in the central processing of simultaneously available peripheral input.

PROCEDURE

Experiment 1 is described in some detail below, but for experiments 2 and 3 only major deviations from the procedure of experiment 1 are indicated.

Experiment 1

In experiment 1, the target-carrying stimuli were a set of 36 disyllabic words, each occurring in a list containing from three to six words. In addition, subjects were presented 10 practice word lists at the beginning of the test the responses to which were ignored, and also, dispersed among the experimental lists, 18 no-response distractor lists and 9 filler lists. All subjects heard exactly the same stimulus material. The target-carrying stimuli were chosen in 12 triplets so that, within each triplet, all three words had a phonemically identical second syllable, and the first syllable of each word had the same general structure in terms of the C and V class affiliation of its segments. A further requirement for a word to be included in a triplet was that at least one further familiar word must exist that diverges from the experimental word with respect the the final phoneme alone, to guarantee that the uniqueness point of the experimental words was not reached until the portion corresponding to the final phoneme.

Each word in each of such highly controlled triplets functioned as carrier of each of the target types Word, Syllable and Phoneme but in three different, rotated target conditions. The target conditions were rotated in such a way that, for a given carrier word, subjects in one condition were given a word target, those in a second condition a syllable target, and those in a third condition a phoneme target. Target assignments were balanced across the conditions so that each triplet yielded three instances of RTs to each target type. Consequently, the RTs to the three target types to be reported were obtained using exactly the same set of words.

In the no-response distractor lists the Word, Syllable and Phoneme targets were similarly rotated, but the Word target specified for a list did not occur in that list. Instead, the list contained a word that deviated from the specified Word target by the last phoneme only. E.g., one such list had the specified targets "HELMA", "MA" and "A", and the list consisted of the words *kuori kuusi potti rove helmi tossu*, in which the penultimate word is the intended distractor. Thus in each no-response list, the distractor conditions were exactly the same for the three target types, and the appearance of finally-diverging distractors in the Word target condition should induce subjects to respond only after a complete analysis of the stimulus words. Subjects should not respond to the distractor lists if they were reacting accurately, and therefore subject reacting to more than a predetermined number of such lists were discarded. 27 of the 30 tested subjects were accepted.

Individual subjects were seated before a computer terminal, and the lists were presented through earphones. Subjects were told that they would hear word lists and that their task was to monitor for whole-word targets, targets consisting of a consonant-vowel sequence, or vowel targets, and they were instructed to press the space key as soon as they were certain that they had heard the target valid for a given word list. Before each new word list, an alert beep was sounded and the (fully phonemic) written target specification appeared on the screen where it stayed 2.5 seconds, after which the list was heard.

For each target-carrying word, the raw RTs were measured from the estimated onset of the final vowel, but the raw values were adjusted to give RTs from the common acoustic end point of the three target types.

Experiment 2

In experiment 1 vowel-final disyllabic real words were used as stimuli, whereas in experiment 2 phonologically well-formed nonsense items were used, to allow for more variable yet native-like structural patterns. Half of the items were disyllabic, half trisyllabic, and within each group, half were vowel-final, half consonant-final. Nonsense items are also insensitive to word frequency effects which were not completely controlled in experiment 1. All subjects again heard exactly the same stimulus material, and the 48 stimuli

carrying the three target types were chosen in 16 triplets following the same criteria as in experiment 1. Subjects were instructed to treat the whole-item nonsense targets as novel words, e.g. as names of new products. Otherwise, the procedure was as in experiment 1, including the use of finally-diverging distractor stimuli and rotation of the target types. There were 24 accepted subjects.

Experiment 3

Experiments 1 and 2 required detection of targets in lists of real and nonsense items that were separated by pauses, and such lists may favour the detection of whole-word targets because the input has already been segmented into stretches that correspond exactly to the target units, whereas the onsets of the phoneme and syllable targets have to be located in coarticulated speech within the stimuli. Therefore, experiment 3 was conducted in which all target types had to be segmented from both preceding and following continuous speech. Experiment 3 contained the same target-carrying words as experiment 1, but this time embedded in short sentences that were semantically fully neutral with respect to the probability of occurrence of either the specified Word target or its implicit, finally-diverging lexical competitor(s). 21 subjects were accepted.

RESULTS

The mean RTs observed in experiments 1-3 are shown in Tables 1-3.

Table 1. Mean RTs (in ms) to detect the target types Word, Syllable and Phoneme in disyllabic vowel-final real words in experiment 1.

Wrd	Syl	Pho	Mean
173	271	314	253

In experiment 1, whole words were detected about 100 ms faster than final syllables, which were in turn detected about 40 ms faster than final phonemes; both of these differences were significant.

Experiment 2 replicated the major results of experiment 1. Thus while responses to consonantal Phoneme targets and consonant-final Syllable and Word targets were faster than responses to vocalic or vowel-final targets, RTs to Word targets were again faster than those to the phonological targets irrespective of the type of final segment, and final syllables were again detected faster than final phonemes (all these differences were significant).

Table 2. Mean RTs to detect the target types Word, Syllable and Phoneme as a function of target-final segment class in nonsense items in experiment 2.

Final segment class	Target type			
	Wrd	Syl	Pho	Mean
Vowel	175	301	344	273
Consonant	156	260	285	234
Mean	165	280	314	253

The results of experiment 3 replicate the major finding: Whole words were detected faster than final syllables and final phonemes, even when also whole words had to be segmented from continuous speech. In experiment 3 all mean RTs are on average about 70 ms longer than in experiment 1; this increase may be due to a greater difficulty of performing target monitoring in material that is semantically coherent.

Table 3. Mean RTs to detect the target types Word, Syllable and Phoneme in disyllabic vowel-final real words in experiment 3.

Wrd	Syl	Pho	Mean
263	340	370	324

DISCUSSION

Against the background that both lexical access through necessary intermediate phonemic and/or syllabic representations and direct lexical access using some alternative sound representations find ample support among researchers of spoken word recognition, the present

series of experiments set out to investigate response times to three acoustically coterminous target types, namely whole words, word-final syllables, and word-final phonemes, in an attempt to distinguish between the rival broad views of lexical access. The results indicate that whole words were detected before their final syllables and final phonemes even when the words were lexically non-unique prior to the last phoneme, and when the possibility of responding on the basis of guessing was eliminated by stringent distractor conditions. The whole-word advantage was observed in experiments whose materials jointly contained variable and phonotactically representative targets of each type, it was observed with real words as well as pseudowords, with words separated from others by pauses and semantic incohesiveness, and with words in connected speech in meaningful sentences. Differences in the manner in which response times and other temporal data on on-line speech behaviour have been measured and reported make it impossible to compare the present results with previous ones, but the temporal distances here observed between the end of a word and the detection of that word are not inconsistent with the intuitive immediateness with which words seem to be recognised outside the laboratory.

On the assumption that shorter RTs reflect earlier processing, I interpret the results as support for the idea that lexical access and word recognition are direct in the sense that they do not involve compulsory intermediate levels of representation in terms of phonemes or syllables. There is no direct evidence from monitoring studies against direct access in any language, e. g. findings to the effect that RTs to whole words are longer than those to their constituent phonemes or syllables, and consequently there is no principled reason for dismissing the present results as specific to Finnish alone (which nevertheless remains a testable possibility).

A counterargument against the above interpretation that I have come across is that the results are most probably irrelevant to the question of the nature of the sound representation involved in lexical access because word detection involves identifying a familiar unit whereas syllable or phoneme detection does not, and therefore, even if phonological units were used implicitly to identify words, it does not necessarily follow that they would be detected faster than words, in an explicit detection task. If this argument is taken as a sufficient explanation of the observed detection advantage of whole-word targets, then obviously the conclusion follows that it is *a priori* impossible to distinguish between lexical access through intermediate units and direct access using the target monitoring paradigm, because the familiar-unfamiliar distinction can always be invoked to annihilate any data that seemingly support direct access. But if phonemes and/or syllables are regularly and compulsorily identified prior to word recognition, can they really be characterised as unfamiliar units, especially in comparison to pseudowords as used in experiment 2? And if a familiar-unfamiliar effect is operative, does its magnitude fully account for the observed word-advantage in response times?

REFERENCES

[1] Pisoni, D. & Luce, P. (1987), "Acoustic-phonetic representations in word recognition". In *Spoken word recognition* (U. Frauenfelder & L. Tyler, editors), pp. 21-52. Cambridge, MA: The MIT Press.

[2] Klatt, D. (1979), "Speech perception: A model of acoustic-phonetic analysis and lexical access", *Journal of Phonetics*, vol. 7, pp. 279-312.

[3] McNeill, D. & Lindig, K. (1973), "The perceptual reality of phonemes, syllables, words and sentences", *Journal of Verbal Learning and Verbal Behavior*, vol. 12, pp. 419-430.

[4] Norris, D. & Cutler, A. (1988), "The relative accessibility of phonemes and syllables", *Perception & Psychophysics*, vol. 43, pp. 541-550.

[5] Suomi, K. (submitted), "Evidence of word recognition using direct lexical access without phonemes or syllables", a paper submitted for publication.

[6] Suomi, K. (1993), "An outline of a developmental model of adult phonological organization and behaviour", *Journal of Phonetics*, vol. 21, pp. 29-60.

ROLE OF THE GRAMMATICAL GENDER IN MENTAL LEXICON ACCESS

*S. Monpiou, **M.-N. Metz-Lutz, *F. Wioland, *G. Brock
*Institut de Phonétique de Strasbourg
22 rue Descartes 67 084 strasbourg, France
monpiou@ushs.u-strasbg.fr
**INSERM U398
Hôpitaux Universitaires de Strasbourg - Clinique Neurologique
67 091 Strasbourg

ABSTRACT

The aim of the present sudy is to evaluate the effect of the grammatical gender in word recognition for French. The focus is on auditory word identification when words are presented within a restricted context, *i.e.* preceded only by the French definite article, singular: "le/la". In this particular context one may hypothesize that this syntactic information could be considered as a prime for accessing the mental lexicon.

INTRODUCTION

Spoken language comprehension is a process of high complexity involving a great number of different processing levels. Among these levels, there is one of particular importance: the level of word identification and recognition. Lexical recognition relies on the mapping of information extracted from the verbal stimulus with a particular representation present in the mental lexicon. Consequently, word recognition consists in accessing the mental lexicon in which representations of the constituents of a given language are stored in long term memory for all speakers of the language.

Models have been elaborated in order to describe the internal mental access process. A good amount of these models posit that recognition of an acoustic element is based on two types of information: acoustic-phonetic information furnished by the input itself and information present in the context of the input. Generally, recognition of words presented auditorily is triggered by acoustic-phonetic information, also called bottom-up information; contextual or top-down information intervene later. Such models, in particular the Cohort model — whose first version was elaborated by Marslen-Wilson and Welsh, based on Morton's model of Logogenes and on Forster's model — aim at accounting for lexical recognition processes regardless of language specificity. However, languages have different organizational structures. Cutler and colleagues [1] have demonstrated that French and English subjects behave differently in segmenting verbal units when recognizing them and that this difference in behaviour is linked to the structure of the language. From such results, it is judicious to think that the relevance of contextual information used in the process of internal lexicon access, depends, to a certain degree, on the structural organization specific to the language. Thus information furnished by the context would be used differently depending on particular linguistic structures of languages which consequently would develop their own routines in the word recognition process [1]. So, one would state that in languages where gender is part of their lexical organization, this linguistic specificity should play a critical role in the mental lexicon access process. The purpose of the gender in a language is to arrange referential elements (nouns and pronouns) into lexical classes according to the formal characteristics encountered by these elements in a sentence. All lexical classes have a particular characteristic or *trait de genre* (gender feature). According to Renault [2] this gender feature does not appear in the lexical unit but in a unit associated with it. In Romance languages, the gender feature is represented by an element that precedes the lexical base, *i.e.* the determinant. From a psycholinguistic point of view, one may wonder to what extent does this contextual element influence access to the internal lexicon in a language such as French? Apart from Grosjean *et al.* [3], studies in this area are very rare, even though the gender in French is omnipresent: there is no substantive that does not possess the masculine or the feminine gender [4]. To bring some light to this question, we decided to study the role of the grammatical gender in structuring the internal lexicon and its influence as contextual information carried by the singular definite article, masculine or feminine (le/la), on the word recognition process. This problem will be addressed experimentally using lexical decision or related paradigms.

EXPERIMENTS

Experiment 1: lexical decision task

This experiment consisted of two tests comparing lexical decision for isolated words and non-words to lexical decision for words and non-words preceded by the French singular definite article "le/la". According to previous studies [5] [6] a faster lexical decision for items preceded by an article was expected.

Method:
Subjects: 19 native French speakers, all students participated in this experiment.
Materials: two different lists of 160 stimuli each were constructed. In each list, the words were mixed in equal number with non-words and presented in random order. All targets in the two lists were separated by a three second silence.
Procedure: The two lists were presented in random order and instructions were given between each list. Subjects were asked to answer as fast and as accurately as possible by pressing a decision-key.
Results: Errors and Reaction-Times (RTs) for correct responses were analysed. As RTs were measured from the onset of each item — in order to compare decision time in the two conditions — the artcile mean duration value ("la"=137 ms ; "le"=144 ms) was substracted from mean decision time to the sequences "article + real word" and "article + non-word". A three-way ANOVA on mean RTs with factors *condition*, *target type* and *gender* showed a significant main effect of *condition* $F(1,128) = 16,992$ $p = .0001$. There was no interaction effect between *condition*, *target type* and *gender*. This means that, a target preceded by the article is recognized faster than a target in isolation. However, would such a priming effect, related to the syntactic information carried by the article, reflect a specific organization of the internal lexicon? To answer this question we elaborated a second experiment.

Experiment 2: gender decision task

This experiment concerned real isolated words: 80 masculine words and 80 feminine ones mixed in a list. The task was to determine the grammatical gender of the targets by pressing a button "masculine" or "feminine".

If the mental lexicon is organized according to word grammatical gender, the gender decision would be performed within the time of lexical decision for isolated lexical items. If the noun grammatical gender is explicitly marked in the mental lexicon, then the gender decision would be longer than the lexical decision for isolated words.

Method: The method is similar to those of the first experiment.
Results: We compared RTs of the sequences "article + real words" (with the article duration mean value substracted) of the second list of the first experiment with RTs obtained in this experiment. A three-way ANOVA on mean RTs taking *condition*, *target type* and *gender* as the variables showed a significant effect of *condition* $F(1,140) = 24,888$ $p = .0001$. The word recognition task is faster than the gender identification one. According to this result, it can be supposed that

mental representations would be ordered in the mental lexicon according to explicit marks. Consequently the grammatical gender decision would be post-lexical. In order to verify this hypothesis a third experiment was designed.

Experiment 3: compatibility decision task

160 "article + word" stimuli were used: for 80 out of them the gender of the article did not correspond to that of the following word (e.g. *le maison instead of "la maison"). This decision task consisted in deciding if in a noun phrase the gender of the article was compatible with the grammatical gender of the following word.

If grammatical gender figures in the mental lexicon, one might suppose that compatibility decision would occur at the same time with lexical decision as regards sequences "article + real word".

Method: the method is similar to that of the two preceding experiments.

Results: Only sequences "article + real words" of the second list in the first experiment and the gender compatible sequences "article + real word" of this experiment were used. Analyses of variance with three factors (*target type* ; *condition* and *gender*) showed a significant effect of *condition* $F(1,140) = 17,42$ $p = .0001$. So compatibility decision is significantly longer than lexical decision.

DISCUSSION AND CONCLUSION

The various results obtained in the three experiments support the claim that the gender plays an active role both in the mental lexicon access and in its organization.

According to results from the first list of the first experiment it is suggested that the processing of a phonological string is the same whether it belongs to the lexicon or not. There is no significant difference in RTs between recognizing an isolated word or an isolated non-word (RTs words = 929 ms; RTs non-words = 1016 ms). Compared to normal verbal communication, in which all elements emitted are supposedly endowed with meaning and would thus help in the recognition of subsequent units, in the first experimental condition there was no contextual cues that may facilitate word recognition process. Thus subjects developed a strategy different from those used in everyday communication. Maximum acoustic-phonetic information to make decision about the stimuli was collected. So words and non-words in isolation are recognized well after their uniqueness point [7] and their deviation point [8] respectively. When such results are analyzed within the Cohort II model [9], it could be suggested that decisions about isolated units can only be made after perfect matching (in the case of real words) between the acoustic input and a mental representation.

Results obtained for real words in the two tests of the first experiment show that the presence of a gender element before the word, significantly influences the process of internal lexical access by accelerating it (RTs for words of list n°1 = 929 ms; RTs for words of list n°2 = 803 ms). According to Marslen-Wilson [9], in the Cohort II model, top-down information, *i.e.* contextual information, would have facilitating effects on lexical access. Facilitating effects would correspond to an increase in activation level of certain candidates compared with others. In this perspective, it can be supposed that the role of the definite article as contextual information, would be that of increasing the activation level for candidates of the initial cohort that share the same gender as that of the definite article. This selective activation would contribute in reducing the number of potential candidates that would fit the context and thus reduce the time needed for lexical recognition. In the case of isolated words, there are no contextual cues that may reduce the initial cohort.

Results from the gender decision task suggest that gender decision is post-lexical as it occurs later compared with the time needed for the recognition of a word preceded by the definite article (RTs for words in experiment n°2 = 927 ms; RTs for words in list n°2 = 803 ms). Consequently one may conclude that access to word gender can be operational only when word access has applied. Thus the strategy used by subjects is not only that of an acoustic-phonetic analysis of the input but more so that of an active search in the lexicon. Such a lexical strategy is strongly influenced by the nature of the task: subjects knew that they only had to deal with words.

Compatibility decision is more complex than lexical decision. RTs obtained in the third experiment are significantly longer (1049 ms) than RTs concerning word recognition in context (943 ms). Such results can be accounted for within the Cohort II model. From the task required of subjects, they knew that the article and the word could sometimes be incompatible in gender. It follows that the influence of the context, *i.e.* of the definite article, was neutralized by the experimental paradigm. Not a single representation present in the initial cohort could have an activation level higher than an other with regards to its gender, since the input article may not correspond to the word that follows. Thus the effect of the grammatical gender is no longer absolute but rather more *probabilistic* [3].

Following these findings, one may posit that the grammatical gender intervenes in the structure of the mental lexicon by ordering representations in relation to their gender. The mental lexicon would thus ressemble a dictionary, solely composed of words accompanied by their gender marking.If such is the structure of the internal French lexicon it can be hypothesized that the organization of the mental lexicon is not universal but specific to the language it belongs to. From such a perspective, access processes would not be universal but determined by the structural characteristics of every language.

To carry out a more thorough analysis of this hypothesis, it should be worthwhile examining the mental lexicon access for languages that have a grammatical gender ressembling that of the French language, but also for those that have other types of gender.

REFERENCES

[1] Cutler, A., Mehler, J., Norris, D., Segui J. (1986), "The syllable's differing role in the segmentation of French and English.", *Journal of Memory and Language*, 25, pp.385-400.

[2] Renault, R. (1987), "Genre grammatical et typologie linguistique.", *Bulletin de la Société de Linguistique de Paris*,82 , 1, pp. 69-117.

[3] Grosjean, F., Dommergues, J.Y., Cornu, E. (1992), "The gender marking effect in spoken word recognition", *Perception & Psychophysics*, 56 (6), pp. 590-598.

[4] Desrochers, A. (1986), "Genre grammatical et classification nominale.", *Revue Canadienne de Psychologie*, 40 (3), pp. 224-250.

[5] Huckel, C. (1991) "De l'importance du genre dans la décision lexicale". *DEA (pre-doctoral dissertation), Université des Sciences Humaines de Strasbourg.*

[6] Monpiou, S. (1992), "Contribution du genre - porté par l'article défini - dans la reconnaissance auditive des mots". *Masters dissertation, Université des Sciences Humaines de Strasbourg.*

[7] Marslen-Wilson, W. (1984), Function and process in spoken word recognition. A tutorial review. In H. Bouma and D.G. Bouwhuis (eds) Attention and Performance X: Control of language processes, Hillsdale, NJ: L. Erlbaum, pp. 125-150.

[8] Radeau, M., Mousty P., Bertelson P (1989), "The effect of the uniqueness point in spoken word recognition.", *Psychological Research*, 51, pp.123-128.

[9] Marslen-Wilson, W. (1987), "Functionnal parallelism in spoken word-recognition.", *Cognition*, 8, pp.1-71.

MENTAL LEXICON ACCESS: INITIAL PHONOLOGICAL & MORPHEMIC SYLLABLE IN AUDITORY RECOGNITION

P-Y. Connan*, M-N. Metz-Lutz, F. Wioland*, G. Brock***
**Institut de Phonétique, ERS 125 CNRS*
Université des Sciences Humaines de Strasbourg - FRANCE
***Clinique Neurologique, INSERM U398*
Hôpitaux Universitaires de Strasbourg - FRANCE

ABSTRACT

This study deals with an auditory lexical decision task that should enable a better understanding of auditory word recognition. Interactive processes of perception and comprehension of spoken language are examined. A major question addressed here, is to find out if auditory word recognition is facilitated when a word or a non-word prime or target share the same initial sequence, whose status, whether phonological or morphemic, may change access conditions to the mental lexicon.

Results show a lack of a phonological priming effect and the specific status of the initial morphemic syllable as a factor that facilitates lexical decision. The data are discussed in relation to lexical recognition models such as the Cohort Theory. They differ from results usually obtained in the domain.

INTRODUCTION

Spoken word recognition is an extremely complexe phenomenon. It is plausible to assume the existence of a mental lexicon, whose components are accessible through different mental processes. This complexity is due, on one hand, to the numerous steps and functions (access, selection and integration, as described in many recognition models) that constitute the spoken language processing; on the other hand, to multiple relationships that exist within the different units of mental representations: phonological, morphological, syntactic and semantic dimensions can interact at different levels and time in these processes.

A good amount of studies and interactive models — like the Cohort model [1]— have shown the priority of acoustic-phonetic analysis of the incoming speech signal (bottom-up information) and the importance of acoustic features of word onsets during the access phase [2]. Such a model suggests that this initial sensory input could activate the representation of the signal itself, as well as all other words with common properties.

Thus, if word onsets «do have special status in the lexical access of spoken words» [2], a major question concernes the nature of the onsets of verbal sequences: do morphemic structure (prefixed words for e.g.) undergo a different processing during word recognition processes, compared with words sharing a similar syllabic onset? In other terms, could morphemic relationships be represented explicitly in the mental lexicon.

Several studies have dealt with this morphemic structure but largely using visually presented material. It has been shown [3] that prefixided and non-prefixided words (but not pseudo-prefixed words whose mean decision times are longer) are processed equally rapidly, «indicating that a decompositional process (left to right) is efficient». Interesting results in both visual and auditory modalities [4] have been attested, to demonstrate that «prefixed words are recognized, after the prefix has been removed, via a representation of their stem». Moreover, certain authors [5] tried to distinguish morphemic relationships from semantic and formal ones. The question, now, is to know if significant effects are just «a convergence of semantic, orthographic and phonological relationships». This study shows that morphemic priming is «a separate dimension along which two words can be related».

METHOD

With an aim to evaluate the role and the importance of word onset during the mental lexical recognition process, a lexical Decision task was used. For the two experimental conditions, subjects were required to listen to different pairs of Words (W) and/or Non-words (Nw) and to judge whether or not the second stimulus (the target) was a word of the French lexicon. The duration between W or Nw prime and W or Nw target was approximatively 400 ms.

Corpus

• Syllabic priming condition.
Prime-target pairs of this first priming condition consisted of 138 pairs of syllabic French Words and Non-words. Nw were created from real words by displacing one phoneme. For this first experimental condition, 8 different combinations were used (see Table 1), where the status (W or Nw) of the prime and target may change with the presence or not of a syllabic priming effect. For each combination, 15 different pairs of verbal sequences were presented.

• Morphemic priming condition.
For this second experiment also, 8 different combinations of Word and Non-word pairs (see Table 2), with a total of 96 bi or tri-syllabic pairs of verbal stimuli were used. Note that the morphemic structure, here, is the "prefixe".

Subjects

A group of 30 normal subjects participated in the experiment. This group was composed of the same number of male and female subjects. All subjects were, monolingual, French speakers and young adults, from 18 to 30 years old, chosen among volunteer students at the University of Strasbourg. None of them had hearing loss or neurological impairement. The same group of subjects participated in the two experimental conditions.

Procedure

A work-station for the lexical decision task was specially built to accumulate data with a 100% fiability. The stimuli were recorded on the first channel of a Tascam Tape and were presented through headphones. Subjects, a maximum of 3 at a time, were carried out the tests in a sound-proof anechoic room, and had to press, the most rapidly and the most exactly possible, two buttons (labelled "yes" & "no") on an individual board. On the second channel, a "target-impulse activated a millisecond counter, localized on a digital acquisition card of a microcomputer. This inaudible signal started exactly at the onset of the target. The counter was stopped when subjects pressed one of the buttons. Finally, target counting, Reaction Times calculation and file creation for later statistical analysis were executed by sofware.

Table 1. Combinations, examples & results for syllabic priming

Type	Combination	Resp.	Example	mean RTs (ms)
1	W-W syll.+	yes	galop - gamin	779,96
2	W-Nw syll.+	no	bouton - [buʒɛj]	936,72
3	W-W Ø	yes	sapin - propos	782,39
4	Nw-Nw Ø	no	[ʀakaj] - [guʒɛ̃]	914,14
5	Nw-W Ø	yes	[gafo] - milieu	772,26
6	Nw-Nw syll.+	no	[mupɛ̃] - [mutwaʀ]	928,98
7	Nw-W syll.+	yes	[fimɛl] - figure	791,76
8	W-Nw Ø	no	salon - [pʀode]	898,94

RESULTS

Mean error rates were calculated for each experiment, subject and types of combinations. For syllabic and morphemic priming conditions, a mean error rate of respectively 2,19% and 3,04% was found. No subject was eliminated due to a high error rate.

Separate statistical analyses of variance (one-way ANOVA) were conducted for all combination types in the two experimental conditions. We first compared mean RTs for "yes" and "no" responses and, as expected, significantly faster RTs for "yes" in both experiments were found ($p<0.05$).

In the first experiment, no significant effect ($p=ns$) of syllabic priming (Type 1 *vs.* Type 2) was found. The same result was found in Exp. n°2 (Type 5 *vs.* Type 3) in which the W-W Syllabic (pseudo-morpheme, here) condition, *vs.* W-W neutral condition were duplicated.

In the morphemic priming condition, statistical analyses revealed that mean RTs were significantly faster for W-W priming pairs than for neutral W-W pairs ($p<0.05$). On the contrary, no significant effect was found comparing W-W morphemic priming pairs *vs.* W-W syllabic priming pairs, intra and inter experimental conditions ($p=ns$).

Table 2. Combinations, examples & results for morphemic priming

Type	Combination	Resp.	Example	mean RTs (ms)
1	W-W morph.+	yes	déboucher - déranger	**808,36**
2	W-Nw Ø	no	convenu - [desãʒe]	**934,03**
3	W-W Ø	yes	travail - prévenir	**865,32**
4	Nw-Nw Ø	no	[vɛdaj] - [sivɔ̃]	**918,94**
5	W-W syll.+	yes	infantile - incapable	**827,80**
6	W-Nw syll.+	no	emporter - [ãtaʀke]	**994,96**
7	Nw-W Ø	yes	[ɛfape] - déranger	**830,50**
8	Nw-Nw syll.+	no	[ãsivɔl] - [ãʀimit]	**1036,59**

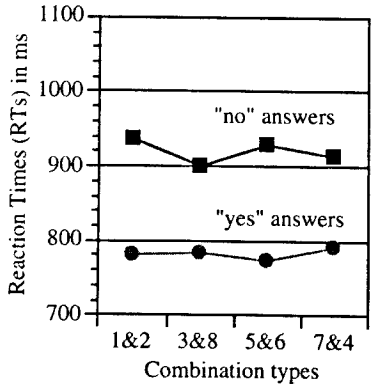


Figure 1. Results for syllabic priming

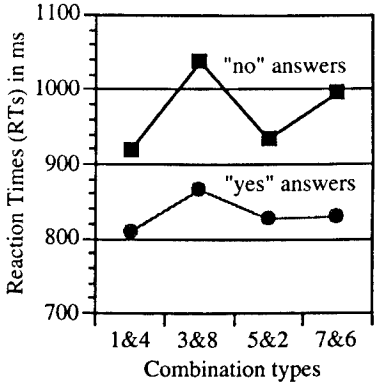


Figure 2. Results for morphemic priming

DISCUSSION AND CONCLUSION

Results obtained in the syllabic priming condition (*i.e.* a lack of phonological effect) confirm earlier findings [6]. A few other studies also invalidate, to a certain extent, one of the hypothesis of the Cohort model: the special status, in word recognition, of word onsets during the primary activation process of candidates that are phonologicaly similar. However, although our findings, as has been reported elswhere [7] strongly suggest «no evidence of facilitation in response to targets preceded by primes that shared word-initial phonological information with the target», evidence for inhibition could not be demonstrated: mean RTs obtained for W-W pairs without any priming effect, were not significantly longer than W-W syllabic prime pairs. We are inclined to favour the hypothesis of a desactivation process more than an inhibiting one. Lack of RT increase show that the final decision is not influenced by the prime *i.e.* activation level has become neutral again.

Results for the second experiment, demonstrate a facilitation of lexical decision when prime-words and target-words share the same morphemic onset (Type 1 *vs.* Type 3). This result seems to indicate that identification of a prefixe — and not simply of a phonological similarity — produces a different activation (facilitation) of the cohort candidates. But, on the contrary, when W-W pairs with a same morphemic or pseudo-morphemic (similar to syllabic priming) onset are compared, there is no significant effect. Taft & Forster hypothesized that RTs must be longer for pseudo-affixed words due to a complexe pre-lexical morphological analysis: our recent results do not confirm such a finding. Other studies, cited in the literature [8], also advocated that a prefixed-word prime does not facilitate the identification of a second word with the same prefixe (e.g. préface/prénom) compared to the condition in which the same word is presented after a pseudo-prefixed word sharing the same initial syllable (e.g., préfet/prénom).

Finally, the significant effect of morphemic priming found in Exp. n°2, could be interpreted in terms of a specific process within a word recognition model. This hypothesis is being explored using more data.

REFERENCES

[1] Marslen-Wilson W.D. (1987), "Functional parallelism in spoken word recognition", *Cognition 25, 71-102.*

[2] Marslen-Wilson W.D., Zwitserlood P. (1989), "Accessing Spoken Words: The Importance of Word Onsets", *Journal of Experimental Psychology : Human Perception and Performances, vol. 15, n°3, 576-585.*

[3] Bergman A.W., Hudson T.W., Eling P. (1988) ,"How Simple Complex Words Can Be: Morphological Processing and Word Representations", *The Quaterly Journal of Experimental Psychology, 40A, (1), 41-72.*

[4] Taft M., Hambley G., Kinishita S. (1986) "Visual and Auditory Recognition of Prefixed Words", *The Quarterly Journal of Experimental Psychology, 38A, 351-366.*

[5] Napps S.E. (1989),"Morphemic relationships in the lexicon : Are they distinct from semantic and formal relationships", *Memory & Cognition, 17 (6), 729-739.*

[6] Connan P.Y., Metz-Lutz M.N., Wioland F., Brock Gilbert (1994), "Rôle des informations lexicales dans la reconnaissance des mots. Etude chez les sujets sains", *TIPS, 24, 1-20.*

[7] Slowiaczek L.M., Pisoni D.P. (1986) "Effects of phonological similarity on priming in auditory lexical decision", *Memory & Cognition, 14 (3), 230-237.*

[8] Colé P. (1988), "Le traitement des mots dérivés : une analyse morphologique sélective", *L'année Psychologique, 88, 405-418.*

PHONOLOGICAL PRIMING EFFECTS IN LEXICAL COMPETITION

N. Bacri and F. Isel, Laboratoire de Psychologie Expérimentale, Université René Descartes & CNRS, Paris, France

ABSTRACT

Cross-modal phonological priming between auditory pseudoword primes and visual word targets gave rise to two experiments. Targets were carrier trisyllabic words in which two shorter words were embedded. Interference effects were greater than facilitation, and the embedded words were never activated. It is suggested that decision bias effects could explain the weakness of phonological priming.

Lexical ambiguity inherent in carrier words in which are embedded shorter words may be resolved either according to a specific segmentation routine [1] or as a by-product of excitation and inhibition of overlapping lexical items [6]. In fact, lexical embedding is very frequent in English [4], and is probably still more important in French, a language which presents around 85% of polysyllabic words. Gating studies, as well as simulations, suggest that the amount of acoustic information is a main determinant of lexical access as far as an interactive activation process is assumed to command which lexical candidates are momentarily activated, and which of them gets the strongest activation [2]. In the TRACE model proposed by McClelland and Elman [6], the degree of matching information determines the issue of lexical activation at each processing cycle. Intermediate units such as phonetic features or phonemes are progressively activated or desactivated, irrespective of the lexical status of the input signal. Exhaustive alignment of the processing system on the input produces competition between multiple lexical hypotheses [2].

Phonological priming paradigm is well suited to test certain assumptions of this model. A residual activation of lexical hypothesis corresponding to the embedded words may facilitate their recognition. On the other hand, the longer carrier word offers the best conditions of phonological overlap between prime and target, and might present the highest activation. The fact that primes are pseudowords (except in the repetition condition), and that targets are words ought not to diminish priming effects since these effects are based on phonological overlap.

However, it is worth noting that phonological priming data do not afford clear-cut evidence [7]. In experiments bearing on monosyllabic words [3, 9] or on bisyllables [7], when both primes and targets were auditorily presented, interferences were often more important than facilitations. Still more often, there was no consistent priming effect [3, 7, 9]. We used a cross-modal priming technique, where subjects see a visual target immediately following an auditory prime, in order to pick up residual effects of the activation of successive segments in the auditory input. If priming is modality-independent and intervenes at the level of lexical entries, an auditory prime may affect a visual target as well as an auditory target [5]. As for lexical embedding, it offers a better opportunity to follow the time course of competition between lexical hypotheses than the non ambiguous items that usually constitute the test materials [8].

Several predictions were tested. The multiple representations of lexical hypotheses at different moments in time allows different types of segmentation to operate [2]. The amount of matching information will interact with the presence or the absence of an alignment between the onset of auditory primes and the representation of word targets. Activation of the initially embedded word, although transient, could be sufficient to trigger a lexical access attempt at the monosyllabic target, since it is initially aligned with the input. The longer carrier word will receive an increasingly overwhelming amount of activation, since it satisfies to the two requisits of matching and alignment.

However, the finally embedded bisyllabic word could become weakly activated, at least from the Isolation Point (according to Marslen-Wilson's definition of the Isolation Point [5]) onwards, because its match with part of the input and despite it is not aligned with the input. A sufficient amount of matching could thus compensate for the non alignment. In a second experiment, primes corresponding to the carrier words will contain a parsing cue in order to diminish the advantage of the long words and to enhance the chances of embedded words of being recognized [2]. Alternatively, the absence of phonological priming might be due to processing differences between auditory and visual modalities. It might thus support the assumption that post-access processes are implied both in the priming paradigm and in the lexical decision task.

Two cross-modal priming experiments were run, both with a lexical decision task. Experiment 1 aimed to compare the respective activation levels of each embedded word and of the carrier word when the input signal and the phonological representations of each target were progressively overlapping. In experiment 2, an interval of silence was introduced in the repetition condition to reinforce the lexical hypotheses corresponding to the embedded words.

EXPERIMENT 1

Method

Subjects: 18 native speakers of Parisian French participated.

Materials

Test stimuli were 15 low-frequency trisyllabic words, constructed so that both the syllable in initial position and the two following syllables were high- or middle-frequency meaningful words, e. g. "chapelure" = "chat" + "pelure". Syllabic structure of the monosyllable was CV, CVC or CCV (C: consonant, V: vowel). Fifteen sets of five trisyllabic items served as primes. A set of primes involved three pseudowords and two words: (1) a pseudoword beginning with a phoneme of the same broad phonetic category as the initial phoneme of the target ("sebojim"), (2) a pseudoword beginning with the same first syllable as target ("chabojim"), (3) a pseudoword beginning with the same sequence as target until target Isolation point ("chapeleun"), (4) the carrier word, (5) a word unrelated with the target (control condition). Targets were in turn one of the embedded words or the carrier words. In addition, 15 fillers were presented as primes with a pseudoword as target. All the items were stored and digitized at 10 kHz with 12 bit resolution.

Procedure and design:

Each of the 18 subjects participated to the five conditions, the order of presentation of the pairs being balanced within each subgroup of 3 subjects. No subject heared the same prime twice or saw the same target twice. The visual probe followed the auditory target immediately, and the speeded lexical decision was performed on the visual target. Response times (RTs) were measured from the offset of the auditory prime. The types of primes (5 modalities) and the format of targets (monosyllables, bisyllables, trisyllables) were between-subject factors.

Results

Error rates and RTs longer than 1,500 ms did not exceed 1.5%, and were discarded from the analysis. Mean RTs were faster for a monosyllabic target (501 ms, sd = 98 ms) than for either a bisyllabic (569 ms, sd = 130 ms) or a trisyllabic target (595 ms, sd = 137 ms). ANOVAs run on RTs showed that the main effect of target format was significant, overall and for each overlap condition, both by subjects and by items. More important, pairwise comparisons between test and control conditions for each overlap and each type of target showed that a significant effect of phonological priming never appeared, excepted in the repetition condition for a trisyllabic target (Fig. 1). Thus, a partial overlap either had no effect, when prime and target were sharing a phonetic category for their initial phoneme, or gave rise to weak interferences for the two embedded words, even when the one-syllable overlap corresponded to the initially embedded word and when the Isolation-point overlap gave enough acoustic information to access the finally embedded word. The next experiment aimed to help the subjects parse the carrier words into their components.

EXPERIMENT 2

An interval of silence was introduced inside the carrier words in order to facilitate lexical access attempts to the embedded words. Its duration (from 18 ms to 34 ms) had been evaluated for each carrier word in a previous discrination experiment so that subjects could parse the signal in 50% of occurrences.

Method

18 native speakers of Parisian French participated. Except the introduction of an interval of silence in the trisyllabic primes for repetition trials, the apparatus and procedure were the same as those of the preceding experiment.

Results

1.3% erroneous data or RTs longer than 1,500 ms were discarded. Mean RT values were of the same order as previously, RTs to monosyllabic targets being significantly faster for each overlap condition than RTs to bisyllables and trisyllables. As shown in Fig. 2, comparing each overlap condition for each type of target to the control condition, the introduction of a silence strengthened the interference effects. Planned comparisons showed that these effects were significant for all types of word targets in the phonetic-category- and the one-syllable-overlap conditions, and for both embedded words in the repetition condition (for F1(1, 17), $p<.01$; for F2(1, 14), $p<.02$). The weak facilitation effect of the carrier-word target in the repetition condition did not reach significance. An interval of silence sufficient to be auditorily perceived did not trigger a lexical parsing of the carrier word into its constituents. This result pointed out that the interval of silence has not been processed as a potential boundary cue.

GENERAL DISCUSSION

The present research aimed to evaluate the time course of lexical competition when a carrier word contains embedded words. Phonological priming was not efficient to trigger lexical access attempts: A facilitatory effect appeared in just the repetition condition for the longer word. Initial overlap had no effect in most previous studies [3, 7, 9]. Contrary to our expectations, processing of ambiguous words did not enhance the weight of residual activation, if any. Whatever, our data are not inconsistent with all the predictions derived from TRACE. The amount of inhibition for the initially embedded word decreased progressively as the overlap increased, except in the repetition condition. The importance of onset alignment has been assessed for all the overlaps. The processing system ignored parsing cues. However, even if interference effects may correspond to the issue of lexical competition, the difference between the two experiments suggest that interferences may be due to the discrepancy between pseudoword primes and word targets [7], increased by the presence of a silence. The presence, in the experimental set, of plurisyllabic suits segmented into their lexical components disturbed the search of a congruency between primes and targets. Subjects might have not processed the primes containing an interval of silence as two words, but as a long word containing a mismatch by misalignment. When they saw the word targets, they needed more time to respond "word", because they were waiting for a pseudoword. This point could explain the observed longer RTs in the repetition condition (Exp. 2). These results do not ascertain that critical matching points, such as a one-syllable overlap or an overlap between prime and target until the Isolation Point of the finally embedded word, play no role in lexical access. They suggest that a lexical decision task taps only into the phonological output representation, and not into the lexical hypothesis elaborated during prime processing. The weakness of phonological priming could thus be due to a decision bias.

REFERENCES

[1] Banel, M-H., Bacri, N. (1994), "On metrical patterns and lexical parsing in French", *Speech Communication*, vol. 15, pp. 115-126.

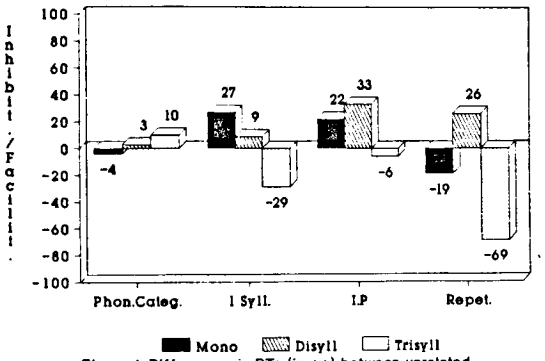


Figure 1: Differences in RTs (in ms) between unrelated and experimental conditions as a function of the amount of prime-target overlap for each type of target (Experiment 1)

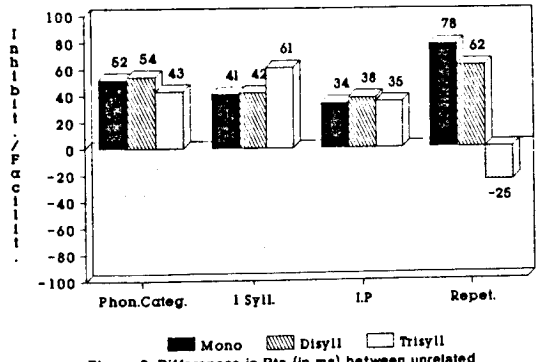


Figure 2: Differences in Rts (in ms) between unrelated and experimental conditions as a function of the amount of prime-target overlap for each type of target (Experiment 2)

[2] Frauenfelder, U., Peeters, G. (1990), "Lexical segmentation in TRACE: An exercise in simulation", in G. Altmann (Ed.), *Cognitive models of speech processing*, Cambridge, MA: The MIT Press, pp. 50-86.

[3] Goldinger, S., Luce, P., Pisoni, D., Marcario, J. (1992), "Form-based priming in spoken word recognition: The roles of competition and bias", *J. of Experimental Psychology: Learning, Memory, and Cognition*, vol. 18, pp. 1211-1238.

[4] Luce, P. (1986), "A computational analysis of uniqueness points in auditory word recognition", *Perception & Psychophysics*, vol. 39, pp. 155-158.

[5] Marslen-Wilson, W. (1990), "Activation, competition, and frequency in lexical access", in G. Altmann (Ed.), *Cognitive models of speech processing*, Cambridge, MA: The MIT Press, pp. 148-172.

[6] McClelland, J., Elman, J. (1986), "The TRACE model of speech perception", *Cognitive Psychology*, vol. 18, pp. 1-86.

[7] Radeau, M., Morais, J., Dewier, A. (1989), "Phonological priming in spoken word recognition: Task effects", *Memory and Cognition*, vol. 17, pp. 525-535.

[8] Shillcock, R. (1990), "Lexical hypotheses in continuous speech", in G. Altmann (Ed.), *Cognitive models of speech processing*, Cambridge, MA: The MIT Press, pp. 24-49.

[9] Slowiaczek, L., Hamburger, M. (1992), "Prelexical facilitation and lexical interference in auditory word recognition", *J. of Experimental Psychology: Learning, Memory, and Cognition*, vol. 18, pp. 1239-1250.

DO METRICAL AND PHONOTACTIC SEGMENTATION CUES COOPERATE IN SPOKEN WORD RECOGNITION?

Bacri N. & Banel M-H.,
Laboratoire de Psychologie Expérimentale, Université René Descartes & CNRS, Paris, France

ABSTRACT

A Metrical Segmentation Strategy has been proposed to account for the segmentation of stress-timed language like English [3] whereas for syllable-timed languages like French segmentation should depend on syllabic structure. Three word-spotting experiments supported a different assumption. Detection of words embedded in nonsense strings either in initial (Exp. 1 & 3) or in final position (Exp. 2) provided evidence of an effect of prosodic structure on the phonotactically based mechanism of parsing.

Speech segmentation is basically supported by two types of cues: prosodic cues [3, 5, 6] and phonotactic cues [7]. Both are determined by language-specific constraints. Lexical parsing is assumed to be stress-based for languages contrasting stressed and unstressed syllables. It is founded on distributional phonotactic properties as evidenced by juncture misperception data [2]. Syllabification was almost equally affected by the type of phonotactic string and by the stress pattern of the stimuli [7]. In French, in which stress falls regularly on the lengthened final syllable, durational inter-syllabic differences may have a functional role in speech segmentation. The parsing of bisyllabic words into their monosyllabic constituents showed that the usual short-long (iambic) pattern produced recognition of bisyllables more often than did the reverse long-short (trochaic) pattern [1]. Subjects' behavior was well adapted to the trailer-timed structure of French language [8] and to the fact that 85% of lexical items are polysyllabic. Syllable internal structure had no clear effect on parsing, but the parsing of long-long spondees was facilitated by CVC syllables. Listeners were induced to apply a syllable-based device when lacking metrical cues. However, data were compatible with two different accounts. The efficiency of prosodic cues might be due to rhythmic expectancies and intervene post-lexically. On the other hand, a metrical segmentation device might have had a direct effect on parsing, either cooperating or conflicting with phonotactic cues.

Well-formedness conditions for syllable-final consonant clusters imply, at the phonological level, that the first but not the second consonant belongs to the same syllable as the preceding vowel: French codas cannot contain more than one consonant. "garde" would be syllabified as /gar # dØ/. But in the specific case of the obstruent + liquid legal clusters, whatever their position, the two consonants are tautosyllabic: "livre" is syllabified /li # vrØ/ [4]. If we consider an illegal medial cluster, e. g. /VpzV/, the first assumption leads to syllabify it as /Vp # zV/, whereas the second one forbids the processing of /pz/ as tautosyllabic, and blocks a /V # pzV/ parsing. These phonotactic constraints constitute powerful segmentation cues.

The aim of the present experiments was to determine whether or not metrical and phonotactic segmentation devices are simultaneously available. A word-spotting task offers a good opportunity to address this question [3]. Listeners have to detect real monosyllabic words in nonsense bisyllabic strings. The medial consonant cluster, when present, was illegal, thereby providing a segmentation cue. Metric structure realized an iambe, a trochee or a spondee. According to the metrical segmentation hypothesis, the bisyllable string will be more easily segmented when bearing a trochaic pattern, whereas it will be processed as a whole when bearing an iambic pattern. If a metrical segmentation device intervenes prelexically, detection of the word in initial position will be facilitated in the first case when both cues cooperate, and interfered with in the second case by the conflict between phonotactic and prosodic cues, as compared with the prosodically neutral condition.

EXPERIMENT 1

Method:

Subjects:

37 native speakers of Parisian French.

Materials:

The materials consisted in two sets of nonsense bisyllables. The 18 items of the first set had an embedded high frequency CVC word target in initial position. Word final consonant was an obstruent. Consonant sequence in medial position cannot form a syllabic coda or onset, e. g. /bk/, /bv/, /pz/, /tl/, /tn/. The 18 CVC-CVC items in the second set were nonsense fillers without embedded word target.

Three metrical patterns were realized for each test item: A short-long iambic pattern, a long-short trochaic pattern, and a long-long spondee. All the fillers were spondees. All the stimuli were recorded by a Parisian French female speaker, at a regular rate and without salient F0 movements. They were stored and digitized at 10 kHz with 12 bit resolution. Monosyllabic mean duration was about 520 ms when long, and about 350 ms when shortened with a compression rate of 35%.

Procedure and design:

Subjects were told that they would hear nonsense sequences, some of them having a real CVC word inside, in initial position. On detecting a word, they had to press a button as quickly as possible, and to say it aloud. Detection times were measured from the burst of the word final consonant. Three experimental tapes were constructed, one for each metrical pattern, each having 32 stimuli (18 word targets plus 18 fillers). The order of stimuli was balanced across the three tapes. The three practice tapes included 32 different bisyllabic strings (16 with, and 16 without a CVC word), bearing the same metrical pattern than the corresponding test tape. Subjects were given a feedback on their performance for the first half of the practice tape. Each experimental condition combined an illegal cluster with (1) a neutral pattern ("one cue" control condition), (2) an iambic pattern (conflicting cues), (3) a trochaic pattern (cooperating cues). Each subject heard the three metrical patterns, but each item was presented bearing only one pattern.

Results and discussion:

Seven subjects, having detected less than 50% of the targets, were discarded, leaving 10 subjects for each tape. Missing data (3.9%) and reaction times (RTs) greater than 1500 ms (2.4%) were replaced. RTs were faster for a trochaic pattern than for either an iambic pattern or a neutral spondee (Figure 1, left panel). The main effect of Condition was significant (ANOVA by subject: $F1(2, 58)= 9.3$, $p<.001$; by item: $F2(2, 34)= 21.2$, $p<.001$). Pairwise comparison showed that RTs were reliably faster for a trochaic pattern than for an iambic and a neutral pattern, respectively. The tendency towards slower RTs for a neutral pattern than for an iambic pattern did not reach significance.

The order of performance levels supported the predictions derived from the application of a metrical segmentation device. Word extraction is facilitated when a long-short pattern is correlated with an illegal cluster. In as much as the cluster cannot be tautosyllabic, it is parsed into its constituent units, and an attempt at lexical access, initiated near the beginning of the input in the case of a trochaic pattern, can be easily achieved. In the case of iambic patterns, lexical access is delayed, since the nonsense string is processed as a whole on the basis of its metrical structure. However, conflicting cues did not reliably increase RTs as compared with the control condition. The lack of a reliable interference when cues were conflicting suggests that the two types of information have been processed separately. But the redundancy gain when both cues cooperate is consistent with the view that both dimensions may nevertheless influence each other. The purpose of the next experiment was to determine whether this influence is symmetrical or not.

EXPERIMENT 2

To address this issue, the functional relations between the two segmentation cues were reversed, by locating the target word in final position. The former "conflicting cues" (an illegal cluster plus an iambic pattern) might now facilitate a lexical access attempt on the final syllable, whereas the former

"cooperating" cues (an illegal cluster plus a trochaic pattern) should now deter the listeners from pursuing a lexical search after having parsed the bisyllable and found a nonword. The presence either of interferences (slower RTs with a trochaic pattern than in the control condition) or of a redundancy gain (faster RTs when iambic than in the control condition) will permit to decide whether both cues are processed conjointly and symmetrically. On the other hand, if the two cues are processed separately, results will show neither interference, nor redundancy gain.

Method:

Subjects: 30 Parisian French speakers.

Materials and procedure:

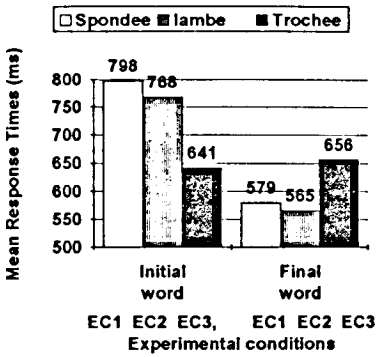
18 new nonsense bisyllabic strings were created by reversing the order of the first and the second syllable. Medial illegal clusters assembled obstruents and liquids so that they could not be tautosyllabic in French, e. g. "lamp(e) + zok" became "zor + lamp(e)" to avoid a /kl/ legal cluster. The apparatus and procedure were similar to those of the preceding experiment.

Results and discussion:

Missing data (6%) and RTs exceeding 1500 ms (3.1%) were replaced. As in the previous experiment, the main effect of Condition was significant ($F1(2, 58) = 5.2$, $p<.01$; $F2(2, 34) = 7.1$, $p<.01$). The 14 ms difference between RTs obtained for iambic patterns and RTs for the control condition did not reach significance, but RTs for trochaic patterns were reliably slower than RTs for the two other conditions (Fig. 1, right panel). The order of performance levels was reversed by comparison with the preceding experiment. There was no redundancy gain when iambic patterns induced subjects to process the whole string. But the convergence of both cues to parse the input increased RTs and slowered lexical access. These results suggest a mutual interference between prosody-based and phonotactic-based segmentation devices.

EXPERIMENT 3

The purpose of the next experiment was to evaluate the weight of the phonotactic cue, respective of metrical pattern, by comparing the processing of CVCCVC strings and of CVCVC strings.



EC1: one segmentation cue
EC2: conflicting cues
EC3: cooperating segmentation cues

Figure 1: Mean RTs as a function of experimental conditions and word position (left panel Exp.1, right panel Exp. 2).

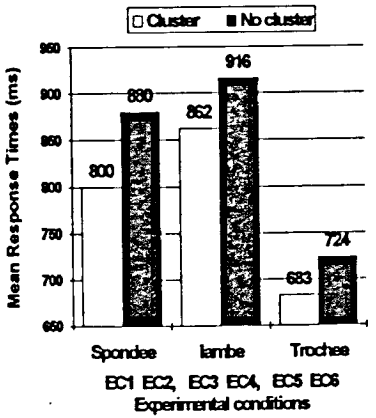
In the later case, word coda is assembled with the following vowel, e. g. "lampoc" is syllabified /lã # pok/. The target word is in initial position. A lexical access requires a resyllabification that is impeded by an iambic pattern indicating not to parse the bisyllable, but also by a trochaic pattern which induces to parse the string in such a way that the parsing produces two nonsense syllables.

Method:

Six groups of ten subjects were tested on 18 test items, with or without a medial cluster, depending on the condition, and bearing one of the metrical patterns. The apparatus and procedure were as previously described.

Results:

Two test items were missed by 50% of the subjects. Corresponding data were discarded. The overall analysis of variance showed a significant main effect for both the phonotactic structure ($F1(1, 54) = 4.4$, $p<.04$; $F2(1, 15) = 21.8$, $p<.001$) and the metrical pattern ($F1(2, 54) = 15.7$, $p<.001$; $F2(2, 30) = 22.5$, $p<.0001$). The interaction between the two factors did not reach significance. The order of RTs was the same for both types of phonotactic structure (Fig. 2). It approximated the order obtained in Experiment 1, and yielded the same levels of significance. The weight of a phonotactic cue is more important in the control condition. However, the 54 ms (iambic pattern) and the 41 ms lengthenings (trochaic pattern) indicate merely a tendency towards an impairment of lexical access attempt when listeners are lacking a strong phonotactic segmentation cue.



EC1: one relevant segmentation cue.
EC2: one irrelevant segmentation cue.
EC3: conflicting cues.
EC4: converging irrelevant cues.
EC5: cooperating segmentation cues.
EC6: cooperating irrelevant cues.

Figure 2: Mean RTs as a function of experimental conditions (Exp. 3).

GENERAL DISCUSSION:

The ability to segment speech input and to detect a word merged into a nonsense string crucially depends on the position of the word in the string. The target position determines the functional relations between metric and phonotactic structures, and thus the efficiency of the metrical segmentation device. However, together with the redundancy gain when both cues are cooperating (Exp. 1), and with a mutual interference when they give competing information (Exp. 2), the last experiment clearly demonstrates a joint influence of metrical contrast and of phonotactic constraints. This joint influence cannot be the result of interactions in the perceptual system (Exp. 3). We suggest that metrical structure might activate the segmentation procedure which is in turn triggered by a phonotactic disjuncture. McQueen et al.' proposal (1994) should be modified to take the specificity of French metrics into account. A metrical segmentation device would permit to anticipate word boundaries, but makes no assumption about the mechanism of parsing. The interference effect, although limited to certain condition, would be best interpreted if metrical and phonotactic information are processed in parallel.

ACKNOWLEDGEMENTS

Financial support for this research was provided by the GDR 957 "Sciences Cogntives de Paris" (CNRS).

REFERENCES

[1] Banel, M-H., Bacri, N. (1994), "On metrical patterns and lexical parsing in French", *Speech Communication*, vol. 15, pp. 115-126.

[2] Cutler, A., Butterfield, S. (1992), "Rhythmic cues to speech segmentation: Evidence from juncture misperception", *J. of Memory and Language*, vol. 31, pp.218-236.

[3] Cutler, A., Norris, D. (1988), "The role of strong syllables in segmentation for lexical access", *J. of Experimental Psychology: Human Perception and Performance*, vol. 14, pp. 113-121.

[4] Dell, F. (in press), "Consonant clusters and phonological syllables in French", *Lingua*, 95.

[5] Grosjean, F., Gee, J. P. (1987), "Prosodic structure and spoken word recognition", *Cognition*, vol. 25, pp. 135-155.

[6] McQueen, J., Norris, D., Cutler, A. (1994), "Competition in spoken word recognition: Spotting words in other words", *J. of Experimental Psychology: Learning, Memory and Cognition*, vol. 20, pp. 621-638.

[7] Treiman, R., Danis, C. (1988), "Syllabification of intervocalic consonants", *J. of Memory and Language*, vol. 27, pp. 87-104.

[8] Wenk, B., Wioland, F. (1982), "Is French really syllable-timed?", *J. of Phonetics*, vol. 10, pp. 193-216.

SPOKEN WORD SEGMENTATION IN ARABIC LITERATE AND ILLITERATE SUBJECTS: A PSYCHOLINGUISTIC APPROACH

Mohamed FARID
Laboratoire de Phonétique, U. F. R de Linguistique, Université Paris 7, 10, rue Charles V, 75004, Paris, France

ABSTRACT

An experiment was conducted to study the effect of alphabetic literacy on developing the ability of speech segmentation. Both Arabic literate and illiterate subjects were asked to segment progressively spoken Moroccan Arabic sentences

The results showed that literates were able to reach the level of phonemes in their segmentation, whereas illiterates reflected a syllabic procedure of speech segmentation and were unable to segment phonemically.

We conclude that the level of phonological awareness, that is the ability to consciously recognize the internal phonemic structure of spoken words, is higher in Moroccan Arabic literates than in illiterates. This result speaks in favour of literacy having a crucial role in determining the level of processing which a listener can reach.

INTRODUCTION

The cognitive processes underlying speech segmentation make up a central topic in psycholinguistic studies. In the last decade, cognitive psychologists have been interested in studying the ability to segment speech signal into its component units. Some psycholinguists proposed the notion of phonological awareness [1]. Phonological awareness refers to a special kind of phonological representations. It is a type of phonological knowledge which differs from the phonology used in language production and comprehension. In other words, phonological awareness refers to conscious representations of the phonological properties and constituents of speech. Some studies claimed that phonological awareness is logically related to reading and spelling acquisition in an alphabetic system [2]. More recently, some researchers [3] have considered this ability to be a crucial component of reading and spelling. Its development is dependent on the learning of reading and spelling.

There are three levels of phonological awareness: word awareness, syllable awareness and awareness of phones (sub-syllabic units like onset and rime). The phonological awareness hypothesis is supported by some psycholinguistic studies showing that preschool children like illiterates were unable to manipulate speech segments at a sub-morphemic components level [4 and 5.]. These subjects were good at manipulating syllable units but poor at segmenting speech into phones.

Two types of studies were proposed to test the phonological awareness hypothesis: (1) experiments using metaphonological tasks such as rime judgement, syllable addition or deletion; (2) longitudinal or correlational studies.

More recently, data obtained with Portuguese illiterates suggested that that population could not analyse speech explicitly as a sequence of phones. Thus, awareness of segmental structure of speech does not arise spontaneously in the course of cognitive growth, but in the learning of reading [4].

The tasks used in Morais et al. [4] consisted in adding or deleting a phone (consonant) at the beginning of a non-word. The results obtained suggested that illiterate subjects were unable to delete or add a consonant, but these tasks were easily performed by ex-illiterate adults who learned to read. Thus, illiterate speakers did not represent speech as a sequence of phones.

This awareness is probably provided by learning to read in an alphabetic system. Other experiments have demonstrated that literacy training has an effect on speech segmentation [5]. Illiterate subjects, unlike literates, displayed an incapacity to deal with phonetic segments (initial consonant) in a detection task and in a progressive free segmentation task [5]. But their performance was better with syllables. Thus, the capacity to analyze intentionaly and explicitly speech at a segmental level is developed in an alphabetic code [6]

Morais et al.[5] suggested that reading acquisition is correlationaly significant with the ability to deal with sub-morphemic units of speech such as syllables and phonemes.

The aim of the present study is to assess the segmentation capacities of literate and illiterate Moroccan speakers in a progressive segmentation task. Previous studies with Arabic literate and illiterate speakers [7 and 8] have demonstrated that phonological awareness develops with literacy acquisition. Literate people have a good performance when processing speech units at very difficult level. For example, the processing of intervocalic geminates or long consonants seems to vary as a function of educational level [9]. Derwing [9] investigated syllable boundaries in semiliterate and ilitterate Arabic speakers (Cairo). His results showed that literate subjects processed geminate consonants as ambisyllabic bisegments, but this tendency was much reduced in subjects with lower educational levels (semiliterates). Thus, judgments about syllable boundaries depend on educational level in general and on literacy in particular [9].

According to the present hypothesis, reading ability has a strong effect on subjects performance in a speech segmentation task at a sub-syllabic level. Generally, speakers living in poor cultural environment can not develop metalinguistic capacities such that they can perform well on segmental or metaphonological tasks.

PROGRESSIVE SEGMENTATION TASK

Method

Stimulus Material

Ten Moroccan Arabic sentences were used as stimulis in the experiment. They were five to seven words long. The long word in the ten sentences was trisyllabic and frequent structure was bisyllabic as [CV-CV] *"daru" (his house)*; the short word had a [CV] structure like *"wa" (and)*. An example of sentence-stimulus is:

<< had llaʕib faz belkura ddahabiyya ʕla laʕbu lmumtaz >> (" This player was awarded the gold ball for his excellent performance").

Subjects

The experiment was run in Paris. Two groups of subjects participated in it: illiterate adults and literate adults. The illiterates were eight subjects (2 females and 6 males aged 30 to 65. They were Moroccan immigrants having lived in Paris for many years. They were all of peasant origin and none had received any reading instruction at any time.They speak poor French. The literate subjects were administered a reading test at the end of the experiment. It consisted in reading as fast and as accurately as possible 120 arabic words, most of them, nouns (65), the majority of which were bisyllabic (52) or trisyllabic (42). The results showed a clearly discontinuous distribution, sug)gesting the presence of two types of subjects who will be called better and poorer readers. Better readers read over 60 words/min and did not make errors. Poorer readers read less than 60 words/min and made errors. Ten better readers (2 females and 8 males) aged 22 to 31 were selected. They were students in a Paris university and had received, at least, bilingual instruction in reading and writing both Arabic and French .

The poorer readers were eight subjects (3 females and 5 males) aged 21 to 51. All were workers and had stopped their

schooling in primary school. They read and wrote poorly in Arabic and French.

Task and Procedure

The subjects listened to recorded sentences and were asked to say only part of a sentence, then only a subpart of the part, and so on, until they could not go any further: each subject segmented progressively all the sentences that served as trials.

Results

Mean percents of segmentation types are presented in Table 1. These were based on the number of responses produced by each subject on each sentence. Five types of isolated linguistic units were selected for the analysis: (1) phones (consonants), (2) syllables, (3) one word, (4) two words, and (5) sequences of words (more than two words).

Illiterate subjects had a higher performance in units "one word" (32.54 %). This isolated linguistic unit is very significant in the process of segmentation in illiterates. Performance with "phones" significantly differed from "syllables" ($t(9)= 1.48$, $p<.005$). Also performances on phone and "one word" were significantly different ($t(9)= 6.20$, $p<.001$).

Nevertheless, poorer readers showed a similar performance in segmenting sentences in relevant linguistic units. But, one notices that this group of subjects had a higher performance in isolating "more than two words" (35.17 %). No difference was revealed between isolating "one phone" and "one syllable" ($t(9)= 0.77$), but the difference was significant between "one phone" and "one word" ($t(9) = 2.16$, $p<.05$). Poorer readers performed well progressive segmentation from "one word" to "more than two words".

Better readers performed well on all types of segmentation. They reached the phone level. This sub-syllabic unit was rarely produced by illiterates and poorer readers as opposed to better readers. But at the word level, all subjects (illiterates, poorer and better readers had similar performance.

Better readers reached, without difficulty, the phone and the syllable levels. This gives further support to the hypothesis that better readers have the ability to reach the phonemic and syllabic units in a progressive segmentation.task. Analysis of variance (ANOVA) performed on subject's responses yielded a significant effect of alphabetic literacy ($F(4,25) = 10.84$, $p<0.005$).

Table 1. Progressive segmentation of speech. Percentage of final responses of each type.

Isolated units	Illiterates	Poorer readers	Better readers
One phone	23, 12	26, 63	50, 25
One syllable	25, 08	26, 76	48, 16
One word	32, 54	33, 90	33, 56
Two words	27, 40	31, 51	41, 10
more than two words	20, 91	35, 17	43, 92

DISCUSSION

In the Arabic alphabet, it is difficult to segment a syllable into a consonant and a vowel because vowels are represented by diacritics in the writing system. The diacritics do not have an independant status as consonants do. For example the syllable [ka] is written in Arabic as a consonant plus a diacritic mark. This concerns the written syllable in Arabic. For the spoken syllable, the problem of analysis is not similar.

The results obtained in the present experiment showed that better readers have a more developed phonemic awareness than poorer readers and illiterates. They were able to isolate correctly the small sub-lexical units (phonemes and syllables) which are components of the phonemic structure of words and sentences. The development of this awareness is explained by their reading and spelling practice in an alphabetic system. Thus, cognitive capacities can help the speaker-hearer manipulate speech units. These manipulation of the segmenal structure of words is a result of a conscious and intentional processing of speech elements. Moreover, both reading and spelling imply, in addition to the ability to perceive minimal phonetic distinctions, an explicit knowledge of the phonetic structure of speech. Furthermore, to segment progressively spoken sentences requires that subjects develop a special strategy in the segmentation process. First, they must memorize the whole sentence and then processe it according to their metalinguistic and linguistic knowledge. Illiterates and poorer readers do not have sufficient metalinguistic knowledge to reach such sub-lexical units. The fact that illiterates are not aware of the phonetic structure of speech does not imply, of course, that they do not use segmenting routines at this level when they listen to speech [4].

The hypothesis that reading and spelling knowledge may develop the capacity to segment speech into its small components is confirmed. This study is a comparison of performances between illiterates and literates in speech segmentation. It deals with the effect of alphabetic literacy on spoken word recognition and segmentation. It is a contribution to understand cognitive processes and the mechanisms of language processing in general, and speech segmentation in particular.

REFERENCES

[1] Morais, J. (1991), "Phonological Awareness: A Bridge Between Language and Literacy", .in D.J. Sawyer & B.J. Fox (Eds.) *Phonological Awareness in Reading: The Evolution of Current Perspectives*. Springer-Verlag.Pp. 31-71

[2] Morais, J. (1991), "Constraints on the development of phonemic awareness", in S. Brady & D. Shankweiler (Eds.), *Phonological processes in reading: A tribute to Isabelle Liberman*, Hillsdale, NJ: Lawrence Erlabaum.

[3] Morais, J. (1985), "Literacy and awareness of the units of speech: implications for research on the units of perception", *Linguistics*, vol. 23, 707-721.

[4] Morais, J., Cary, L., Alegria, J., & Bertelson, P. (1979), "Does awareness of speech as a sequence of phones arise spontaneously?", *Cognition*, vol. 7, 323-331.

[5] Morais, J., Bertelson, P., Cary, L., & Content, A. (1986), "Literacy training and speech segmentation", *Cognition*, vol. 24, 45-64.

[6] Read, C., Zhang Y., Nie H., Ding B. (1986), "The ability to manipulate speech sounds depends on knowing alphabetic spelling", *Cognition*, vol. 25, 21-52.

[7] Idrissi-Bouyahyaoui, B. (1987), *Metalinguistic awareness in literate and illiterate children and adults: A psycholinguistic study*. Ph.D. University of Edinburgh.

[8] Farid, M.. (1991), *Quelques Aspects de la segmentation et de la perception de la parole chez des sujets lettrés et illettrés adultes: Etude psycholinguistique*. D. E. A. in Phonetics, Université Paris VII.

[9] Derwing, B. L. (1992), "A 'pause-break' task for eliciting syllable boundary judgments from literate and illiterate speakers: preliminary results for five diverse languages", *Language and Speech*, vol. 35 (1,2), 219-235.

THE PERCEPTION OF THE SINGLE-GEMINATE CONSONANT CONTRAST BY NATIVE SPEAKERS OF ITALIAN AND ANGLOPHONES

Bernard L. Rochet and Anne P. Rochet
University of Alberta, Canada

ABSTRACT

The results of a psychoacoustic experiment suggest that native speakers of Italian distinguish between intervocalic single and geminate consonants (as in fato and fatto) on the basis of the duration of these consonants, and not in terms of the duration of the preceding vowels, while anglophones perceive the same contrast not in terms of the duration of these consonants but in terms of the duration of the preceding vowels.

INTRODUCTION

The difference between Italian words like fato ('fate') and fatto ('fact') is generally stated in terms of the opposition between a single consonant (in fato) and a double consonant (in fatto), or between a short and a long consonant.

It is well known, however, that this consonantal length difference is accompanied by a vowel length difference: Italian geminates are preceded by short vowels, while their single intervocalic counterparts are preceded by long vowels [1, 2, 3, 4].

As in all cases where two phonetic characteristics covary, the question arises as to whether one of those two characteristics is perceptually more salient than the other.

Although it is generally assumed that native speakers of Italian are sensitive to the difference in consonantal length, it has been suggested that the primary perceptual cue in distinguishing between words like fato and fatto is the difference in vowel duration that characterizes such words [4]. Neither claim has been tested experimentally.

Contrary to Italian, English does not make use of an opposition between single and double consonants, and it is reasonable to assume that English speakers learning Italian do not perceive the difference between words like fato and fatto in terms of the duration difference between [t] and [tt].

The purpose of this study was to establish

1. whether the difference between words like fato and fatto is perceived by native speakers of Italian as residing in the consonant or the vowel; and

2. whether this difference is perceived in the same way by anglophones learning Italian.

PROCEDURE

Stimuli

The stimuli were prepared in the following way. A token of fato recorded by a native speaker of northern Italian on a good quality cassette recorder was low-pass filtered at 8.8 kHz to preclude aliasing, and digitized at 22 kHz. The digitized signal was then modified by means of a waveform editor (SoundEdit) to produce 7 stimuli by decreasing the length of the vowel from 215 to 92 ms in steps corresponding to two pitch periods (17-18 ms); each pair of pitch periods was removed from the middle portion of the vowel to leave the CV and VC transitions intact.

For each stimulus thus obtained, 5 new stimuli were produced by increasing the length of the silent portion of the intervocalic consonant ([t]) in 30 ms steps from 100 ms to 220 ms. This yielded a total of 35 stimuli (7 vowel durations x 5 consonantal durations).

Subjects

Subjects were twelve native speakers of northern Italian enrolled at the University of Bologna, and twelve native speakers of Canadian English attending the University of Alberta. Their ages ranged from 21 to 24 years.

Experimental Task

Subjects were asked to identify as fato or fatto the stimuli described above, in which the durations of the intervocalic consonant ([t]/[tt]) and the vowel preceding it ([a]) were varied systematically. They listened to 10 repetitions of each stimulus played in random order via a computer program and delivered through good quality headphones. The subjects' task was to identify each token as either fato or fatto by clicking in the appropriate box on a computer screen using a mouse-driven cursor.

RESULTS

The results for the Italian and English listeners are presented below by means of identification functions, with consonant duration varying in some, and vowel duration in the others.

Italian Listeners

The identification functions in Figures 1 and 2 are representative of those obtained from all the Italian listeners, with minor variations.

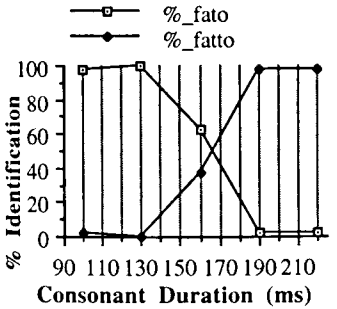


Figure 1.

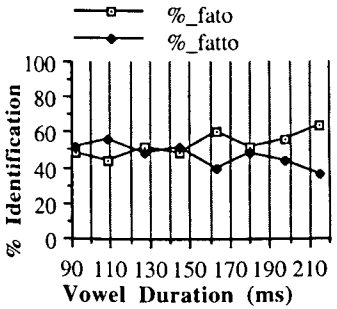


Figure 2.
The well-defined identification functions in Figure 1 (where the variable is consonant duration) and the undifferentiated identification functions in Figure 2 (for vowel duration) suggest that the Italian listeners distinguished between fato and fatto on the basis of consonant duration but not on the basis of vowel duration.

English Listeners

The identification functions in Figure 3 are representative of those obtained from all the anglophones, with minor variations. These undifferentiated identification functions suggest that when the varying dimension was consonant duration, the anglophones were unable to distinguish fato from fatto.

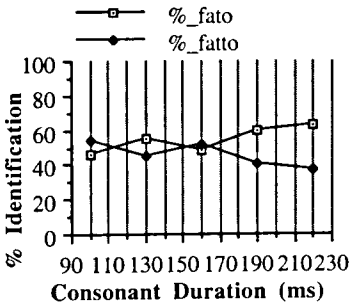


Figure 3.
When the varying dimension was vowel duration, two patterns of identification emerged for the anglophones, as illustrated in Figures 4 and 5.

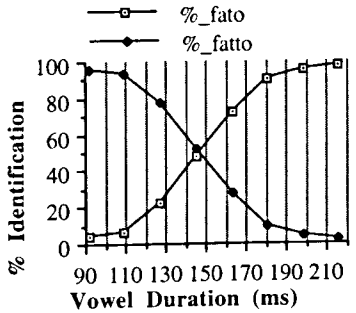


Figure 4.
Figure 4 suggests that some anglophones (n=8) used vowel duration as a perceptual cue to distinguish between fato and fatto. Because they associated fato with a perceived long vowel and fatto with a perceived short vowel, it can be said that those listeners' identification of fato and fatto was

essentially correct, in spite of the fact that they used vowel duration instead of consonant duration as a perceptual cue. For those listeners, a long vowel signaled a following short (or single) consonant, and a short vowel signaled a long (or geminate) consonant, in keeping with the duration characteristics of Italian.

On the other hand, Figure 5 suggests that some anglophones (n=4) equated vowel duration and consonant duration in a direct way. They identified as fato tokens with a short vowel and as fatto tokens with a long vowel. This is contrary to the facts observed about Italian quantity, and results in misidentification of fato and fatto.

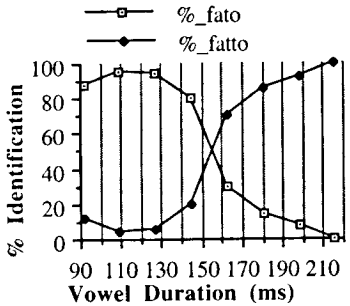


Figure 5.

DISCUSSION

These results suggest that Italian listeners may be sensitive to differences in consonant duration and not to co-occurring differences in vowel duration in the process of distinguishing between words like fato and fatto. This finding lends support to phonological analyses that describe the opposition between words like fato and fatto in terms of differences in consonant duration, and view the associated vowel duration differences as allophonic variation.

On the other hand, it appears that anglophones may not be sensitive to differences in consonant duration and that they may rely instead on differences in vowel duration when they are asked to make the fato-fatto distinction. This is not surprising in view of the fact that English speakers do not distinguish between long and short consonants, but distinguish vowels that may be differentiated in terms of their durations (e.g., /i/ vs /I/, /u/ vs. /U/), and rely on preconsonantal vowel duration differences in perceiving voiced-voiceless consonant contrasts [5, 6, 7].

Although the difference between the two subgroups of anglophones who used vowel duration in different ways to identify fato and fatto may reflect different levels of linguistic aptitude, it remains a superficial difference and does not affect this study's basic findings. Neither subgroup used consonant duration as a perceptual cue to distinguish between fato and fatto; both used vowel duration. The subjects who equated vowel and consonant duration appear to be sensitive to quantity differences only as they pertain to vowels, as suggested by the fact that they attributed to consonants the duration differences they perceived among the vowels. The subjects who were able to detect that vowel and consonant duration were negatively correlated appear to have shown some sensitivity to consonant duration, but they used vowel duration differences as the primary perceptual cue.

Although English learners appear to be capable of distinguishing between Italian single and geminate consonants on the basis of concommitant vowel duration differences, their continued reliance on the latter in production is likely to prevent them from being understood, or perceived as native or near-native by Italian listeners, because the latter appear not to be sensitive to vowel duration differences but only to consonant duration differences.

CONCLUSIONS

The results of this experiment suggest that listeners from different language backgrounds may perceive the same phonetic input in different ways, i.e., by using different characteristics of the acoustic signal as perceptual cues. One implication of these results is that foreign language pronunciation training should consider the linguistic background of the learners, both to understand their pronunciation mistakes, and to devise instructional tools that focus on the language-specific phonetic features that need attention.

In addition, it appears that although foreign language learners' inaccurate pronunciations may reflect faulty articulation, it is also possible that such mispronunciations are the consequence of a faulty perception of the target sounds, i.e., a perception of those sounds in terms of the learners' native language categories [8]. Consequently, it appears that auditory training must play an important part in foreign language pronunciation instruction.

The results of this study also suggest that pronunciation training should take into account not only the perceptual biases of the learners (to adopt the appropriate instructional procedures that make it possible to modify those biases), but also the perceptual expectations of the native speakers of the target language, to make sure that learners produce the appropriate phonetic cues, i.e., those that are recognizable by the target language speakers. This in turn provides support for a type of contrastive phonetics that derives its explanatory power from a thorough examination of foreign language learners' perceptual and articulatory behaviours, and the perceptual consequences of those behaviours on target language listeners [9].

ACKNOWLEDGEMENT

The authors gratefully acknowledge the support of a Central Research Fund grant and an SAS grant from the Faculty of Arts at the University of Alberta; the staff and students of the Centro Interfacoltà di Linguistica Teorica e Applicata "Luigi Heilmann," at the Università degli Studi di Bologna; and especially Dr. Grazia Busa for her generous help in recruiting the speakers of northern Italian and providing the facilities for their recordings in Bologna.

REFERENCES

[1] Josselyn, F. M. (1900), *Etude sur la phonétique italienne*, Paris.
[2] Panconcelli-Calzia, G. (1911), *Italiano*, Leipzig.
[3] Metz, C. (1914), *Ein experimentell-phonetischer Beitrag zur Untersuchung der italienischen Konsonantengemination*, doctoral dissertation, Glückstadt.
[4] Parmenter, C. E. and Carman, J. N. (1932), "Some remarks on Italian quantity," *Italica*, vol. 9, pp. 103-108.
[5] Lauefer, C. (1992), "Patterns of voicing-conditioned vowel duration in French and English," *Journal of Phonetics*, vol. 20, pp. 411-440.
[6] Gottfried, T. L. (1982), *Perception of French and American vowels: A cross-language study*. PhD dissertation, University of Minnesota.
[7] Crowther, C.S. and V. Mann (1992), "Native language factors affecting use of vocalic cues to final consonant voicing in English," *Journal of the Acoustical Society of America*, vol. 92, pp. 711-722.
[8] Rochet, B. L. (1991), "Perception of the high vowel continuum: A cross-language study," *Actes du XIIème Congrès International des Sciences Phonétiques*, Aix-en-Provence: Université de Provence, Service des Publications, vol. 4, pp. 94-97.
[9] Kohler, K. J. (1981), "Contrastive phonology and the acquisition of phonetic skills," *Phonetica*, vol. 38, pp. 213-226.

PERCEPTUAL CONFUSION BETWEEN SOUTH AFRICAN AND BRITISH ENGLISH VOWELS

A. Traill, M. J. Ball, and N. Müller†*
*University of the Witwatersrand, *University of Ulster, †University of Central England*

ABSTRACT

South African English demonstrates a shift in the realizations of front vowels when compared to most British English accents. This study includes a perception experiment, where British English listeners are presented with tokens containing the vowels in question recorded by 4 SAE speakers. Then an acoustic analysis of the vowels is presented, which demonstrates the similarity between the British English set, and their SAE counterparts.

SOUTH AFRICAN ENGLISH

South African accents of English (SAE) demonstrate a shift in the realizations of front vowels to a generally closer or more central position than those found in most British English accents (see, for example, Wells [1]). This results in a set of vowels /æ, e/, and some contexts of /l/ that may potentially be confused by British English listeners with their set /e, ɪ, ə/, with /ə/ in monosyllables probably heard as /ʌ/. The question then arises as to what happens to /ʌ/; does this vowel also shift in SAE, and if so where to?

Normally context disambiguates any potential perceptual confusions between SAE and other accents; nevertheless, there are many possible homonymic clashes.

A further point to be considered is that SAE like many other English accents has a velarized variant of /l/ following vowels ([ɫ]: the 'dark-l'). As noted, for example, by Gimson [2], following dark-l has a tendency to centralize the preceding vowel. It is possible, therefore, that the short-vowel shift of SAE may be affected by this, in that centralization may have differential effects on perception.

The authors decided to design a perception experiment to test whether these short front vowels in SAE would indeed be perceived by British English listeners to be the shifted values when all contextual cues as to meaning are removed. This would also prove an opportunity to test the /ʌ/ vowel, and see whether it too would be perceived as shifted, and if so, to which vowel.

This experiment would then be follow-ed by an acoustic analysis of the SAE vowels, and a comparison of their formant values with those of British English.

PERCEPTION EXPERIMENT

Method

Four speakers were used in the perception experiment. Details of the speakers are given in Table 1.

Table 1. Details of the SAE Speakers.

	sex	age	area
speaker 1	M	20	Jo'burg
speaker 2	M	19	Durban/J'burg
speaker 3	M	20	Jo'burg
speaker 4	M	20	Durban

All the speakers were students at the University of the Witwatersrand, and were first language speakers of English.

The speakers were all recorded in good acoustic conditions on a DAT recorder by the first author. The material recorded consisted of a set of twenty different words embedded in the phrase "say the word ___ again". The words used are given in Table 3 below, and it can be seen that as well as the front vowels noted above, a wider range of vowels was included. This allowed investigation of whether other vowel confusions were present, as well as acting as distractors from the main set.

The subjects involved in the listening task were all first year Speech and Language Therapy students between the ages of 18-30, at two institutions in the UK. All had followed a course of one semester in phonetics and practical phonetics, but had not studied different accents of English. The details of the listeners are given in Table 2.

The perception task was undertaken by the last two authors at their institutions. An answer sheet was prepared (see Table 3), which listed for each token the target pronunciation, a first foil (in the case of the front vowels, the predicted changed version), and a second foil, more distant phonetically. Listeners had to mark which of the three words they heard in each instance. Targets and foils were randomized for each token. Separate answer sheets were used for each of the four speakers. Listeners were informed that the four speakers would not necessarily use the same targets, and that repetitions of targets by individual speakers was also possible.

There was a short gap between each token to force immediate responses from the listeners. All four speakers were presented in a continuous session, but no learning effect was seen in the results. Listeners were told to make a choice in each instance, and very few unmarked examples were found.

Table 2. Details of the Listeners.

	male	female	total
S. England/RP	2	10	12
Midlands	0	6	6
N. England	0	4	4
N. Ireland	0	13	13
S. African Eng	0	1	1

Table 3. Answer Sheet for the Perception Task. The Targets are in italics, and the First Foils are underlined.

1	hit	*heat*	heart
2	*hit*	hot	hut
3	putt	pat	*pit*
4	pal	*pill*	pull
5	*pool*	Paul	pull
6	pit	part	*pet*
7	still	*sell*	seal
8	soul	sell	*Sal*
9	pet	*pat*	port
10	pout	port	*pot*
11	*putt*	pat	peat
12	poot	part	*put*
13	pit	*peat*	part
14	pill	pall	*peal*
15	pod	*paired*	ped
16	*part*	port	put
17	pot	peat	*port*
18	beat	*boot*	but
19	*bird*	bead	bud
20	Paul	pull	*pearl*

Results

Results for all 36 listeners for all four speakers were calculated for each token in the experiment. While there was a certain amount of difference between the scores of the listeners (some of which appears to be attributable to their regional back-ground), and between the scores given to the four speakers, there was generally good agreement. It is hoped to explore what differences there were in greater depth elsewhere. In Table 4 below the total scores for all listeners for all speakers are given. The maximum possible score for any one token is 144; as noted above, a few instances of non-scoring occurred, and this, together with rounding percentages up or down, accounts for why the scores for some tokens do not reach 100%.

Table 4. Results for All Listeners and All Speakers in % out of 144 for each Token.

	%		
	Target	1st Foil	2nd Foil
1. heat	97%	3%	0%
2. hit	100%	0%	0%
3. pit	13%	88%	0%
4. pill	1%	98%	0%
5. pool	92%	8%	0%
6. pet	15%	85%	0%
7. sell	95%	2%	1%
8. Sal	35%	28%	37%
9. pat	7%	93%	0%
10. pot	90%	8%	2%
11. putt	59%	41%	0%
12. put	90%	10%	0%
13. peat	97%	3%	0%
14. peal	92%	8%	0%
15.paired	86%	14%	0%
16. part	58%	37%	3%
17. port	92%	7%	1%
18. boot	99%	1%	0%
19. bird	99%	1%	0%
20. pearl	84%	15%	1%

These results confirm that the major area of perceptual confusion for the British English listeners was with the short front vowels. For example 'pit' had an 88% score for its first foil ('putt'), 'pet' an 85% score for the main foil 'pit', and 'pat' a 93% score for the foil 'pet'. This confirms the predicted pattern of change. We were also interested in the behaviour of the central vowel /ʌ/, and if we examine the score for 'putt' we find that while on 59% of occasions it was

heard as 'putt', there was a considerable number of identifications (41%) as 'pat', which would suggest at least that these four vowels are shifting in a circular fashion. It would certainly appear important to include /ʌ/ in the acoustic study.

An exception to the trend just reported occurs with the token 'hit', where no instances of identification as 'hut' were recorded. This compares with 'pit' where, as just noted, only 13% identified the token as containing the /ɪ/ vowel. In South African English, however, /ɪ/ is noted as abetting raising, but blocking lowering of /ɪ/ [1]; this result confirms that characteristic, and a comparison of the acoustic aspects of the two allophones of this vowel is given below.

The effect of following dark-l on vowel identification is quite striking in these results. The token with the target high vowel, 'pill', was almost always heard as 'pull', while 'sell' was not heard as the raised equivalent 'sill', but was correctly heard as 'sell' in 95% of occasions. The low target vowel in 'Sal' caused the most confusion (and indeed showed quite an amount of variation between the four speakers). Both the second and third foils ('sell' and 'soul') scored well, though differentially between the speakers). The explanation for all these results clearly lies in the centralizing effect of following dark-l, as noted in the introduction. This effect will reinforce the movement of target /ɪ/ (and the increase in gravity lead to a perception of the rounded /ʊ/), but centralization of target /e/ does not bring it into conflict with any of the vowels in the foils. With target /æ/, the increased gravity with dark-l will lead listeners to expect a retracted but rounded vowel, thus causing the confusion seen in the results.

These results also show some con-fusion with several other target vowels, including 'paired', 'part', and 'pearl'. Some of the difficulty with these is possibly due to interference from the listeners accents (e.g. rhoticity in some cases), however they may reflect aspects of SAE as well. It is hoped to explore these results more fully elsewhere.

It is interesting to note that the one South African English listener did score higher on identifying target vowels, but was only marginally better than the average.

Table 5. Formant Values in Hz for the Test Words.

		F1	F2
'hit'	spkr 1	330	2175
	spkr 2	395	1656
	spkr 3	433	2047
	spkr 4	350	1798
	SBS	374	2165
'pit'	spkr 1	364	1498
	spkr 2	434	1494
	spkr 3	358	1672
	spkr 4	391	1601
	SBS	433	2056
'pill'	spkr 1	378	762
	spkr 2	399	798
	spkr 3	447	964
	spkr 4	492	886
	SBS	524	1296
'pet'	spkr 1	443	1671
	spkr 2	392	1891
	spkr 3	363	1979
	spkr 4	367	1844
	SBS	690	1996
'sell'	spkr 1	522	1383
	spkr 2	514	1322
	spkr 3	594	1351
	spkr 4	601	1201
	SBS	707	1735
'pat'	spkr 1	541	1960
	spkr 2	598	1721
	spkr 3	554	1809
	spkr 4	600	1565
	SBS	784	1615
'Sal'	spkr 1	663	1026
	spkr 2	564	1194
	spkr 3	602	1534
	spkr 4	581	1303
	SBS	783	1501
'putt'	spkr 1	568	1605
	spkr 2	571	1430
	spkr 3	581	1598
	spkr 4	614	1319
	SBS	795	1230
'put'	spkr 1	466	1240
	spkr 2	444	1414
	spkr 3	369	1576
	spkr 4	396	1236
	SBS	454	988
'pull'	SBS	486	695

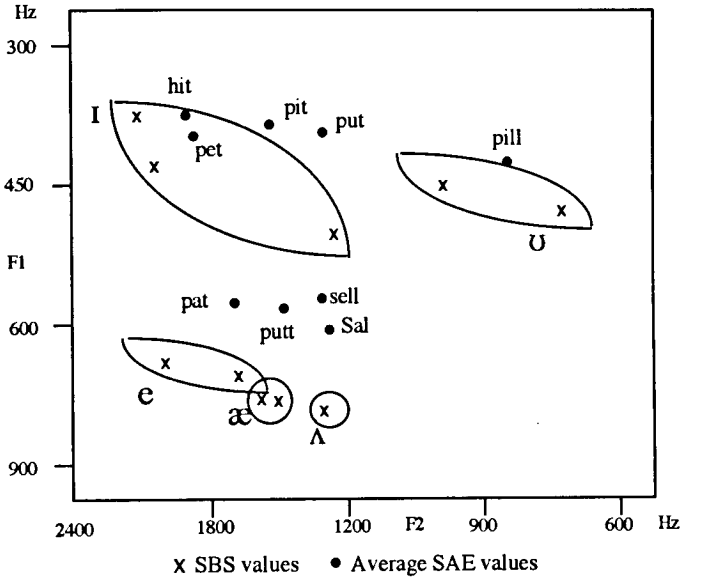


Figure 1. F1-F2 Plot for the SBS and SAE vowels.

ACOUSTIC STUDY

The acoustic study investigated the first two formants of the vowels in the following tokens: (2) 'hit', (3) 'pit', (4) 'pill', (6) 'pet', (7) 'sell', (9) 'pat', (8) 'Sal', (11) 'putt', (12) 'put', and for comparison the token 'pull' from the SBS speaker only.

The tokens were analysed on the Kay CSL™ 4300, software version 4.01, using the FFT function. Vowels were measured by placing the cursor on a central part of the vowel, avoiding formant transitions. In Table 5, the formant values for F1-F2 are given for the four SAE speakers, and for one SBS speaker (the second author), recorded reading the same list of words in the same conditions as the SAE speakers. Values from previous studies of SBS vowels [3] were similar to those reported here.

Figure 1 shows the average F1-F2 values for all four SAE speakers, compared to the SBS speaker. This clearly shows the reasons for some of the identifications, but suggests that other may well have been an artefact of the foils presented. It also suggests that switching from one identification to another may require further acoustic movement with some vowels as compared to others, as the difference between 'pit' and 'putt' demonstrates.

The figure also shows a general centralization of many vowels, not solely those with following dark-l. Nevertheless, the considerable raising of /æ/ and /e/ in 'pat' and 'pet' is clearly demonstrated, together with a centralization of /ɪ/ in 'pit'. The /ʌ/ in 'putt' fronts and raises only slightly, which may account for the ambiguous response to this vowel from the listeners. The other allophone of /ɪ/, that does not undergo centralization ('hit') is raised compared to SBS.

The authors hope to explore this whole topic more fully in further work, including the full set of monophthongs and diphthongs in SAE.

REFERENCES

[1] Wells, J. (1982), *Accents of English 3: Beyond the British Isles*, Cambridge: CUP.

[2] Gimson, A. (1989), *An Introduction to the Pronunciation of English*, 4th ed., London: Edward Arnold.

[3] Fry, D. (1979), *The Physics of Speech*, Cambridge: CUP.

PERCEPTION OF VOWEL QUALITY BY GERMAN-ENGLISH BILINGUALS

Frederika Holmes
Dept. of Phonetics and Linguistics, University College London, UK

ABSTRACT

German-English bilinguals' labelling of the front vowel space in each of their language modes was investigated using a synthesised vowel continuum. Results showed that bilinguals achieved native-like performance in the English condition, but that their performance in German was affected by their experience in English.

INTRODUCTION

This study examines the extent to which perceptual categories for vowels are language-specific, and investigates how bilinguals process potential conflicts between the phonological categories of their two languages.

The conflict investigated in the present experiment was the division into phonological categories of the front vowel space. Both English and German have an open-mid front vowel /e/ and an open central vowel /a/ which are phonetically similar in both languages [1]. (The symbol /a/ will be used for the vowel in German *Bad* and in English *but* (usually transcribed /ʌ/), to avoid using different symbols for similar vowels).

However, the English front open vowel /æ/ has no equivalent in German, and evidence indicates that German speakers have difficulty in developing a stable category for English /æ/, and either identify it with an adjacent German vowel [1], or use a category based on different acoustic dimensions to those of English natives [2].

The question of interest in the present study was to what extent the German-English bilinguals had succeeded in acquiring and maintaining native-like categories for the front vowel space in each of their languages, or whether the co-existence of the two systems affected perceptual categories in one or both of the languages.

METHODOLOGY

Test material

The stimuli consisted of synthesised CVC syllables in which the formant structure of the vowel portion was varied to create a continuum between three fixed points corresponding to /e æ a/.

The vowel portions of the three fixed points, based on acoustic measurements of similar syllables spoken by native English and German speakers, were synthesised through the cascade branch of a Klatt synthesis system. The acoustic characteristics of these three vowels were as follows:

/e/ F1 650 Hz, F2 1900 Hz, F3 2640 Hz; F4 4000 Hz; F5 4500 Hz
/æ/ F1 800 Hz, F2 1550 Hz, F3 2460 Hz; F4 4000 Hz; F5 4500 Hz
/a/ F1 700 Hz, F2 1250 Hz, F3 2550 Hz; F4 4000 Hz; F5 4500 Hz

The fundamental frequency for all three vowels was 100 Hz at onset and 85 Hz at offset; the amplitude was 45 dB and the duration of the vowel was 100 msecs.

A continuum of vowel quality was then created by logarithmic interpolation of a further five values for F1, F2 and F3 between each pair of fixed points. Other characteristics were held constant. This resulted in a thirteen-point continuum of formant structure, ranging from /e/ through /æ/ to /a/.

These vowel tokens were inserted between consonants synthesised through the parallel branch of a Klatt synthesis system, to produce CVC syllables. The consonant frame used for the English condition was /b_t/, giving the possible English words *bet*, *bat* and *but*; for German the context was /f_st/, giving the possible German words *fest* and *fast*.

The entire continuum consisted of 13 synthetic syllables, which were presented in 10 randomised blocks, giving a total of 130 stimuli, preceded by a practice block consisting of an additional 13 randomised steps.

Subjects

Subjects were 12 German-English bilinguals with a range of language-backgrounds and patterns of acquisition. Some were childhood bilinguals; others had acquired the second language as adults. All spoke both languages to a very high level, had spent time living in both countries and used both languages on a regular basis. Bilingual subjects were matched for language-dominance on the basis of data extracted from a questionnaire (after [3]).

In addition six monolingual speakers of each language were tested. The English monolinguals were first year Speech Science students at University College London; the German monolinguals were students of the Fachbereich Computerlinguistik at the Universität des Saarlandes in Saarbrücken. All subjects were paid for their participation.

Test procedure

Testing took place in soundproofed rooms at UCL, and at the University of Saarbrücken.

Monolingual subjects were tested in a single session, and bilingual subjects in two sessions, one in each language. Half the bilinguals were tested in English first, the other half in German first; in either case, the two sessions were conducted at least two weeks apart.

The tests were conducted as part of a wider series of tests, and considerable care was taken to place subjects in the appropriate language mode. All conversation and instructions took place in the test language, and subjects were asked to read aloud several texts in the test language before the experiment began.

The stimuli were played on a Marantz audio cassette recorder, and presented to subjects binaurally via Sennheiser HD414 headphones. The task was an open-labelling one: subjects were asked to write down on the response sheet the word in the test language which most resembled each stimulus heard; the practice block was conducted first to ensure that subjects had understood the task and had the chance to familiarise themselves with the material.

RESULTS

A Maximum Likelihood Estimate procedure was used to produce a cumulative normal function (probit analysis) for each subject's set of responses, and the parameters of phoneme boundary (PB) and function gradient (slope) were extracted to characterise the categories perceived by the subjects. Mean results for each group were then established.

Table 1. Location of phoneme boundaries and function gradient for different subject groups

		/e-æ/	/e-a/	/æ-a/
Eng mono	PB	4.127		8.199
	slope	2.302		-2.650
Ger mono	PB		6.468	
	slope		1.383	
Eng biling	PB	3.948		8.306
	slope	1.748		-2.348
Ger biling	PB		5.725	
	slope		2.620	

The most immediately striking feature of this data is the clear split between the responses in the English condition, in which all subjects perceived two phoneme boundaries, and the responses in the German condition in which all subjects perceived one boundary only.

Figures 1-4 below characterise the different categories used in each of the language conditions. The boundaries between the three categories for the English monolinguals are marked by steep curves and low inter-subject variability (as measured by the one standard deviation error bars). This suggests that for the English monolinguals the three categories spanned by the continuum were stable and clearly defined, with sharp boundaries.

For German monolinguals the boundary between the two categories in the German condition is marked by a much shallower curve, with a high degree of inter-subject variability. This suggests that that part of the continuum corresponding to English /æ/ was not reliably identified by German subjects. The location of the phoneme boundary in German confirms this impression, since it occurred between steps 6 and 7, which corresponds exactly to the midpoint of the continuum, the default location for a boundary which is not mediated by phonetic considerations [4].

The labelling behaviour of the bilinguals is more complex. Although all bilinguals had a category corresponding to English /æ/, the boundary curves are shallower than those for English

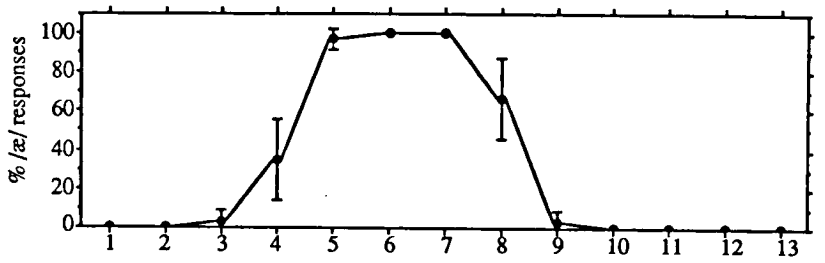


Figure 1. % /æ/ responses: English monolinguals

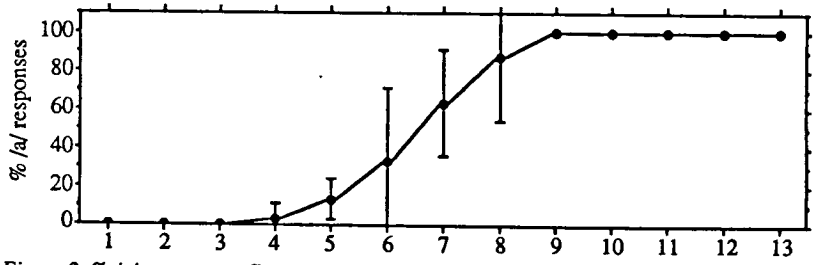


Figure 2. % /a/ responses: German monolinguals

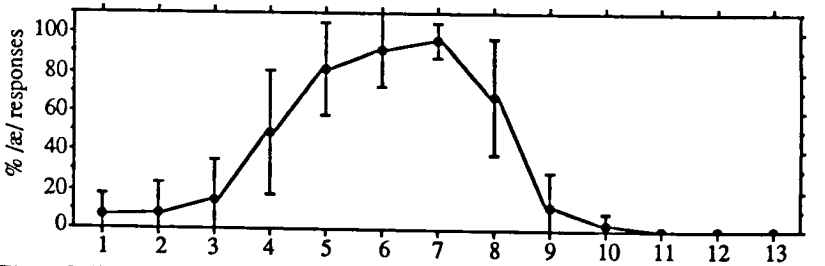


Figure 3. % /æ/ responses: Bilinguals English condition

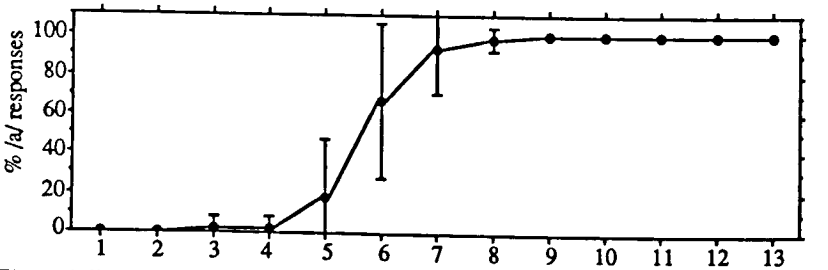


Figure 4. % /a/ responses: Bilinguals German condition

monolinguals, and characterised by a higher degree of variability, particularly for the /e-æ/ boundary. This suggests that the bilinguals' categories in the English condition were less stable than those of the monolinguals.

A further interesting finding is the labelling behaviour of the bilinguals in the German condition. Unlike the German monolinguals, for whom the /æ/ category was simply divided at the midpoint of the continuum, the boundary curve for bilinguals in the German condition shows a clear skew to the left, and is steeper than that of the monolinguals. This suggests that the bilinguals' categorisation of the front open vowel in the German condition was mediated by their linguistic experience of the English /æ/ category, although this would not have been relevant to the German task.

A statistical analysis in the form of a t-test for two independent samples was performed on the values for phoneme boundary and slope extracted from the MLE procedure. Comparison of the bilinguals' results in each language condition confirm that they were able to match the performance of the English monolinguals in labelling English categories. There was no significant difference between monolinguals and bilinguals in the English condition either with regard to the location of phoneme boundary for the /e-æ/ boundary ($t=0.50$ ($df=16$) $p>0.05$) nor the /æ-a/ boundary ($t=0.48$ ($df=16$) $p>0.05$). The slope of the two boundaries for the bilinguals also did not differ significantly from that of the monolinguals ($t=1.22$ ($df=16$) $p>0.05$ and $t=0.96$ ($df=16$) $p>0.05$ respectively).

However, in the German condition, the performance of the bilinguals did differ significantly from that of the monolinguals, both for PB ($t= 2.89$ ($df=16$) $p<0.05$) and slope ($t=2.42$ ($df=16$) $p<0.05$).

DISCUSSION

The fact that all bilinguals perceived the appropriate number of categories in each language condition shows that they were able to code-switch in their perception according to the language-set they were in. This clear between-language difference is all the more striking in view of the fact that the task was an open labelling one, which did not predispose subjects to choose a particular number of categories for their responses.

Moreover, the finding that the group results for bilinguals in the English condition matched the performance of English monolinguals shows that the categories they had developed were sufficienctly accurate and stable to enable them reliably to label vowels corresponding to the English sequence /e-æ-A/. Since most of the subjects were not childhood bilinguals, this suggests that a late age of learning is not necessarily an obstacle to the formation of new phonological categories.

In contrast, the findings for the German condition show that bilinguals as a group do differ from monolinguals in their categorisation of stimuli corresponding to English /æ/. The leftward shift in the bilingual labelling function as compared to the German monolingual labelling function suggests that the bilinguals are using phonemic criteria in their labelling behaviour, since use of acoustic criteria would produce a boundary at the midpoint of the continuum. Since there are no German phonetic criteria which are relevant to this category, it appears that the bilinguals' experience with English influenced their categorisation of the continuum in the German condition.

It seems that the bilinguals' ability to acquire and maintain English categories in the face of the competing German standard was matched by a move away from monolingual categorisation in the German condition. In other words, it is possible to acquire native-like categories for a second language, but that improved performance in the L2 may be matched by decreasing nativeness in L1.

ACKNOWLEDGMENT

This work was funded by a research studentship from the Medical Research Council.

REFERENCES

[1] Barry, W.J. (1989), "Perception and production of English vowels by German learners: Instrumental-phonetic support in language teaching", *Phonetica*, vol. 46, pp. 155-168.

[2] Bohn, O.-S. and Flege, J. (1990), "Interlingual identification and the role of foreign language experience in L2 vowel perception", Applied Psycholinguistics, vol. 11, pp. 303-328.

[3] Hazan, V. and Boulakia, G. (1993), "Perception and production of a voicing contrast by French-English bilinguals", Lang. & Speech, vol 36 (1), pp. 17-38.

[4] Rosen, S. (1979), "Range and frequency effects in consonant categorisation", *J. Phonetics*, vol. 7, pp. 393-402.

PERCEPTION OF SYNTHETIC VOWEL STIMULI WITH CHANGING ONSET F2-TRANSITION BY RUSSIAN AND FINNISH SUBJECTS

*V. Kouznetsov**, *A. Iivonen***
**Moscow Linguistic University, Moscow, Russia*
***Helsinki University, Helsinki, Finland*

ABSTRACT

The aim of the present paper is to investigate the perceptual role of the direction of the onset F2-transition in determining the phonetic vowel quality and to study language specific aspects of auditory analysis.

INTRODUCTION

Contemporary theories on vowel perception may be roughly divided into two major classes: those that consider information conveyed by onset and offset formant transitions essential to vowel identification (dynamic-specification models), and those assuming that all necessary information is contained at the vowel nucleus (target models).

Russian language provides a good opportunity to test these hypotheses due to the specific allophonic variation of stressed vowels in the context of palatalized and nonpalatalized consonants. Figure 1 depicts five russian vowel phonemes pronounced in the symmetrical contexts of palatalized and nonpalatalized fricative [s]. Formant frequencies were measured at the middle of the vowels [1]. Apostrophes surrounding vowels indicate palatalized environment. The plot reveals that several pairs of allophones can not be discriminated using F-pattern at the vowel nucleus, for example: ['U'-ɨ, ɨ-'O']. None the less, listeners rarely confuse these vowels. This is explained not only by the fact that the vowels occur in different consonant environments but by the differences of transition sections as well: at the onset of ['U'] F2 glides down, while at the onset of [ɨ] (Russian symbol for this vowel is [ы]) it rises up or stays level. Attempts to assess the perceptual significance of the direction of F2 transition in recognition of ['U] and [ɨ], using forced-choice identification of synthetic stimuli, produced contradictory results [2,3].

In the present study the same problem is addressed, employing different synthesizer and experimental procedures. To discover traces of language specific behavior the experiments reported below were conducted on russian and finnish subjects.

EXPERIMENT 1

In this experiment a modified ABX procedure was used to test the listeners ability to discriminate stimuli on the basis of the direction of F2-transitions.

METHOD

Stimuli. Test stimuli were synthesized using Klatt software synthesizer. Frequency and bandwidth of F1, F3, and F4 were kept constant: F1 = 300 Hz, B1 = 50 Hz; F3 = 2300 Hz, B3 = 200 Hz; F4 = 3300 Hz, B4 = 250 Hz. The frequency of the F2 was changed from 700 to 1900 Hz in 100 Hz step. The stimuli had either steady F2 or an onset F2-transition of ± 200 Hz or ±100 Hz was added to them. Stimulus duration was 150 ms, transition duration - 55 ms. A falling F0-contour was used: 127-100 Hz. The voice amplitude increased from 55 to 60 db during the first 10 ms and fell down to 31 db on the last 35 ms. The stimuli were sampled at 10 kHz via a 10-bit D/A converter. A spectrogram of two tokens is presented in Figure 4.

Two experimental tapes were recorded using a speech processing tool-kit ISA (designed by R. Toivonen). Each tape contained 84 randomly ordered triads composed of stimuli with F2 transition either of ± 200 or ± 100 Hz. A triad consisted of stimuli differing only in direction of F2 transition. The third triad element was also realized by a vowel with an appropriate steady F2 (Z-element). Thus there were three types of triads: ABA, ABB, and ABZ. Both temporal orders of the first two elements were used in triad construction. The ISIs were 500 ms within triads, 2 s between triads and 5 s after each group of 15 triads.

Subjects and Procedure. Russian listeners were 14 students (19-20 years old) who took up an introductory course in phonetics. Ten finnish subjects took part in the experiments, their age varied from 23 to 52 years. Presentation was over loudspeakers in a quite room. The task was to identify the third vowel in a triad as A or B by circling the appropriate response on an answer sheet.

The tape composed of the stimuli with ΔF2 = 200 Hz was presented first. It was preceded by 9 stimuli for familiarization and 10 practice triads. All perception test, including paired comparison of the Experiment 2, were conducted in one session that lasted about 45 minutes.

RESULTS

In order to find out whether the group of subjects was able to discriminate between stimuli A and B a χ^2-test was applied to each type of triads separately. Under the null hypothesis (no discrimination) the subject responses would split evenly between A and B and the expected frequency would be equal to half the number of subjects in a group.

The best discrimination performance of russian listeners was on ABB triads with ΔF2 = 200 Hz: the probability of correctly rejecting the null hypothesis was 0.93 (χ^2 = 39.3 with 28 d.f.). On 19 (out of 29) ABB triads not less than 9 subjects identified the third stimulus correctly. Only in one case the listeners majority made a mistake.For the other two types of triads the level of significance was well above 0.10.

On ABA triads the russian listeners responded consistently (distribution of responses was at least 9/5 or 5/9) in 14 trials out of 28, but in half of them their judgement was incorrect. The same degree of response consistency was reached 14 times on the ABZ triads where Z was identified as B (in 10 cases B was realized by a stimulus with rising F2-transition).

The discrimination of stimuli with ΔF2 = 100 Hz was more poor, but the general pattern of the responses was the same as described above.

The data of the finnish listeners is not reported here for it requires special treatment since the subjects were allowed not to take any decision in case of doubt.

EXPERIMENT 2

Trying to assess in this experiment the role of the direction of F2-transition in identification of ['U] and [ɨ] we did not consider it correct to ask russian and finnish listeners to perform the task of recognition because [ɨ] is an alien sound to Finnish language while ['U] is "unknown" to naive russian listener as a distinct vowel category for it does not occur in isolation and therefore its phonetic quality must be abstracted from the context. The perceptual importance of the direction of F2-transition may be studied by asking listeners to judge (dis)similarity between stimuli with rising and falling F2-transition and the two standard stimuli with steady F2 that are most similar to ['U] and [ɨ].

METHOD

In this experiment the same set of stimuli was used as in Experiment 1.

The stimulus with a steady F2 = 1700 Hz was used as a token of ['U]. Its formant frequencies are quite close to those of Finnish short [y] (see Figure 1. [4]). The second standard stimuli had F2 = 1400 Hz and as it was revealed by pilot observations its phonetic categorization was uncertain. If a rising F2-transition was added to it, it was definitely perceived as [ɨ].

Every stimulus with F2-transition was paired with the two standards in both orders. 10 times each standard was paired with itself. Separate tapes, each containing 130 pairs, were created for stimuli having ΔF2 = 200 and 100 Hz. The stimuli with greater F2-transition were presented first. All the other experimental conditions were the same as in Experiment 1.

Procedure. The subjects were asked to judge each pair for its dissimilarity on a five-point scale on which 1 was considered most similar and 5 least similar.

RESULTS

Figures 2 and 3 show the medians of the dissimilarity judgements for each pair of stimulus and standard (regardless of their order) plotted against F2 values at the stimulus steady-state. Minimums of the curves point at the F2 of an appropriate standard. Examination of the Figures leads to the following conclusions. For both groups of listeners the main factor affecting similarity estimate is the difference between stimulus F2 value at the steady-state and that of the standard. The effect of the direction of F2-transition if ΔF2 = 100 Hz is negligible. When ΔF2 = 200 Hz and stimulus F2 was lower than that of the standard, stimuli with falling F2-transition were judged more similar to both standards than stimuli with rising F2. The relationship was reversed when stimulus F2 was higher than that of the standard. It seems, that the listeners based their judgement of similarity not only on comparison of the physical properties of the sounds but on categorical decisions as well: in Figure 3 stimuli having F2 ≤ 1000 – 1100 Hz were estimated equally similar to the standards though the difference in F2 continued to grow.

From the results reported and discussed above it is apparent that in the present experimental paradigm no evidence was obtained supporting the hypothesis that russian listener had some language specific rules for categorical interpretation of the direction of F2-transition.

ACKNOWLEDGEMENT

We are indebted to our students J. Välikangas, K. Laasonen and T. Sertun for their help in preparation of the experimental tapes, carrying out the experiments and a part of data analysis. Due to the telecommunication problems the major part of the responsibility for the paper contents lies with the first author.

REFERENCES

[1] Kouznetsov, V., and Ott, A. (1987), "Spectral properties of russian stressed vowels in the context of palatalized and nonpalatalized consonants", *Proc. of XIth ICPhS*, v.3, pp. 117-120.

[2] Kouznetsov, V. (1991), "A cross-linguistic study of perception of vowel stimuli with changing F2" (in Russian), *Proc. of the XVIth All Union seminar on automatic recognition of auditory images*, pp. 160-162.

[3] Lublinskaya, L. (1993), "Application of phonetic criteria in speech psycho-acoustics" (in Russian), In: *Issues in phonetics I* (eds. T. Nikolaieva et al.), pp. 274-287.

[4] Iivonen, A., and Laukkanen, A. (1993), "Explanations for the qualitative variation of finnish vowels", *Studies in logopedics and phonetics*, v.4, pp. 29-54.

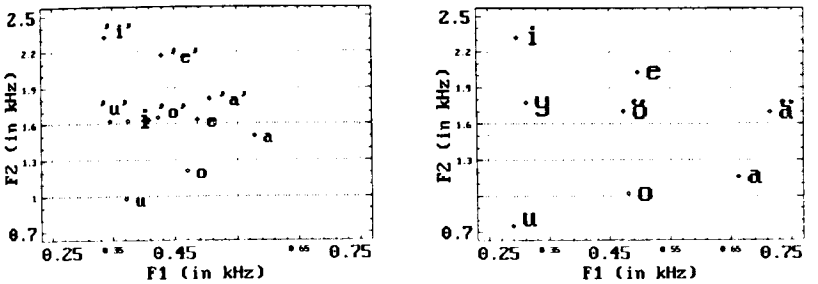


Figure 1. Vowel formant space. Left - russian stressed vowels in the context of palatalized and plain consonants. Right - finnish stressed vowels (data is averaged across long and short vowels).

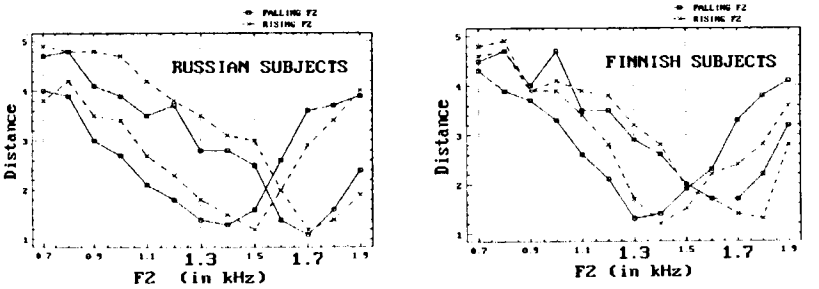


Figure 2. Median distances between two standards (steady F2 = 1400/1700 Hz) and stimuli with F2-transition of 200 Hz. Left - russian data, right - finnish data.

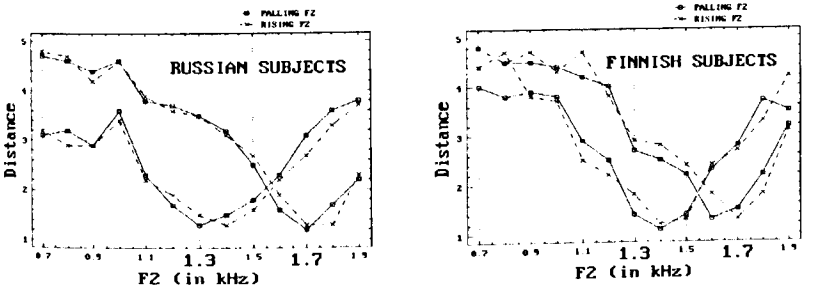


Figure 3. Median distances between two standards (steady F2 = 1400/1700 Hz) and stimuli with F2-transition of 100 Hz. Left - russian data, right - finnish data.

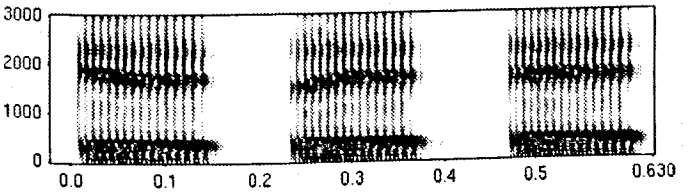


Figure 4. Spectrogram of test stimuli.

CROSS-LINGUISTIC COMPARISON OF DURATIONAL PATTERNS IN FINNISH AND FINLAND-SWEDISH

Ilse Lehiste
Ohio State University

ABSTRACT

The paper describes durational patterns in Swedish and Finnish trochaeic verse produced by speakers of Swedish and Finnish in Turku. Both groups of speakers exhibit a combined prosodic system that contains more contrasts than either of their two languages.

INTRODUCTION

The paper is one of a continuing series of reports about my study of the phonetic manifestation of metrical structure in orally produced poetry. The rationale for the study, as well as the methodology, have been described in a number of previous publications [1]. In the study, I investigate the ways in which the same metric structures are realized in languages that have different supra-segmental structures.

At the same time, the paper deals with another continuing research interest - the topic of language contact, especially in the region of the Sprachbund around the Baltic Sea [2]. The special focus of the current report is the relationship between tonal and durational patterns found in the speech of Turku Finns and Swedes, and the Swedish spoken in Stockholm.

MATERIALS AND METHODOLOGY

The texts analyzed in this context include the poems "Vastavirtaan" by Juhani Siljo and "Bonden Paavo" by J.L. Runeberg. The recordings were made in September 1988 in Turku and in June 1989 in Stockholm. The tapes were analyzed at the Ohio State University's phonetics laboratory in Columbus. Measurements include the duration of lines, pauses between lines, metric feet, and various subparts of the metric feet. Three fundamental frequency measurements were made for every metric foot that consisted of a disyllabic word: Fo value at the beginning, peak, and end of the Fo curve of every syllable, as well as the position of the Fo peak within the syllable. The present paper describes patterns found in medial (i.e. non-initial, non-final) metric feet. Results are averaged from productions by four speakers in each group.

FUNDAMENTAL FREQUENCY

The basic fundamental frequency patterns found in the speech of the Stockholm speakers have been reported before [1]. Table 1 (below) presents average Fo values for a subset of Stockholm speakers in words with Accents 1 and 2, and in the same words produced by speakers of Turku Swedish (for whom Swedish is first language) and Turku Finns (for whom Swedish is second language). The difference in absolute values depends on characteristics of the speakers and is irrelevant with regard to the shape of the fundamental frequency contour.

As can be seen from the table, the Stockholm speakers maintain a distinction between the two accents, the characteristic feature of which is the low Fo value at the end of the first syllable in disyllabic words with Accent 2 (printed in boldface). The Turku Swedes and Finns have essentially the same pattern in words expected to carry the two accents.

Both groups of Turku speakers used the same patterns in their production of Finnish metric feet.

DURATION

Figure 1 (below) shows the average durations of short and long syllable nuclei in metric feet consisting of Finnish disyllabic words. Finnish has four possible quantity patterns in disyllabic words, determined by syllable length: Short-short, Short-long, Long-short, and Long-long. For the sake of comparison, syllables of the same contrastive duration occurring in the same position were averaged together: CV- stands for the initial short syllable in Short-short and Short-long words, CVV- stands for the initial long syllable in Long-short and Long-long words etc.

As may be seen from the figure, both groups of Turku speakers maintain a clear distinction between contrastive short and long syllable nuclei in their Finnish productions.

Figure 2 (below) shows the average durations of first vowels in Swedish metric feet with (intended) Accent 1 and 2 produced by Swedish speakers in Stockholm and Turku, and by Finnish speakers in Turku reading the same material. As may be seen from the figure, all three groups of speakers employ comparable durations. The same is the case for vowels in second syllables, whose duration is similar to that of the Finnish short syllable nuclei.

DISCUSSION

The study shows that both groups of Turku speakers keep the durational systems of their two languages clearly apart. The Turku Swedish speakers share the Stockholm speakers' long vowel durations, and have acquired the durational system of Turku Finnish. On the other hand, the Turku Finnish speakers have acquired the Swedish long vowel duration - intermediate between the Finnish contrastive short and long durations - and use it in their Swedish.

The intimate contact situation has led to additional similarities. For example, both groups of speakers use the half-long vowel in the second syllable of words like yli. Finns reading Swedish distinguish two intervocalic consonant durations in words like diken vs. döden (119 msec vs. 50 msec for the Finns, 110 vs. 52 msec for Turku Swedes, and 147 vs. 59 for Stockholm Swedes), and both groups use a third duration in the word miekka (210 msec for Finns, 233 msec for

Finland-Swedes). Both groups also pronounce the cluster /lv/ in the words tulva and tulvaa with a short epenthetic shwa inserted within the cluster. (The Swedish text did not contain any /lv/ clusters; it would have been very interesting to see whether the speakers carried this phenomenon over into their production of Swedish.) There were also occasional realizations of Swedish words like lade, sade, and hagel with a short first syllable vowel, again by both groups of Turku readers.

Both groups of speakers exhibit a combined prosodic system that contains more contrasts than either of their two languages. The short vowels of the two systems can be considered identical, but the speakers distinguish two contrastive durations in Finnish and have a long vowel duration in Swedish that is intermediate between the short duration and the Finnish long duration. The short intervocalic consonants of the two systems can be likewise considered the same, but both groups of speakers have two additional long intervocalic consonants - one for Swedish words like maka, the other for Finnish words like miekka. The Finnish of Turku Swedish speakers is clearly dialectal and thus orally acquired, as shown by the occurrence of short shwa in the intervocalic /lv/ cluster that speakers of the local Finnish dialect produce, but do not hear [3]. The uniformity of the patterns within the two groups shows that the systems are quite stable.

REFERENCES

[1] Lehiste, I. (1992), "The phonetics of metrics", Empirical Studies of the Arts, vol 10(2), pp. 95-120.
[2] Lehiste, I. (1978), "Polytonicity in the area surrounding the Baltic Sea", Nordic Prosody. Papers from a Symposium. Travaux de l'Institut de linguistique de Lund 13, Lund University. Pp. 237-247.
[3] Kalevi Wiik, personal communication.

Table 1. Average Fo values in medial metric feet in the Swedish-language poem "Bonden Paavo" read by Swedish subjects in Stockholm (St.Sw.) and Turku (T.Sw.), and Finnish subjects in Turku (T.F.). A 1 = Accent 1, A 2 = Accent 2.

	First syllable	Second syllable
St.Sw.,A 1	160 - 172 - 161	157 - 158 - 151
St.Sw.,A 2	167 - 168 - 134	160 - 165 - 156
T.Sw.,A 1	184 - 192 - 175	168 - 171 - 155
T.Sw.,A 2	176 - 184 - 166	168 - 170 - 154
T.F., A 1	116 - 120 - 108	94 - 102 - 91
T.F., A 2	114 - 118 - 106	102 - 103 - 93

Figure 1. Average duration of short and long syllable nuclei occurring in Finnish disyllabic metric feet produced by speakers of Swedish and Finnish in Turku. CV- = short initial syllable, CVV- = long initial syllable, -CV = short second syllable, -CVV = long second syllable.

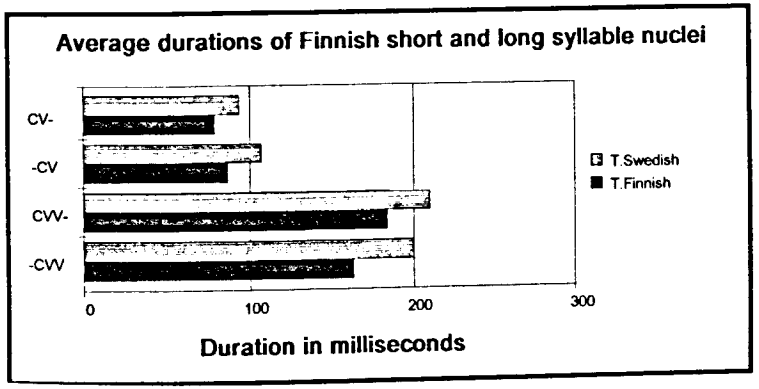
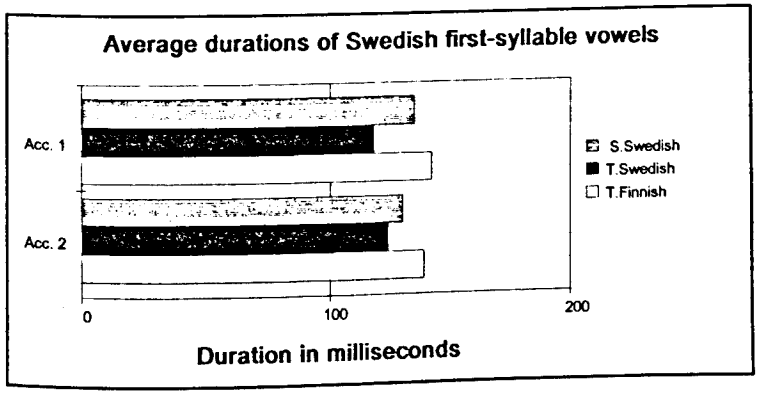


Figure 2. Average durations of first-syllable vowels in Swedish disyllabic metric feet, produced by speakers of Swedish in Stockholm and Turku, and by speakers of Finnish in Turku.



PROSODIC PATTERNS IN SINGAPORE ENGLISH

Ee Ling Low† and Esther Grabe
†Department of Linguistics, University of Cambridge, Great Britain
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

ABSTRACT

Research on Singapore English has concentrated on segmental rather than prosodic aspects although it is prosody that contributes most to its distinctive character. The few existing analyses of Singapore English intonation are based on the "British tradition" of intonation analysis. This paper investigates whether the British model is suitable for Singapore English.

INTRODUCTION

The nature of Singapore English has sparked off much interest and research in the past two decades. Kachru's [1] notion of nativisation accounts for its distinctive character. Nativisation is the process of acculturation of a language into a society which gives the language a distinct identity. As a result of this process, we find systematic differences between Singapore English (SE) and British English (BE). These differences involve the syntactic structure of SE, the lexicon, and especially its pronunciation.

A number of studies have investigated the pronunciation of SE [2,3,4]. Most authors focus on segmental rather than prosodic aspects. However, as Brown [5] points out, it is the prosodic aspects that contribute most to its distinctive character. The few existing analyses of Singapore English intonation [6,7,8] are based on the British model of intonation analysis [9]. This model has developed the concept of the 'tone unit' as a unit of intonational analysis. Tone units are stretches of utterance consisting of an obligatory element, the nucleus, and three optional elements, the prehead, the head and the tail. The definition of the tone unit relies on a set of underlying assumptions about the prosodic structure of English, in particular a distinction between unstressed, stressed and accented syllables. A stressed syllable is perceived as prominent in relation to other syllables in a given tone unit. In BE, this distinction is closely related to the one between full and reduced vowels [10]. Accented syllables are considered more prominent than stressed syllables and characterised by some degree of pitch movement. Accents may be prenuclear or nuclear, i.e. distinctions are made among accents.

A number of studies have applied the British model to Singapore English. This suggests that the authors assume that SE exhibits the prosodic parameters relevant to a successful application of this model, i.e. that we find unstressed, stressed and accented syllables. However, Brown [4] and Deterding [8] have cited the absence of reduced vowels and a lack of prominence contrasts as factors contributing to what has been termed the rhythmic 'staccato-effect' of SE. Our informal auditory analysis of SE confirmed that a clear distinction between stressed, unstressed and accented syllables cannot be established. Moreover, nuclear accents - crucial to the British model - could not be identified with any degree of certainty. At the functional level, we found that SE did not exhibit deaccenting. These observations shed doubt on the applicability of the British framework of intonation analysis to SE. In the following sections, we present experimental work attempting to provide an acoustic explanation for two of the observed aspects of SE prosody:
(i) the perception of 'staccato' rhythm
(ii) the apparent lack of the deaccenting function

STACCATO RHYTHM

Previous research [2,3,4,6] has explained the 'staccato' rhythm of SE by suggesting that SE is in fact syllable-timed, unlike BE which is frequently referred to as stress-timed. Syllable-timing is attributed to languages perceived to have near equal duration of syllables while stress-timing characterises languages perceived to have near equal intervals between prominent syllables. Here, we assume this distinction to be representative of a continuum between languages which are prepared to make durational adjustments for rhythmic purposes ("stress-timed") and those that do not. Adopting this view, it seems reasonable to look for an acoustic explanation for the perception of syllable-timing in SE.

Method

Yeow [11] measured syllable duration in SE and failed to find acoustic correlates for the perception of syllable-timing. Taylor [12] suggests that the acoustic correlate of SE's syllable-timed rhythm is primarily one of nearly equal vowel duration in syllables, not near-equal syllables. This view receives support from Brown's [4] and Deterding's [8] comments on the absence of reduced vowels. This lead us to test whether in the acoustic domain, a measure of vowel duration reflected the rhythmic structure of SE more accurately than one of syllable duration. We hypothesised that SE vowel were more nearly equal in duration than BE vowels.

Three British and three Singaporean subjects read a set of sentences. In order to test our hypothesis, a measure was needed to summarise the patterning of vowel durations in the two samples. We considered using the standard deviation of vowel durations, but, although a larger variation from syllable to syllable as expected in "stress-timing" would yield a high standard deviation of vowel duration, this would not unambiguously demonstrate the durational patterning SE vowels exhibit. It could, in principle arise if vowel duration became steadily longer as an utterance progressed. A measure which more securely reflected alternations of longer and shorter vowels would be the mean absolute difference between successive pairs of vowels in an utterance. This can be expressed as

$$\left[\sum_{k=1}^{m-1} | d_k - d_{k+1} | \right]/(m-1)$$

where m = the number of vowels in the utterance
d = the duration of the kth vowel

Informally, the difference in duration between each successive pairing of words in the utterance (d1 and d2, etc) is calculated, and the absolute values taken (by discarding the negative sign where it occurs). The mean difference is calculated by summing the differences, and dividing by the number of differences (i.e. one less than the number of vowels) and this is expressed in terms of an index.

Results

Table 1 and 2 show the vowel duration index values obtained for SE and BE speakers. A t-test showed that the difference between the overall index in SE and BE was highly significant. Clearly, durations of adjacent vowels in SE are more nearly equal in SE than in BE and we suggest that this lack of difference in successive vowel durations is largely responsible for the perception of syllable-timing in SE.

Table 1. Index for Singapore English

Speakers	*1*	*2*	*3*	*All*
Sentence 1	34	49	49	
Sentence 2	34	43	51	
Sentence 3	33	31	35	
Sentence 4	32	37	51	
Sentence 5	31	40	41	
Sentence 6	29	29	30	
Average	*32*	*38*	*43*	*38*

Table 2. Index for British English

Speakers	*1*	*2*	*3*	*All*
Sentence 1	40	74	59	
Sentence 2	74	71	70	
Sentence 3	63	60	51	
Sentence 4	60	56	50	
Sentence 5	56	48	43	
Sentence 6	35	30	29	
Average	*55*	*56*	*50*	*54*

DEACCENTING

Cruttenden [13] discusses deaccenting in the context of given and new information. In BE, given information is frequently deaccented. The most obvious type of given information involves verbatim repetition. Our materials included sentences with repeated lexical items. We hypothesised that the absence of pitch obtrusion in a lexical item representing given information would signal deaccenting. Hence, we predicted that in SE, repeated items would exhibit a step-up in fundamental frequency (f0) from preceding unstressed syllables whereas they would not do so in BE.

Method

Sentences (1) and (2) illustrate the results. (1) contains a lexical item repeated at the end of the sentence. (2)

acts as a control; it contains new information at the end of the sentence. In (1) *I went to the shop to buy sweets but they'd totally run out of sweets*, SE subjects accented the second mention of *sweets*, whereas BE subjects did not. In the control sentence (2) *I wonder why Chinese girls are better speakers than Chinese boys*, both SE and BE subjects accented the new information *boys*. In (1), peak f0 was measured on *out* which was perceived to be the last accented syllable in BE. The values were compared with peak f0 on the following two syllables. In the control sentence (2), peak f0 was measured on the last three syllables.

Results

Figures 1-4 illustrate peak f0 on the last three syllables of (1) and (2). Figure 1 shows that in BE, speakers 1 and 2 deaccented *sweets*, corresponding to a peak f0 lower than that of the preceding unstressed syllable. Speaker 3 accented *sweets* and this is reflected in the rising f0 from *out* to *sweets*.

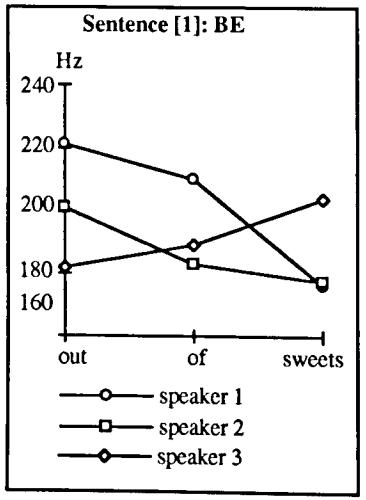


Figure 1. Given information - peak f0 in syllable

In Singapore English, the given information *sweets* was accented. Figure 2 shows that in this case *sweets* is characterised by an f0 value higher than that of the preceding syllable for all three speakers.

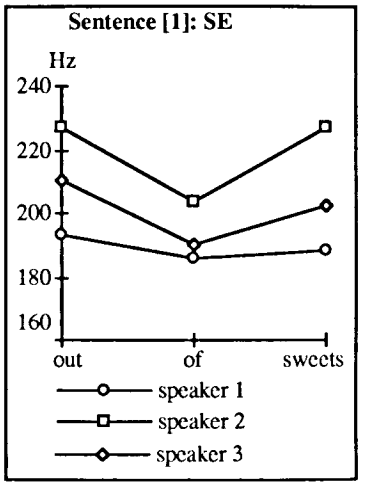


Figure 2. Given information - peak f0 in syllable

In the control (2), the new information *boys* was accented by all speakers. Again, peak f0 was measured on the last three syllables. Figures 3 and 4 show that in BE and SE the accent on *boys* corresponds to a step-up in f0 from the preceding syllable.

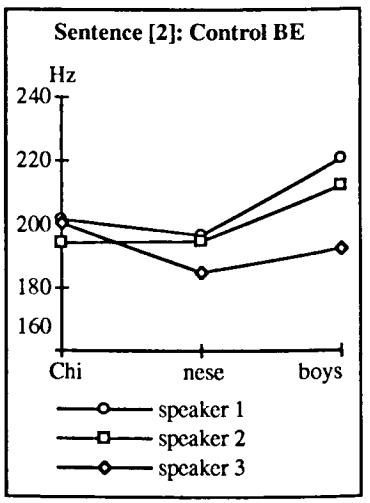


Figure 3. New information - peak F0 in syllable

We conclude that while BE assigns accent to new information and frequently deaccents given information, SE does not exercise this distinction.

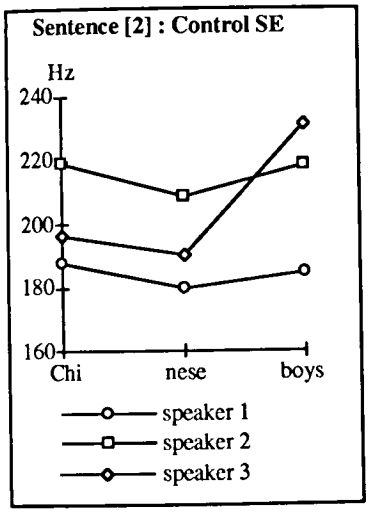


Figure 4. New information - peak f0 in syllable

CONCLUSION

The successful application of the British model of intonation analysis relies on the prosodic characteristics of British English. British English is a stress-accent language; i.e. we find a distinction between stress and accent, and accents are anchored to stressed syllables [14]. Our auditory analyis of SE suggested the absence of these distinctions. An acoustic investigation confirmed that at least two aspects of the prosodic system of SE differ crucially from that of BE, namely rhythm and the use of the deaccenting function. In Beckman and Edwards' [10] prosodic prominence hierarchy for English, the lowest level of distinction is characterised by the difference between full and reduced vowels, which in turn is closely related to that between stressed and unstressed syllables. Our results show that SE does not exhibit a comparable distinction, and suggest that the two varieties of English are not characterised by equivalent stress systems. This non-equivalence results in a lack of comparable anchor points for accents in SE, and contributes towards explaining the different distributions of accent placement. As a result, the overall rhythm of BE is perceived to be nearer "stress timing" and the one of SE produces nearer "syllable timing". Future research will focus on the notion of accent in SE and aims to establish how SE makes use of a phonological distinction in pitch.

REFERENCES:

[1] Kachru, B.B. 1986. *The Alchemy of English: The Spread, Functions and Models of Non-native Englishes*. Oxford: Pergammon Press.

[2] Tongue, R. 1979. *The English of Singapore and Malaysia*. Singapore: Eastern Universities Press.

[3] Platt, J. and H. Weber. 1980. *English in Singapore and Malaysia: Status, Features and Functions*. KL: OUP.

[4] Brown, A. 1988. The staccato effect in the pronunciation of English in Malaysia and Singapore. In J. Foley (ed) *New Englishes*. 1988. Singapore: Singapore University Press.

[5] Brown, A. 1991. *Pronunciation Models*. Singapore: Singapore University Press.

[6] Tay, M. 1982. The phonology of educated Singapore English. *English Worldwide* 3, 135-145.

[7] Rathi, D.A. 1983. Forms and functions of intonation in Singapore English: and auditory and instrumental study. Masters Dissertation: National University of Singapore.

[8] Deterding, D. 1994. A review of the characteristics of SE pronunication. *Review of Educational Research and Advances for Classroom Teachers* 1994.

[9] Crystal, D. 1969. *Prosodic systems and intonation in English*. Cambridge: CUP.

[10] Beckman, M.E. and Edwards, J. 1994. Articulatory evidence for differentiating stress categories. In P.A. Keating (ed) *Papers in Phonology III: Phonological structure and Phonetic form*. Cambridge: CUP.

[11] Yeow, K.L. 1987. Stress, rhythm and intonation in educated Singapore English. Masters Dissertation. National University of Singapore.

[12] Taylor, D.S. 1981. Non-native speakers and rhythm of English. *IRAL* 19:3.

[13] Cruttenden, A. 1986. *Intonation*. Cambridge: Cambridge University Press.

[14] Beckman, M.E. 1986 *Stress and non-stress accent*. Dordrecht: Foris.

PERCEPTION OF FOCUS IN STRESS ACCENT LANGUAGE (GERMAN) AND NON-STRESS ACCENT LANGUAGE (JAPANESE)

R. Hayashi and S. Kiritani
Research Institute of Logopedics and Phoniatrics,
Faculty of Medicine, University of Tokyo, Tokyo, Japan

ABSTRACT

Japanese is a language with lexical 'pitch accent,' and German with a 'stress accent.' The present study investigated the characteristics of production and perception of stress patterns in German SVO sentences by Japanese learners.

In the 'object-focused' utterance produced by German subjects, the pitch level of the object noun was found to be generally higher than that of the subject noun. Japanese subjects also emphasized the object, but the pitch level of the object was still lower than that of the subject.

However, in the perceptual experiment, Japanese gave higher rates of correct responses to the 'object-focused' utterance than native speakers.

These results suggest that Japanese learners of German produce the focus in German sentences with pitch cues which are sufficiently high for Japanese speakers but not for native speakers of German.

INTRODUCTION

The fundamental frequency (Fo) is an important acoustic correlate of word accent in many languages, especially in Japanese, which has lexical 'pitch accent.' There are several studies which claim that Fo contour plays the most important role in realizing word accent even in 'stress accent languages' including German [1]. It is also believed that each language has a language-specific Fo realization. In fact, when native speakers of Japanese learn 'stress accent languages' such as English and German, they frequently have trouble producing the focus in a sentence. Their utterance sound monotonous, and the focus in sentences is not produced with clear prominence.

To understand the differential role of pitch pattern in the realization of focus in German and Japanese, the production and perception of German utterances were compared between native and non-native speakers of German.

ANALYSIS OF PITCH PATTERN

Before conducting a perceptual experiment, a preliminary acoustic analysis of the pitch patterns produced by 6 German subjects and 11 Japanese subjects was performed.

Fig. 1 shows the utterances produced by a native male speaker of German. He produced 'neutral' utterances without any marked local pitch peak representing a word accent. When the focus was introduced, the highest pitch peak in each sentence coincided with the stressed syllable of the focused word. These characteristics were obtained in all series of utterances produced by native speakers.

In contrast, all the Japanese subjects produced neutral utterances with local pitch peaks which were similar to that of focused words produced by German subjects. The pitch peak for the object noun was generally lower than that for the subject. When the focus was put on the subject, the pitch peak for the subject word was the highest in the entire utterance. On the other hand, if the focus was put on the object, the pitch peak for the object noun was enhanced but was still lower than that for the subject. Thus, in these cases, the highest pitch peak in the utterance did not coincide with the position of focus. Fig. 1 shows an example of the utterances produced by a male Japanese subject as well.

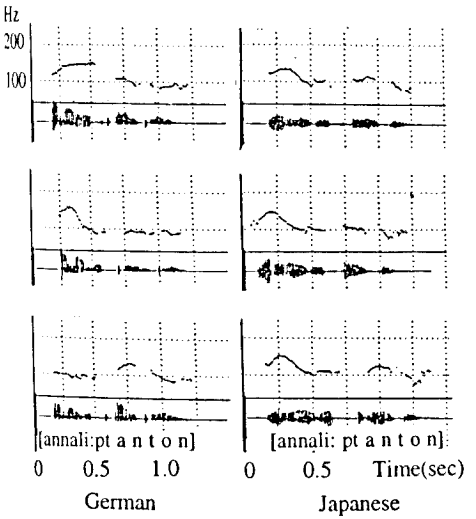


Fig. 1 Example of Fo contour for the utterances of "Anna liebt Anton." produced by a German male subject, and a Japanese male subject. Top row, 'the neutral utterance'; middle row, focus on the subject noun; bottom row, focus on the object noun.

PERCEPTUAL EXPERIMENT

A perceptual experiment on the identification of focused words was performed, with test stimuli with different pitch patterns constructed by editing natural speech.

Speech samples

The speech samples used were simple SVO sentences: "Anna liebt Anton." and "Anton liebt Anna." with seven different pitch patterns. Test stimuli were constructed by editing natural speech using a High-Speed Speech Analysis System on a personal computer [2].

Natural SVO utterances with differing positions of focus (the 'neutral' utterance, focus on the subject word, and focus on the object) were produced by a male and a female native speakers of 'standard German.' The utterances were divided between SV and O. These SV and O sequences from different utterances were combined to construct seven speech samples; Stimulus (S) 1 to stimulus 7 (S1 to S7), in which the relative levels of pitch in the SV and O sequences were varied. (Fig. 2)

Subjects

There were two groups of German subjects and two of Japanese: 7 native speakers of German and 30 Japanese university students who were learning German. Three of the native speakers were German teachers, and four were German university students living in Japan who did not major in linguistics. Eight of the Japanese subjects were advanced learners of German. They were graduate students, who had studied in Germany for more than one year, and therefore had many chances to speak German. The remaining twenty-two were undergraduate students who had learned German as a second foreign language for one or two years and considered as beginners in German.

Method

The subjects received two series of the stimulus sounds; one series from a male speaker, and another from a female speaker. Each series was composed of five trials. In each trail, 14 stimuli were presented in random order (seven types of stimuli: S1-S7 for each of the two sentence, "Anna liebt Anton." and "Anton

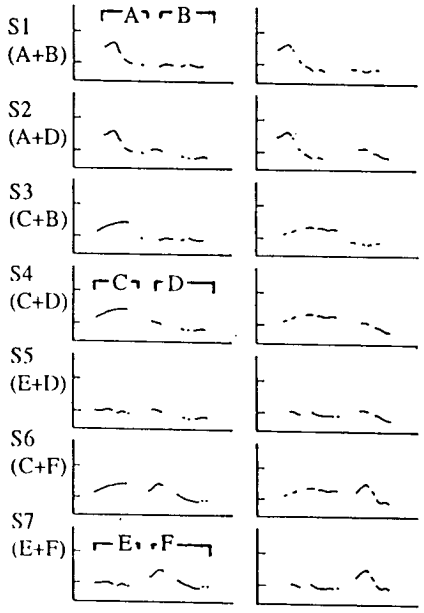


Fig. 2 Stimuli for the perceptual experiment. S1, S4, and S7 were the natural utterances. S2, S3, S5, and S6 were the cross-spliced speech samples as indicated above.

liebt Anna."). Thus, the subjects were presented with 70 stimuli in total for a series of stimuli produced by one native speaker. The subjects were instructed to mark on an answer sheet whether 'Anna' or 'Anton' sounded like the focus of the sentence. They were told to mark as quickly as possible after hearing a stimulus, even if they could not be certain of the position of focus.

RESULTS

Fig. 3 shows the result of the perceptual experiment. There was no significant difference in the responses regardless of the sentences used or the stimuli produced by a male or a female speaker.

As shown in Fig. 2, the relative pitch level of the object became higher in the order of the seven stimuli: S1 - S7. Fig. 3 confirms that the rate of 'object-focused' judgments tend to increase in this order both for the German and the Japanese subjects.

The German teachers responded almost perfectly to the natural utterances. The rate of correct response to S1 (the 'subject-focused' natural utterance) and S7 (the 'object-focused' natural utterance), was 100% and 95%, respectively.

Compared with the German teachers, German students showed lower rates of correct responses to S1 and S7, namely 81% and 70%, respectively. Naive subjects did not always perfectly judge the position of focus in the natural sentences, even though they were native speakers.

In the case of Japanese subjects, the rate of correct response to S1 by the beginners was 79%, which was similar to that of German students. Unexpectedly, the rate of 'object-focused' judgement to S7 was as high as 94%, which was higher than that of German students.

The advanced learners gave differential responses relative to the beginners. The rates of correct responses were 91% for S1, and 95% for S7, which were nearly the same as those obtained in the German teachers. Language learning might be responsible for the higher rates of correct responses by the advanced learners than those of beginners.

In addition, the rate of 'subject-focused' judgement by the advanced learners for S4, the neutral utterance, was 87%, and this was also higher than that of beginners. For S4, the rate of 'subject-focused' judgements by German teachers was 66%. The rate of 'subject-focused' judgements by advanced learners was higher than that by German subjects.

DISCUSSION

In the present study, Japanese subjects were found to produce un-focused words

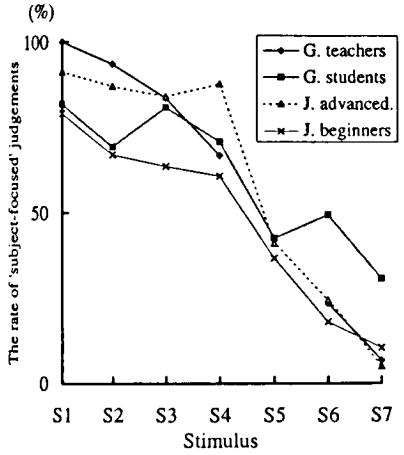


Fig. 3 Judgements of the position of focus. Each value shows the mean percentage of 'subject-focused' judgements by each subject group. (S5 was not included in the experiment for German teachers)

with local pitch peaks in such a way as German subjects pronounced focused words. At the same time, in the object-focused utterance, the highest pitch peak in a whole sentence did not coincide with the position of focus.

These characteristics were in agreement with those of Japanese prosody. In Japanese sentences, all content words show the pitch peak representing the word accent nuclei, and the pitch peak of the first content word is considerably higher than that of the second. When the focus is put on the second content word, enhancement of the local pitch peak and suppression of the pitch level in the rest part of the sentence are not so large as that in German. Consequently the pitch level of the second content word is not always the highest in the sentence.

It was also found that in the perceptual experiment, Japanese subjects had a tendency to show a higher rate of 'object-focused' judgements for S6 and S7 than the naive native speakers, in which the relative pitch level of the object word was higher than that of the subject. It appears that Japanese subjects tend to recognize the object word as 'marked', if the relative pitch level of the object word becomes a little higher, and they identify it correctly. This can be also explained by the characteristics of Japanese utterance with regard to small pitch change accompanying the focus.

Another point to be noted was that the rate of the 'subject-focused' judgment for S4 by advanced learners was much higher than that of German subjects. This phenomenon might be interpreted as an over-generalization response in the process of learning that the pitch level of the subject noun relative to the object in S1 was signaling the position of focus.

It seems that the inability of Japanese subjects to produce sufficient prominence for focus in German is not because they cannot perceive the German stress. Rather, these results suggest that Japanese learners of German produce focus in German sentences with pitch cues that are sufficiently high for Japanese speakers but not for native speakers of German.

REFERENCES

[1] Isacenko, A.V. & Schädlich, H.J. (1966), *Untersuchungen über die deutsche Satzintonation*, Studia Grammatica 7

[2] Imagawa, H. & Kiritani, S (1989), *High-speed Speech Analysis System Using a Personal Computer with DSP and its Applications to Pronunciation Training*, Ann. Bull. RILP. 23.

[3] Altmann, H. (1988b) *Intonationsforschungen*, Max Niemeyer, Tübingen

The Prosody of Mauritian Creole: Some Experimental Aspects

Philippe Martin
Experimental Phonetics Laboratory
Department of French, University of Toronto
Toronto, Ontario, Canada

ABSTRACT

A link does exist between the syntactic structure of a sentence and the sequence of prosodic contours located on the stressed syllables. More specifically, in French, patterns of melodic rises and falls located on stressed syllables do correlate with the syntactic hierarchy, independently of the syntactic categories involved. This preliminary study of Mauritian Creole prosody examines patterns of such prosodic contours in simple SN-V-SN configurations.

PURPOSE

Phonosyntactic theories of intonation link the syntactic structure of the sentence with specific prosodic contours located on the stressed syllables of the word. These contours encode a prosodic structure which enters into a complex relationship with syntax varying to homomorphy to total independence, depending on the style of the discourse (i.e. read sentences vs. spontaneous speech, with continuous variations between these extremes). In French, this approach leads to the discovery of a grammar of intonation, describing prosodic contours in terms of rising or falling fundamental frequency, syllable duration and intensity, which manifest abstract markers of the prosodic structure [1].

Creole languages appear to be of considerable interest to linguists as they demonstrate intriguing similarities on the syntactic level, even between varieties quite apart geographically and in time, such as Haitian and Mauritian Creoles. These similarities prompted a famous theoretical dispute, as to assign the generation of Creole languages ex nihilo by the existence of a bioprogram which would supply basic syntactic rules in the absence of any mother language [2], or (perhaps more convincingly) by applying universal rules that would result in similar word order in the absence of morphological markers [3].

From these two perspectives, the study of Creole intonation appear to be of some interest as 1) the absence of morphological markers indicates that the decoding of the syntactic structure can only be ensured by word order and intonation cues (letting aside semantic markers), and 2) properties of universal grammar can be perhaps found in the intonation grammar as well.

Thus, the quasi absence of morphology may give a more dominant configuration to the prosodic structure than in SF, and the presence of universal characteristics of syntactic encoding may indicate the presence of universal characteristics in the prosodic structure.

METHOD

Two speakers of Mauritian Creole (CL and ML) have been recorded reading about 50 sentences containing

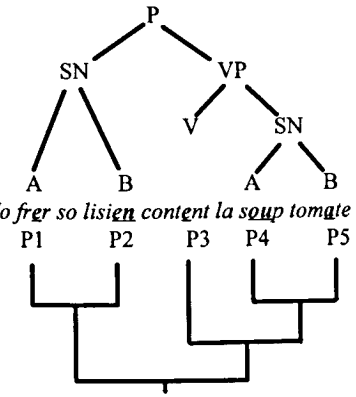
between 3 and 5 prosodic words (i.e. effectively stressed syllables), such as:

Mo frer so lisien content la soup tomate

(FS): Le chien de mon frère aime la soupe aux tomates
(My brother's dog likes tomatoes soup)

Each set of sentences was read 3 times in order to check consistency between prosodic realizations. Orthography has been somewhat modified here (from standard KM conventions [4]. Perceived stressed vowels are bold and underlined.

Most sentences were designed with the simple syntactic hierarchy



to be correlated with a 2 or 3 level prosodic structure organizing 5 prosodic words P1 P2 P3 P4 P5 (squared tree representation) through stressed syllable association. Most SN were of the Adj + N or (Det) N + N type in view that Mauritian Creole allows equivalent constructions such as

Cecil so frer content ser Asin.

Frer Cecil content ser Asin.

with no or very little change in meaning (FS: "le frère de Cécile aime la soeur d'Asin").
(Cecil's brother likes Asin's sister)

Acoustical analysis of the recordings were made with an real time fundamental frequency visualalizer (model PM1000) which allows easy readouts of Fo and duration values.

Since the informants were speaking Creole in their families, and that their language at work was English, it was assumed that interference between Standard and Creole French was minimal, although both informants could speak SF occasionally.

Experimental results

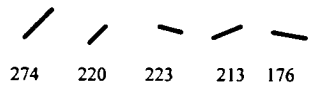
Experimental results in terms of melodic contours showed for both speakers striking similarities for prosodic patterns associated with the subject SN.

For example, comparative data for speaker CL are, for the 4 structures (the first 2 lines in each table correspond to the fundamental frequency values at the beginning and the end of each contour, the third line represents the contour duration in cs):

Cecil so frer content ser Asin (1)
(CL)

277	228	200	166	Hz
288	240	212	160	Hz
9	8	9	15	cs

Cecil so frer content Asin so ser (2)
(CL)

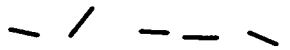


274	220	223	213	176	Hz
302	235	200	219	156	Hz
13	12	12	11	19	cs

Frer Cecil content ser Asin (3)
(CL)

204	250	200	166	Hz
200	266	212	161	Hz
11	10	9	15	cs

Frer Cecil content Asin so ser (4)
(CL)



224	264	223	213	176	Hz
214	278	200	219	156	Hz
10	9	12	11	19	cs

(Cecil's brother likes Asin's sister)
(Same meaning for all sentences)

Stress clashes in (1) and (3) explain the occurrence of 4 instead of 5 contours.

One example with a 2 levels expansion to the left is:

Cecil so tonton so lisien content la soup tomate
(5) (ML)

264	243	238	217	210	200
355	260	242	222	200	194
25	12	20	11	9	23

(Cecil's uncle dog likes tomato soup)

Two patterns clearly emerge from the data: a falling-rising on SN groups with expansion without a determinant, such as (1) and (2), and a rising-rising pattern, appearing on SN groups with expansion involving a determinant or a person's name as in (3) and (4). These regularities show up despite the rhythmical differences between the two informants, ML having a much slower speech rate and using more syllable duration than melodic variation for stress encoding. Examples are:

Cecil so frer content la soup tomate
(ML)

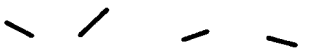
268	230	213	212	184	Hz
302	236	209	204	180	Hz
13	11	7	11	160	cs

Cecil so frer content la soup tomate
(CL)

268	228	221	204	201	Hz
295	240	204	201	268	Hz
13	12	20	11	28	cs

Without drawing any conclusions concerning the origin of these melodic regularities (bioprogram or universal), the two patterns can be related to similar melodic sequences found in SF: the fall-rise associated with subject SN (any grammatical category)

Le frère de Pierre a perdu son vélo (5)



C1 C2

(Peter's brother lost his bike)

and rise-rise with the theme-rheme construction (same meaning)

Pierre son frère a perdu son vélo (6)

C1 C2

By contrast with Creole, the subject SN in (6) appears dislocated, whereas in Creole both elements of the SN are equally integrated as in (2) and (3).

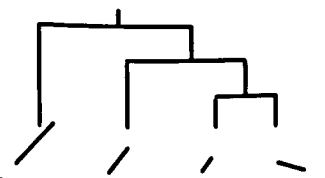
Both informant data displayed similar melodic amplitude contrasts (difference in frequency between the starting and ending points of the contour). The first Fo expansion was consistently higher than the second, although this was more marked for speaker ML than for CL.

The interpretation of these contrasts in term of prosodic structures could lead to the conclusion that the rise-rise sequence is a specific pattern associated with a (Det) N + N syntactic group, opposed to the fall-rising pattern correlated with N + N group

(Det) N + N N + N

In this case, differences in melodic variations could be attributed to the declination effect in the sentence.

Another interpretation would considered this difference between the two contours as indicating a 3 level prosodic structure



Pierre son frère a perdu son vélo (6)

similar to the one found for theme-rheme construction in SF, as in (6).

CONCLUSIONS

Simple read sentences of Mauritian Creole with various syntactic structures of the SN-V-SN type showed regular patterns of melodic contours somewhat different in their distribution from SF. In particular, sequences Det N + Det N were associated with 2 rising contours on the group stressed syllables, whereas examples such as Det N + Adj or N + N were associated with a falling rising pattern. The first pattern is similar to SF dislocated sentence prosody, the second resembles to the more common pattern found in 2 prosodic words subject SN.

REFERENCES

[1] Martin, Ph. (1987) "Prosodic and Rhythmic Structures in French", Linguistics, pp.925-949.

[2] D. Bickerton (1981) *Roots of Language*, Ann Arbor: Karoma Publishers Inc.

[3] R. Chaudenson (1977) in *Pidgin and Creole Linguistics*, ed. A. Valdman, Bloomington: Indiana University Press.

[4] Ph. Baker (1972) *KREOL A Description of Mauritian Creole*, London: C. Hurst & Co.

THE INTONATION OF QUERIES AND CHECKS ACROSS LANGUAGES: DATA FROM MAP TASK DIALOGUES

Martine Grice, Ralf Benzmüller, Michelina Savino, Bistra Andreeva*
Institute of Phonetics, University of the Saarland, Germany
**and CSTR, University of Edinburgh, UK*

ABSTRACT

This paper examines the phonetics and phonology of functional rises in Saarbrücken German, Bari Italian and Sofia Bulgarian . These rises are accounted for as (i) rises up to a boundary associated with a H boundary tone, and (ii) accent rises, with a L+H accent. In all three varieties, H boundary tones function as continuation rises, and L+H accents occur in information-seeking questions (queries) and some, but not all, types of confirmation-seeking question (checks).

INTRODUCTION

The predominance of final rising contours in questions (see e.g. Bolinger's survey [1]) has led to claims by Cruttenden [2] of a universal distinction between rises (signalling "open", meaning) and falls (signalling "closed") meaning. He classifies "open" as being in general non-assertive, examples of which are "listing" and "continuity". He also refers to the endoint of a phrase: in the absence of a rise/fall distinction, non-low and low endings serve to distinguish open from closed meanings.

Ohala [3] goes so far as to say that the correlation between high pitch and such "open" meanings is part of a frequency code for size used by animals in face-to-face encounters. His assumption is that the end point of the contour is of primary importance and it is this which is high. It is also reported that phenomena such as absence or presence of final lowering [4] or declination [5] may serve this function. However, what the universals literature does not deal with are local rises which occur before the phrase end.

Prefinal rises which serve an "open" function have been found in a number of languages, although they are rarely analysed as accentual rises. Exceptions are Transylvanian Hungarian , with a rising (LH) nuclear accent [6] in its question contour, and Palermo Italian which has accent rises for all "open" functions [5]. Yet it is not necessarily the case that a language makes use of only one type of rise. In fact, all three language varieties examined in this paper, Saarbrücken German (SG), Bari Italian (BI) and Sofia Bulgarian (SB), exhibit accentual rises in some "open" functions and high boundaries in others. An analysis of such functions requires speech data which is as natural as possible. Below we discuss how dialogue data were gathered, which "open" discourse functions were selected for analysis, and which types of functional rises are used to express them.

INTONATION IN DIALOGUE

One way of looking at "open" functions is to look at intonation within dialogues, as has been done for English in task-oriented dialogues, by Nakajima and Allen [8] using TRAINS world maps and Kowtko [9] using the Edinburgh map task corpus [10]. Nakajima and Allen parameterise f0 traces in terms of onset and final f0 and peak f0 ratio (a means of determining inter-utterance declination). They distinguish high from low endpoint, rather than rises or falls, thus allowing falls to mid to pattern with rises to mid. High phrase-final endpoints are shown to have one type of "open" function, signalling that the speaker intends to continue the speech act. However, this analysis concentrates on boundary and global f0, an approach not necessitating prior analysis of the intonation, but also not giving information about pitch movement within the utterance. It does not analyse non-final rises.

Kowtko, examining single word phrases in Scottish English, uses six intonational categories and an independent set of discourse categories, defined within the framework of conversational games [11].

The data analysed here follows this methodological lead as far as the dialogue recordings and the analysis of the discourse are concerned, but the analysis of the intonation is different. Rather than global shapes, contours are analysed as comprising mono- and bitonal pitch accents and two levels of boundary tone.

Map Tasks in SG, BI and SB

A modified version of the Edinburgh Map Task was carried out in each of the three language varieties. The task involved verbal co-operation (via auditory channel only) between two participants, each having a map, with the aim of transferring as accurately as possible a given route from one map to the other. There are a number of discrepancies in placement and positioning of landmarks on the maps. Since our aim was to examine intonation contours, the names of the landmarks contained mainly sonorants and were controlled for word stress pattern.

The pairs of subjects were of the same dialect background and knew each other well. They were unaware as to the purposes of the recording. For SG, 8 speakers were used, for BI 6, and for SB 8. Each speaker participated in two map tasks. The recordings were first transcribed orthographically. The orthographic transcriptions were analysed for the occurrence of a small set of discourse functions. Relevant tokens were then digitised and labelled intonationally using a system developed separately for each variety, based on the ToBI system for English [12].

Coding of Conversational Games

The initiating move in the following three game types was coded: 'query', 'check' and 'align'. These are defined as follows (cited from [11], page 4):

QUERY-YN: "Yes-no question asks for new or unknown detail about some part of the task; does not request clarification about instructions (that would be check); e.g. 'Do you have a rockfall?'"

CHECK: "checks self-understanding of a previous message or instructions by requesting confirmation directly or indirectly; makes sure that a complicated instruction is understood. e.g. 'So you want me to go down two inches'"

ALIGN: "Checks the other participant's understanding or accomplishment of a goal; elicits a positive response which closes a larger game; checks alignment of both participants' plans or position in task with respect to goal; checks attention, agreement, or readiness, e.g. 'Ok?' meaning 'Are you with me?'"

Although opening moves of INSTRUCT and EXPLAIN games are usually considered to be "closed", initial portions of them can be considered "open", especially when beginning a list of actions or items. These are also examined.

A preliminary attempt to classify utterances in terms of the kind of functional rise they contain (accentual or boundary-related), along with auditory analysis of the dialogues, led to the need for more differentiation within the category of checks. Three subcategories were proposed:

C0 - very low or no confidence that the information received is correct, therefore incredulous

C1 - medium level of confidence with no clear expectation as to the reply type

C2 - high confidence, not expecting a negation of the proposition but mostly expecting positive feedback. In some, but not all, cases, speakers continue to speak without waiting for such a signal.

Along Cruttenden's "open-closed" dimension, C2 could be said to be closed, C1 open and C0 extra-open. C2 checks, having a "closed" meaning, do not have functional rises. In fact, they all tend to have 'stepped down to' accents; SG has: H+!H*LL%, BI has !H+L*LL%, and SB has !H*LL%. The form of these contours will not be discussed further.

In all three varieties, queries and aligns have accent rises, as do C1-type checks. Continuation contexts in INSTRUCT and EXPLAIN moves have a final H boundary tone in all three varieties;SG additionally has an accent rise. In C0 checks, SG has a high boundary whereas the other varieties have an accent rise.

A closer analysis of the intonation patterns in each variety showed that the accent rises were not all timed in the same way.

TONAL ANALYSIS

A transcription system loosely based on the ToBI system developed for English [12] was used for each variety. In accentual rises, attention was paid to the alignment of the stressed syllable of the accented word with the F0 minimum and maximum corresponding to the L and H tones of a L+H (rising) accent. In addition, two boundary tones were used. For each language, the tonal analysis and morphological and syntactic markers used to distinguish each ot the moves

considered are given below.

Saarbruecken German

Queries (Q) and aligns (A) are generally distinguished from other moves by syntax (verb initial in interrogatives, verb second in declaratives) and intonation. The contour used is L*+H LL%.

1.Hast du ein blaues Wohnmobil? (Q)
Do you have a blue camper?

2.Bist du fertig? *Are you ready?* (A)

Within a dialogue context, the role of intonation is considerable. There are many cases where the verb is superficially phrase initial, through ellipsis of an initial discourse particle, such as 'dann' *then*. These are usually checks of type C2 where the speaker is assertive and confident, as in (ellipsis in parentheses):

3. (Dann...) müssen wir zurück.
(Then...) we have to go back.

Additionally, there are many elliptical queries and checks with no verb which also rely on intonation to distinguish them from other moves.

C1 checks have L+H*LL%, as in:

4. Den, dem linken? *The, the left one?*

C0 checks have a contour which is described in the literature on Standard German as 'Echo-frage' or Satz-Rückfragesatz [13]. It is L* HH%.

5. die Wiese?! *The meadow?!*

Non-final utterances have a combination of a L+H* pitch accent and either HL% or HH%, where an upstep rule as in [10] means that L% following H is high and H% following H is very high. The second contour type is often followed by a hesitation (possibly meaning that the H% is a floor-holding device).

6. ...an der Wiese dann vorbei...
then past the meadow...

Bari Italian

Italian has no morphological or syntactic markers for signalling interrogativity; intonation bears the functional load. Queries and aligns have L+H*LL% on the final potentially stressed syllable in the phrase. The final LL% tones are only fully realised when there is a postaccentual syllable to carry them, also as in PI. Queries (1,2) and an Align (3) are given below (X=focussed, X=accented):

1. Hai aGNELlo?*Do you have 'lamb'?*
2.Hai l'arca di no'E?
Do you have noah's ark?

When the focal accent is earlier in the phrase, a downstepped L+H* follows, as in Palermo Italian (PI) [5], although in PI the accent is L*+H instead of L+H*.

3.Non HAI un ristorante Anima Mia?
Do you not have a restaurant A. M.?

A slight final rise was found in a few emphatic queries, transcribed L+H*LH%. In checks of type C1, L+H* is also used, although deaccenting after the focal accent serves to distinguish these from queries.

4.Non DEVO andare verso la scritta?
I don't have to go towards the writing?

C0 incredulous checks are distinguished from C1 by means of expanded pitch range, as well as by voice quality. Non-final utterances are either L*LH% (5) or H*HH% (6). The latter is only in non-initial and penultimate phrases.

5. Io ho un secondo albero de MEle...
I have a second apple tree...
6. SCENdi...(e siamo arrivati)
Go down...(and we're there)

Canepari [14] claims that Bari questions are falling-rising, a pattern we have not found in our spontaneous data.

Sofia Bulgarian

Queries and aligns have L*+H LL%. A question particle 'li', which is placed after the focussed word, distinguishes them from other moves. Examples of a query, 1, and an align, 2, are as follows:

1. Kragla li e? *Is it round?*
2. Narisuva li go?*Have you drawn it?*

They constitute typical 'li'-question contours" as reported in [15].

C0 checks have L*+H LL% and do not have a question particle:

3. Pravo nagore otivam!?
I go up to the right!?

This is comparable with the description in [15] of emotionally coloured questions where there is a rise-fall, except when the accent is phrase-final, in which case there is a fall-rise. Our analysis of this is that the rising part is accentual and the final low is a boundary tone which, rather like BI, is not realised in the absence of a free syllable to carry it. In our examples, the initial rise begins low in the range, equivalent to level 1 or 2.

C1 checks have L+H*LL%, also with no question particle:

5. Tvoite malini sa moita mina?
Your raspberries is my mine?

Non-finality is expressed with L* HH%:

6. Marshruta ti tragva nagore...
Your route goes upwards...

Summary

Details of accent and boundary type and syntactic and morphological features of the moves across languages are given in table 1. LL% occurs in all cases not specified for boundary tones.

Table 1. Types of rise according to move

Move	SG	BI	SB
Query	L*+H Verb 1st	L+H* -	L*+H Q particle
Check C0	L*HH% -	L+H* ↑range	L*+H -
Check C1	L+H* -	L+H* deaccent	L+H* -
Align	L*+H Verb 1st	L+H* -	L*+H Q particle
Contun	L+H* HL% H%	L*LH% L*HH%	L*HH%

IMPLICATIONS

High boundaries in the majority of the world's languages may serve to determine the illocutionary force of the whole phrase, making it an information- or confirmation-seeking questions, the former corresponding to QUERIES and ALIGNS, the latter to CHECKS. It is therefore sensible for the boundary tone (H%) to be a property of the phrase, as is the case in [4]. The accentual rises shown here serve the same function. This is evidence for assigning pitch accent type at the level of the phrase too.

ACKNOWLEDGEMENTS

We are grateful for comments from J Kowtko, M Johnson and E Grigorova.

REFERENCES

[1] Bolinger, D (1978). Intonation across languages, in Greenberg, Ferguson and Moravcsik (eds.) *Universals of Human Language, vol 2, Phonology*, SUP.
[2] Cruttenden A (1981) Falls and rises: meanings and universals *J. Ling.* 17.
[3] Ohala JJ (1983) Cross language use of pitch: an ethological view *Phonetica*, 40.
[4] Pierrehumbert JB and Beckman ME (1988) *Japanese Tone Structure*, MIT.
[5] Thorsen, N. (1980) Intonation contours and stress group patterns of varying length in ASC Danish, ARIPUC 14.
[6] Ladd DR (1983), Phonological features of intonational peaks, *Language*, 69.
[7] Grice M (1992) The intonation of interrogation in Palermo Italian; implications for intonation theory, PhD, UC London, also (1995) Niemeyer, (L.A. series)
[8] Nakajima S and J Allen (1993) A Study on Prosody and Discourse Structure in Cooperative Dialogues *Phonetica* 50.
[9] Kowtko J (1992) On the function of intonation in discourse, Proc. I of A, 14 6.
[10] Anderson AH et al, (1991)The HCRC Map Task Corpus, *Lang. and Speech*, 34.
[11] Kowtko J, S Isard and G Docherty-Sneddon, Conversational games within dialogue, HCRC RP-31.
[12] Beckman ME and G Ayers, 1994, Guidelines for ToBI Labeling, Dept. Linguistics, Ohio State University.
[13] Altmann H (1984) Linguistische Aspekte der Intonation am Beispiel Satzmodus, Foschungsberichte 19 20, Inst f Phonetik u. spr. Komm der Uni München.
[14] Canepari L (1992) *Manuale di Pronuncia Italiana*, Zanichelli, Bologna.
[15] Tilkov, Dimitar, 1981, Intonaciata v balgarskija ezik, B.A.N., Sofia.

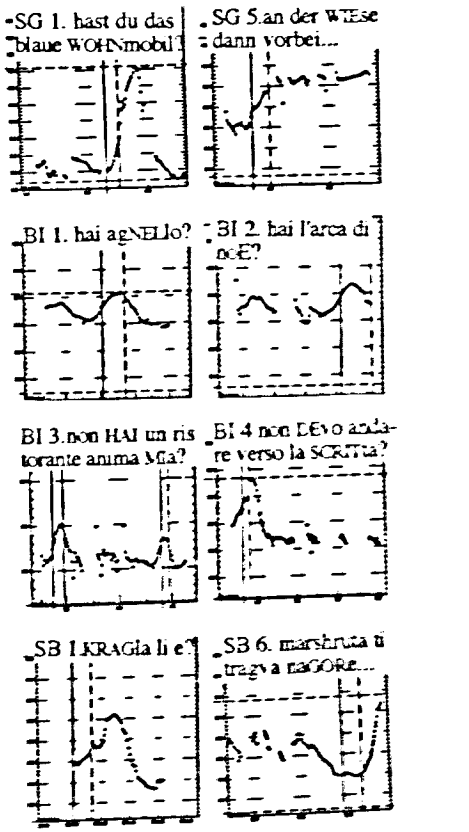


Figure 1: F0 contours, vowel of accented syllable is between vertical cursors.

COMPARING F0 RESET IN TWO PROSODICALLY DIVERSE LANGUAGES
—data from Eskimo and Yoruba—

Y. Nagano-Madsen and A.- C.Bredvad-Jensen***
**Department of Oriental Languages, University of Gothenburg, Sweden.*
***Dept. of Linguistics and Phonetics, Lund University, Sweden*

ABSTRACT

Analysis of H tone reset from text reading was compared for Eskimo and Yoruba. Our preliminary results indicated that the main difference between the two languages were the regularity vs. irregularity in pitch register. In Yoruba, there was a great regularity in the F0 values of the first H tone in the sentence, with reference to sentence length, and in the F0 values of the upcoming reset Hs. Neither of these regularities were present for the Eskimo speakers.

INTRODUCTION

The present paper is a progress report on an on-going project called MULTILINGUAL PROSODIC RULES, —with specific reference to Eskimo, Japanese, and Yoruba—in which three prosodically diverse languages, all non-European, non-stress languages, are examined for their durational and F0 (fundamental frequency) features. A specific hypothesis for the current topic is that, when F0 or duration is used for signalling lexical properties, there must be a limit in the use of the same acoustic property for other purposes such as phrasing, i.e. we expect that a lexical tone language like Yoruba cannot use F0 with the same degree of freedom as Eskimo can, while Eskimo, in which duration is used extensively for contrastive purposes, cannot use duration with the same degree of freedom as Yoruba.

Research on intonational phrasing has been receiving increasing attention in recent phonetic literature, in particular for discourse analysis. Segmenting texts on the basis of F0 organization, with or without additional cues, appears to be a powerful method in many languages. However, most of the works on prosodic segmentation are based on stress languages like English and Swedish. How F0 and duration are utilized to signal phrasing in languages like Yoruba and Eskimo is still a very open question, and is of typological interest.

Tonal and intonational features in Yoruba and Eskimo have been studied at word, phrase, and sentence levels [1 - 5].Some of this research has indicated a number of features which can be related to intonational phrasing in these languages. In the present study we report the results of a pilot study which compared H tone F0 resets in Eskimo and Yoruba using text reading material. To our knowledge, this is the first study of intonation in these languages which analyses a large corpus of text reading.

MATERIAL AND ANALYSIS

Eskimo

The original material was adopted from a textbook of West Greenlandic Eskimo, which consists of mostly declarative sentences of relatively simple syntactic and semantic structure (recording time approx. 3 minutes for each subject). The text included 148 words, 30 sentences, and was divided into 6 paragraphs in written form. Each sentence consisted of between two and seven words.The number of sentences in each paragraph varied from three to nine. Two female speakers of Central West Greenlandic read the text.

Prosodic transcriptions were made for the entire material, marking word property H and L tones as well as short and long pauses. The recorded utterances were digitized at 20kHz and F0 was analysed using the pitch analysis command of the CSL software package installed on a PC. Figure 1 presents a sample F0 contour, showing how F0 values corresponding to H and L tones are obtained.

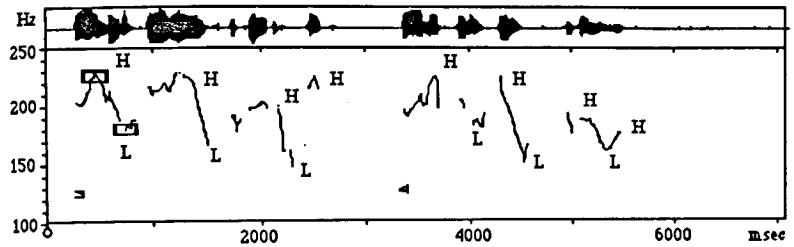


Figure 1. Sample F0 contour and speech waveform for Eskimo sentence "Kaali aqqanilinnik ukioqarpoq, Kaalallu arfineq sisamanik" (Kaali is eleven years old, and Kaala is nine years old-speaker 1), showing how F0 values were obtained for target tones.

Yoruba

The text used for the Eskimo speakers was translated into Yoruba by selecting 18 out of original 30 sentences. The text was read by a male speaker of Yoruba from Lagos in a sound proof studio. The analysis procedure is the same for that of Eskimo. All H, M, and L tones were measured by choosing, in principle, the highest and lowest F0 point for the corresponding syllable, and the F0 value in the middle of the vowel for M tone.

Marking F0 resets

The F0 value of each H tone was examined successively and when it was higher by more than 5Hz relative to the previous H tone, it was marked as a reset H tone.

RESULTS

Number and location of H tone resets

The number of H tone resets in a sentence varied from two to five in both languages, of which two and three resets were most common. In both languages, the number of resets appeared to be related to sentence length - the longer the sentence, the more resets it contains.

As for the location of resets, there was considerable disagreements between the two speakers of Eskimo even though the total number of H resets for the text was extremely similar, i.e. 69 for the first speaker and 68 for the second speaker. The main difference between the two speakers arose from the fact that they had different phrasing strategies in marking syntactic constituents.

While the first speaker tended to mark the beginning of the next constituent by resetting, the second speaker often marked the end of the constituent by resetting. The Yoruba speaker preferred to mark the beginning of the syntactic constituent.

Target F0 values for reset H

F0 values for the reset first H tone as well as the following reset H tones are shown graphically with reference to the number of resets (plus the first H) in a sentence. Figures 2(a)(b)(c) show the results for Yoruba for one, two, and three resets while those for Eskimo are shown in Figures 3 (a)(b)(c).

The main difference between the two speakers of Eskimo and the Yoruba speaker was that of pitch register. For the Yoruba speaker, the F0 value of the first H was fairly steady, clustering around 140-150Hz for majority of sentences, and around 165Hz for longer sentences. Furthermore, subsequent reset H tones had either similar, or lower, but never higher, F0 values. These two characteristics were not observed for the two speakers of Eskimo.

There was a clear indication that sentence length plays a role in determining the F0 target of the first H in Yoruba. The length of sentences in the material varied between 21 morae (sequence of two vowels are counted as two morae) to 51 morae. The first H tone in a longer sentence always had the highest F0 value, i.e. around 165Hz. In Eskimo, on the other hand, no such regularity was observed.

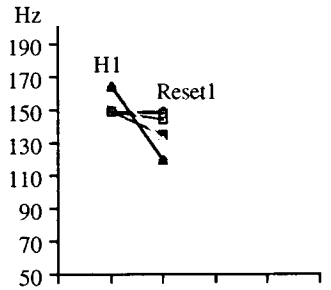


Figure 2 (a). F0 values for the first H and the following reset H in Yoruba.

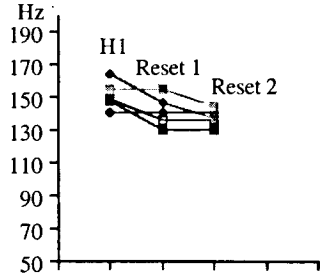


Figure 2 (b). F0 values for the first H and the following two H resets in Yoruba.

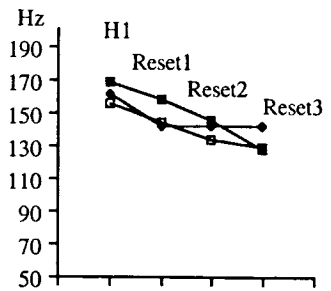


Figure 2 (c). F0 values for the first H and the following three H resets in Yoruba.

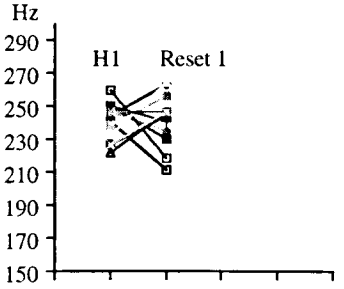


Figure 3 (a). F0 values for the first H and the following H reset in Eskimo (speaker 1).

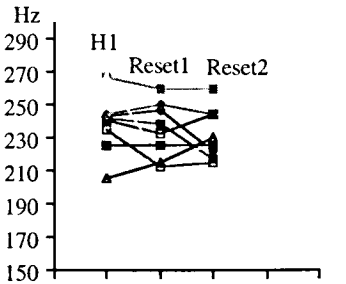


Figure 3 (b). F0 values for the first H and the following two H resets in Eskimo (speaker 1).

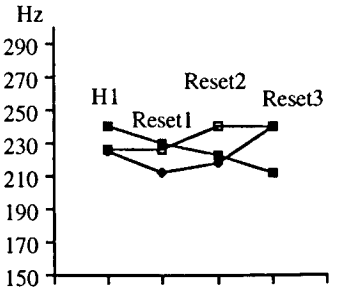


Figure 3 (c). F0 values for the first H and the following three H resets in Eskimo (speaker 1).

SUMMARY AND DISCUSSION

Our pilot analysis of H tone F0 resets in Eskimo and Yoruba revealed both similarities and differences. The fact that the number of F0 resets is related to the length of sentence may imply that speakers in general try to split a stream of speech into segments of reasonable length for comprehension by listeners.

It is interesting to note that the two speakers of Eskimo had different strategies in marking the syntactic boundary by F0 reset, one marking the end of the syntactic constituent, the other marking the beginning of the next constituent. Such a difference has also been reported for Yoruba by Laniran for her two speakers reading identical sentences [4]. Our present speaker of Yoruba preferred to choose the beginning of constituents. Even though this strategy is found to be present in both languages, there seems to be a critical difference. In Eskimo, when the last word of a syntactic constituent is marked by a new reset, the F0 value is often as high, or even higher than that of the previous H. In addition, the total pitch range of the word in question is enlarged by giving an impression of prominence or emphasis. In Laniran's study of Yoruba, the F0 value of such a H reset tone never exceeded the F0 value of sentence initial H.

The most notable difference between Eskimo and Yoruba was that of pitch register. In Yoruba, there was great regularity in the F0 values of the first H tone in the sentence with reference to sentence length, and the F0 values of the upcoming reset Hs. There seems to be a speaker-specific preferred pitch register, i.e. for a long sentence, the first H usually had F0 value of 165Hz and for sentences of moderate length, around 140-50Hz. If the first H had the value of 165Hz or so, the next H reset was likely to be 140-150Hz, and then the value decreases to 130Hz and finally to around 120Hz, being the lowest possible value for H reset for this speaker. The F0 value of the subsequent resets, however, did not decrease step by step to the lowest level. In some sentences, two or three subsequent resets had the same value. There were, however, never more than three sequential resets of the same level. In a few cases, the first H had the highest value (165Hz) and was followed by only one reset of the lowest pitch register (120Hz). How these different levels of pitch register are determined for reset H tones, is not immediately obvious. Possible candidates are (1) tonal structure, (2) syntactic and semantic structure, and (3) the speaker's choice.

Since the results of the present investigation are based on a small number of speakers, we are now continuing to examine some of the points found in this study for more speakers in both Eskimo and Yoruba.

ACKNOWLEDGEMENT

The present research is funded by a grant from Swedish Research Council in Humanities and Social Sciences. The authors are grateful to Gösta Bruce, the project leader, for many useful comments.

REFERENCE

[1] Mase, H. (1973), "A study of the role of syllable and mora for the tonal manifestation in West Greenlandic", *Annual Report of the Institute of Phonetics,* 7, pp.1-98, University of Copenhagen.

[2] La Velle, C.R. (1974), "An experimental study of Yoruba tone", *Studies in African Linguistics.* Supplement 5, pp.185-194.

[3] Connell, B. & Ladd, D.R. (1990), "Aspects of pitch realisation in Yoruba", *Phonology* 7, pp.1-29.

[4] Laniran, Y. (1992), *Intonation in tone languages: the phonetic implementation of tones in Yoruba*, PhD thesis, Cornell University.

[5] Nagano-Madsen, Y. (1993), "Phrase-final intonation in West Greenlandic Eskimo", *Working Papers* 40, pp.145-155, Dept. of Linguistics, Lund University.

PITCH ACCENT PATTERNS IN ADJACENT-STRESS VS. ALTERNATING-STRESS WORDS IN AMERICAN ENGLISH

Stefanie Shattuck-Hufnagel,
Speech Communication Group, MIT
77 Massachusetts Avenue, Cambridge, MA 02139 USA

ABSTRACT

In American English, adjacent-stress words exhibit less regularity than alternating-stress words for both main-stress and phrasal prominence placement. Apparently, words with adjacent stress pose a challenge to the prosodic system, possibly because there is a preference for rhythmic alternation in human motor behavior.

INTRODUCTION

Many words of English exhibit alternating stress, in the sense that full vowel syllables occur next to reduced syllables, as in *open* (Full-Reduced), *information* (F-R-F-R), *Appalachicola* (F-R-F-R-F-R), *about* (R-F), *granola* (R-F-R), *communication* (R-F-R-F-R), etc. In some words, however, two full vowels occur in immediate succession, as in *bisect, Arlene, moron, location,* etc. In both types of words, one of the full-vowel syllables is understood to be the main stress syllable, while full vowels that precede the main-stress vowel are often understood to have other degrees of lexical stress. Following Bolinger [3], we contrast full vowels (which carry some degree of lexical stress) with reduced vowels (which do not); in this framework, we refer to words of the first type as **alternating stress**, and of the second type as **adjacent stress**.

In a study of the placement of pitch accents within intonational phrases, we were puzzled to note that adjacent-stress words did not show the same regularity in pitch accent placement that most alternating-stress words did. In the present paper, we report a more systematic study of this difference in pitch-accent behavior, and relate it to additional differences in lexical stress behavior for words with these two classes of rhythmic or lexical stress patterns.

ANALYSIS 1: LEXICAL STRESS

In earlier work with Ostendorf and Ross [5], we looked at the distribution of pitch accent location within words in a corpus of speech produced by two FM radio newscasters, and labelled for prosodic constituent structure and prominence. We found that a pitch accent did not always occur on the main stress syllable of its word, especially for words that contain a full vowel syllable preceding the main stress syllable, such as *contradict* (F-R-F), *Massachusetts* (F-R-F-R) and *environmental* (?-F-R-F-R)) In such words, a pitch accent often occurred on the early full vowel syllable, in patterns that were constrained by both intonation phrase structure and pitch accent rhythm. Details of database construction, prosodic labelling methods, analysis techniques and results are reported in [5], and a summary appears in Dilley et al. (this Proceedings.)

Although these effects were significant for the more than 400 target words in the corpus, there were a number of cases that did not conform to these general rules. Examining the exceptions, we noted that many of them were target words with two full-vowel syllables in adjacent positions. More careful analysis showed that this phrase-level effect was quite systematic, and that there were parallel differences at the level of lexical stress.

Lexical stress. Candidate words for Early Accent Placement must include a syllable to the left of the main stress syllable to serve as a docking site for a possible early pitch accent [1]. We initially developed a criterion for EAP candidate words that specified the early syllable as a full-vowel, unreduced syllable. In scanning the list of potential EAP candidates in the FM radio news corpus, however, we noticed that for words with alternating stress, such as *institution* and *Mississippi* , it was quite easy to judge whether an earlier syllable was pitch-accentable, and there was good agreement among judges. But for adjacent-stress words like *illegal, trustee* and *statewide,* this judgment was more difficult. (This set of words could be defined as having a monosyllabic foot, which places the head syllable of two feet directly adjacent.) Some consisted of two root morphemes (*sometimes, southwest, shortchange*), some of a prefix plus a root (*predates, rehash* (in the verb form), and in some the two strong syllables appeared in a single root morpheme or sequence of root morpheme plus suffix (*primarily, minority, foundation*).

To resolve this problem for the Early Accent study, we used the secondary stress markings in Webster's 9th New Collegiate Dictionary (1984) to determine whether or not a candidate word had a pre-main-stress secondary stress syllable that could serve as a docking site for a Pitch Accent. However, when we later compared non-main-stress markings in different dictionaries, we again found a contrast: alternating full-vowel words were marked consistently across dictionaries, and their markings almost always corresponded to our common intuitions, but adjacent full vowel words were often marked differently by different dictionaries, or marked with a number of alternate stress patterns, or marked with stress patterns that did not correspond to our intuitions. For example, comparing the stress patterns given by Kenyon and Knott's Pronouncing Dictionary of American English (1944, reprinted in 1953) with those given by Webster's Ninth, for several words in our corpus that were originally candidates for early accent placement, provides the following contrasts ("-" = no stress is marked for this syllable in Webster's, 1 = main stress is marked for this syllable, and 2 = secondary stress is marked for this syllable):

	K&K	Webster
nineteen	-1, 11, 12	(1)1
southwest	21, 12	-1
rehash	-1	(1)1
primarily	12--, -1--	-1--
sometimes	12, -1	12, (2)1

In general, words with adjacent full vowels were less consistently marked for stress than words with alternating full vowels, in two ways: a) they often had more alternative stress patterns listed, and b) these alternative patterns differed not only in whether the early syllable had secondary stress vs. no stress, but also in the location of main stress.

An additional indication that adjacent full-vowel syllables are treated more irregularly by the stress system comes from analysis of words that begin with the prefix *dis-*. When the following syllable is reduced, as in *disagree*, the prefix is regularly marked for secondary stress (e.g. *disability, disadvantage, disaffect, disassociate*, in Webster's Ninth.), But when the following syllable is marked for stress, *dis-* is treated quite variably: in the first ten words of this type listed, 4 have *dis-* marked with no stress (*disable, disband, disbar, disburse*), 4 with possible main stress (*disaggregate, disarm, disbud, disburden*), and three with possible secondary stress (*disadvantageous, disapprobation*).

These observations suggested that, even though we excluded from our analysis of Early Pitch Accent words that lacked a secondary stress marker on a syllable preceding the main stress syllable in the dictionary, these markings might be less reliable for potential adjacent-stress than for alternating-stress words. Thus, it might be of interest to analyse the pitch accent placement data separately for alternating full vowel vs. adjacent full vowel words. Results will be reported here for Early Accent candidate words that were labelled with two pitch accents.

ANALYSIS 2: PITCH ACCENT

Double accented words were not uncommon in this corpus, possibly

because newscasters place pitch accents on a higher proportion of accentable syllables than do nonprofessional speakers. Moreover, double accenting occurs more often for alternating stress candidates (26% double accented, 77/295) than for adjacent stress candidates (8% double accented, 11/132). This difference suggests that speakers are not loath to place accents on two stressed syllables of the same word as long as the two syllables are separated by another syllable, but tend to avoid placing accents on adjacent stressed syllables within the word. If confirmed by further analysis and for additional speakers, this observation provides support for the claim that speakers avoid direct pitch accent clash within the word, just as does our earlier findings suggest avoidance of pitch accent clash across word boundaries (if the two words occur within an intermediate intonational phrase).

We took this analysis by word stress pattern one step further, looking at the set of early prominence candidate words that happened to contain all the accents in their intonational phrase. There were three possible pitch accent patterns for these words: a) pitch accent only on an early full-vowel syllable, b) pitch accent only on the main-stress syllable, or c) a double accent, one on each of these two syllables. Our prediction was that these words would contain a higher proportion of double accents than other candidate words, because they were subject to two influences: the requirement that the nuclear accent of the phrase be placed on the main stress syllable of its word, and the tendency to place the first accent in the phrase as early as possible. (By definition, this subset of EAP target words are never deaccented.) On this view, the first factor encourages placement of an accent on the main-stress syllable, and the second factor encourages one on the earlier secondary-stressed syllable.

Overall, the set of candidate words which contained all the accents in a phrase did have a greater likelihood of double accent: 45% (45/101) of these tokens were double accented, while only 14% of word tokens whose phrases also contained other accents were double accented (44/304). We then separated out the results for alternating stress vs. adjacent stress; results are shown in Table 1.

Table 1. Proportion of EAP target words, containing all accents of their Intermediate Intonational Phrase, that were Early Accented, Main-stress Accented or Double Accented, for a) alternating stress words, and b) adjacent stress words.

	Early	Double	Main	Total
a)	2	41	33	76
b)	7	4	14	25

While 54% of alternating stress target words in this set were double accented (41/76), only 16% of adjacent stress words were (4/25). That is, although speakers tend to double accent EAP candidate words that contain all the accents of the intermediate intonational phrase if they are alternating-stress words, this effect disappears when the main-stress syllable and the earlier full vowel syllable are adjacent, i.e. not separated by a reduced vowel syllable.

DISCUSSION

Although a number of prosodic theories permit double accented words, especially when the word contains the first accent of a phrase (e.g. Gussenhoven [4], Beckman and Edwards [1], Bolinger [3]) we did not expect to find such a high proportion of double accents. It is possible that the FM radio news speaking style has a higher proportion of double accents than other styles, but the fact that they occur freely in this corpus and sound perfectly natural shows that they are fully acceptable according to the grammar of English prosody. Double accents in certain contexts follow naturally from models in which pitch accents are acceptable on the pre-main-stress full vowels of a word, as well as on the main-stress vowel.

The finding that adjacent full vowels or stressed syllables are treated more irregularly by the prominence assigning component of the prosody emerges from both the anecdotal evidence about lexical stress marking in dictionaries, and from the empirical results for Early Accent Placement. An additional small piece of evidence that adjacent stress words are associated with greater irregularity in prominence placement comes from analysis of the target EAP words which did not follow the general rule for locating phrase-final (i.e. nuclear) accents on the main-stress syllable of their word. This general tendency was very strong in our corpus, providing evidence that the prosody labellers were consistently finding the boundaries of intermediate intonational phrases [2], for which the nuclear accent is proposed to fall on the main-stress syllable of its word [2], [4] and others). However, there were five exceptions to this general rule, i.e. five EAP candidate word tokens which contained the nuclear accent of their phrase, but were accented on their early secondary stress syllable only, rather than on their main stress syllable or on both. In contrast, for none of the 90 alternating stress words with a nuclear accent was it placed on the early syllable only. Although the number of exceptions is small, that fact that they occurred only for adjacent-stress words again suggests that this stress pattern challenges the prominence placement rules of English prosody.

CONCLUSIONS

This analysis of Early Pitch Accent in alternating-stress vs. adjacent-stress words in a corpus of continuous communicative speech suggests three points. First, in studying phenomena related to early prominence in the word, either in analysis of the behavior of candidate words in speech databases or in selection of stimulus words for experimental speech elicitation, this contrast in rhythmic stress pattern should be considered. It is possible that alternating stress words will provide a clearer measure of the effects of the various factors that influence Early Accent. In fact, there is much we do not understand about the factors that constrain both secondary lexical stress and Early Accent Placement. For example, some full vowels that are not marked with secondary stress appear to resist early accent (e.g. *Montana*), whereas others accept it freely (e.g. *illegal*, which was often produced with early accent in our FM radio news corpus). The categorization of a pre-main-stress full vowels in words of English as having some degree of lexical stress or not, and as capable of carrying a pitch accent or not, requires further work, at least for words with adjacent full vowels. Second, since it cannot be assumed that a pitch accented word will be accented on its main-stress syllable, studies of the acoustic correlates of different pitch accent types will require labelling of pitch accent locations by syllable rather than by word, at least for Early Accent candidate words that include a full vowel syllable preceding the main-stress syllable. Finally, the observation that speakers prefer not to pitch accent adjacent syllables, either within words or across word boundaries, is consistent with the claim that human speakers prefer to alternate more prominent with less prominent elements in an utterance if they can.

REFERENCES

[1] Beckman, M. and Edwards, J. (1994) Articulatory evidence for differentiating stress categories, in *Phonological Structure and Phonetic Form: Papers in Laboratory Phonology III*, ed. P. Keating, Cambridge University Press

[2] Beckman, M. and Pierrehumbert, J. (1986) Intonational structure in Japanese and English, *Phonology Yearbook* 3, 255-309.

[3] Bolinger, D. (1981) Two kinds of vowels, two kinds of rhythm, distributed by the Indiana University Linguistics Club, Bloomington, Indiana.

[4] Gussenhoven, C. (1991) The English rhythm rule as an accent deletion rule, *Phonology* 8, 1-35

[5] Shattuck-Hufnagel, S., Ostendorf, M. and Ross, K. (1995), Stress shift and early pitch accent placement, *J. Phonetics* 22, 357-388.

NUCLEAR ACCENT TYPES AND PROMINENCE: SOME PSYCHOLINGUISTIC EXPERIMENTS

Gayle M. Ayers
Department of Linguistics, The Ohio State University

ABSTRACT

Two locations of nuclear accent (early and late) and three kinds of nuclear accent in English were considered. In reaction time measures, nuclear accents were faster than non-nuclear accents. However, downstepped nuclear accents were slower than regular and emphatic nuclear accents, suggesting that downstepped accents are less prominent, and that nuclear accent is not a fully uniform category.

INTRODUCTION

This study examines phonetic prominence of nuclear accent types in English using two experimental tasks: cross-modal naming and phoneme monitoring. These tasks provide a way to observe the influence of sentence intonation on the behavior of listeners, from which we can infer the status of the category nuclear accent and the relationship between accent type and prominence values. In addition, they help inform us of the role of intonation in lexical access and sentence processing.

The test materials are sentences produced as single intonational phrases with early ("A BOAT was near the tower") or late nuclear accent ("A boat was near the TOWER"), and with one of three phonologically distinct nuclear accent types. The question of interest is whether these three types of nuclear accent are all equally prominent (the traditional analysis of nuclear accent as a single qualitative level of stress which is independent of accent type) or whether there are differences between the nuclear accent types (e.g., that downstepped nuclear accents are less prominent [1]).

The three accent types can be characterized by the relationship between the pitch levels on the nuclear accent and the preceding accents. See Tables 1 & 2 for sample materials; the intonation patterns are transcribed using high and low tones for accents and phrase boundaries [2]. Figure 1 shows the mean F0 values (in Hz) of early and late position words in four intonation contour types (sentences from Exp. 4). Measurements were taken at the midpoint of the stressed vowel. Filled circles represent nuclear accents. A regular nuclear accent (R) has a pitch level similar to that of the preceding accent (although it may be slightly lower due to final lowering [3]). An emphatic nuclear accent (M) has a dramatic pitch rise on the nuclear accented word. A downstepped nuclear accent (D) is significantly lower than the preceding accent.

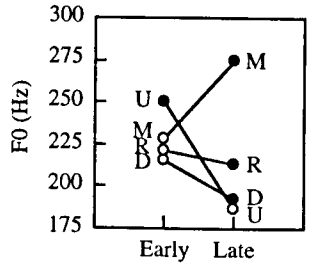


Figure 1. Mean F0 values of emphatic (M), regular (R), downstepped (D), and unaccented (U) contour types.

CROSS-MODAL NAMING

The cross-modal naming task measures the speed of lexical access. It shows effects of lexical priming and sentence position (RT is slower early in the sentence) [4], [5], but effects of intonation have not been systematically explored previously.

Method

Subjects. 84 undergraduate students participated in the two experiments, 42 subjects in each experiment.

Stimuli. 96 critical sentences were used, each containing one prime word. The prime word was either the head noun of the subject (early position) or of the object (late position). Exp. 1 used semantic associate priming, and Exp. 2 used identity priming. Table 1 shows example sentences and targets and the two intonation contour types used: (1) late (regular) nuclear accent, and (2) early nuclear accent.

Table 1. Sample materials used in Experiments 1, 2, and 3. Nuclear vs. non-nuclear accent status in early and late sentence positions. Experiments 1 and 2 used cross-modal naming, and Experiment 3 used phoneme monitoring. The critical words are underlined, and nuclear accented words are in boldface capital letters.

		Exp. 1	Exp. 2	Exp. 3
Early Sentence Position				
(1)	A boat was near the **TOWER**. H* H* L-L%	Related: SHIP	Identical: BOAT	/b/
(2)	A **BOAT** was near the tower. H* L-L%	Unrelated: SHOP	Unrelated: BOX	
Late Sentence Position				
(1)	The baby saw the **CAT**. H* H* L-L%	Related: DOG	Identical: CAT	/k/
(2)	The **BABY** saw the cat. H* L-L%	Unrelated: DUST	Unrelated: CLOCK	

Procedure. Subjects were seated at a computer and wore headphones with a microphone mounted on the headset. The sentences were presented over the headphones, and the target words were shown on the computer screen. At the acoustic offset of the prime word, the computer presented the target word and started a millisecond timer. Subjects named (read aloud) the target word, and the sound of the subject's voice stopped the timer and cleared the computer screen.

Results

Figure 2 shows the mean RTs. The data were analyzed in four-way ANOVAs, by subjects (F1) and by items (F2). For greater detail see [6].

In Exp. 1, the main effects of sentence position (early vs. late) and target relatedness were highly significant, as expected. Accent status (nuclear vs. non-nuclear), however, was only marginally significant, with nuclear accents slower than non-nuclear accents ($F1(1,36)=3.8$, $p=.06$; $F2(1,84)=3.0$, $p=.09$). The two-way interaction of Accent status x Relatedness, where the unrelated targets showed a larger effect of accent status than the related targets, was marginally significant by subjects ($F1(1,36)=3.2$, $p=.08$) and significant by items ($F2(1,84)=4.1$, $p<.05$).

In Exp. 2, the two-way interaction of Position x Relatedness and the main effects of Position and Relatedness were highly significant. However, Accent status was not significant ($F1(1,36)=2.4$, $p=.13$; $F2(1,84)=2.0$, $p=.16$).

Considering Exps. 1 & 2 together, Accent status was significant ($F1(1,72)=6.0$, $p=.02$; $F2(1,84)=4.1$, $p<.05$). Accent status x Relatedness was marginally significant ($F1(1,72)=3.3$,

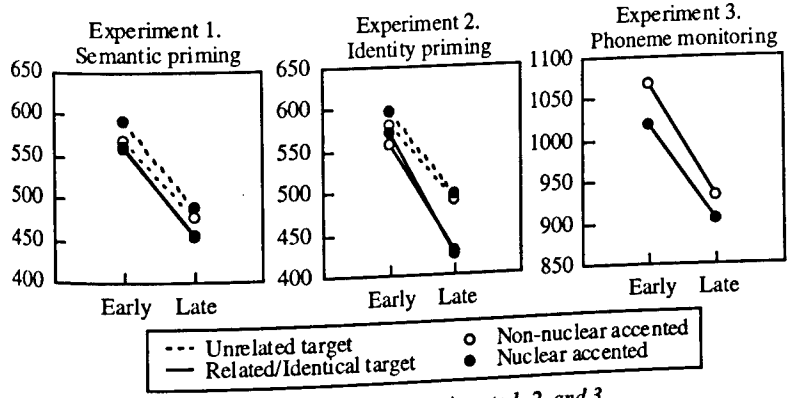


Figure 2. Mean reaction times (in ms) for Experiments 1, 2, and 3.

Table 2. Sample materials used in Experiment 4 (phoneme monitoring). Accent status in early and late sentence position of four intonation contours. Contours are characterized by late position accent status: emphatic, regular, and downstepped nuclear accents, and unaccented (early nuclear accent placement). The critical words are underlined, and nuclear accented words are in boldface capital letters.

			Early	Late
(1)	Emphatic	The poet admired the **CANYON**. H* L+H* L-L%	/p/	/k/
(2)	Regular	The poet admired the **CANYON**. H* H* L-L%	/p/	/k/
(3)	Downstepped	The poet admired the **CANYON**. H* !H* L-L%	/p/	/k/
(4)	Unaccented	The **POET** admired the canyon. H* L-L%	/p/	/k/

p=.07; F2(1,84)=3.3, p=.07). Accent status x Position was marginally significant by subjects; nuclear accents were relatively slower than non-nuclear accents in early position than in late position (F1(1,72)=3.8, p=.06; F2(1,84) =1.8, p=.18).

PHONEME MONITORING

Exps 1 & 2 showed that accent status had very little effect on lexical access. The next two experiments used a task known to be sensitive to differences in accent status. In phoneme monitoring, response times are faster to target phonemes in words with 'sentence stress' than words that are unstressed [7].

Method

Subjects. 100 undergraduate students participated in the two experiments, 20 in Exp. 3 and 80 in Exp. 4.

Stimuli. Target phonemes occurred only once in each sentence, as the initial consonant of a critical word. Exp. 3 used 40 critical sentences with the same intonation contours as Exps. 1 & 2. Sentences had one target phoneme (/p/, /b/, /k/, or /g/) in either early or late sentence position. Exp. 4 used 96 critical sentences with the four intonation patterns described above. The phoneme targets were /p/ and /k/, one in early position and one in late position of each sentence. Sample materials are shown in Tables 1 & 2.

Procedure. Subjects were seated at a computer and wore headphones. The sentences were presented over headphones, and subjects pressed the 'yes' response button when they detected the target phoneme. The computer started a timer at the release burst of the stops, and the button press stopped the timer. In Exp. 3 the target phoneme was specified before each sentence by an auditory phrase, e.g., "Listen for /k/ as in 'car'." In Exp. 4 the target phoneme was specified before each sentence by a visual display of the letter 'P' or 'K'.

Results

Figures 2 & 3 show the mean RTs for Exp. 3 (plotted by accent status and contour type, respectively). The data were analyzed in two three-way

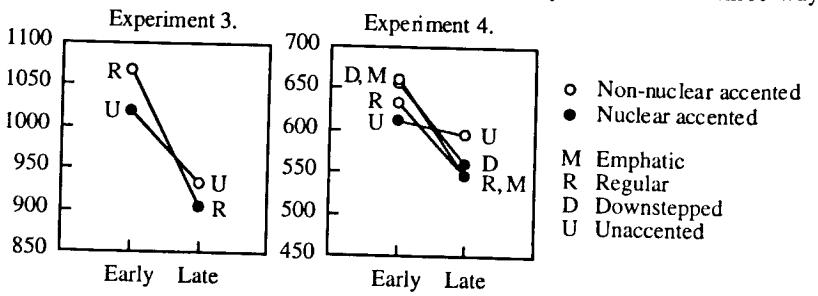


Figure 3. Mean reaction times (in ms) for Experiments 3 and 4.

ANOVAs. The main effect of Accent status was significant (F1(1,18)=5.8, p=.03; F2(1,36)=5.3, p=.03). Nuclear accented words had faster RTs than non-nuclear accented words. Early targets were also significantly slower than late targets. There was no significant Accent status x Position interaction.

The data in Exp. 4 were analyzed in two three-way interaction of Position x Contour was highly significant (F1(3,216)=17.4, p<.001; F2(3,264)= 15.3, p<.001) and the main effect of Contour was significant (F1(3,216)=3.2, p=.02; F2(3,264)=2.5, p=.06).

Condition mean contrasts were calculated in order to explore the two-way interaction. In both early and late position the RT to words with nuclear accent was faster than those with non-nuclear accent (Early: F1(1,216)=19.2, p<.001, F2(1,264)=15.4, p<.001; Late: F1(1,216)=32.7, p<.001, F2(1,264)= 31.1, p<.001). As in Exp. 3, early nuclear accented words were faster than the prenuclear accented words of the regular contour (F1(1,216)=5.2, p=.02; F2(1,264)=5.1, p=.02). The prenuclear accented words of the regular contour were also faster by subjects than those of the emphatic and downstepped contours (F1(1,216)=4.9, p=.03; F2(1,264)=2.7, p=.10). In late position, the downstepped nuclear accents were significantly faster than the unaccented words (F1(1,216)=12.9, p<.001; F2(1,264)= 12.9, p<.001), and the regular and emphatic nuclear accents were marginally faster than the downstepped nuclear accents (F1(1,216)=3.5, p=.06; F2(1,264)=2.8, p=.09).

DISCUSSION

In the cross-naming experiments, accent status did not strongly affect lexical access. For lexical priming, basically a word is a word, no matter whether it is nuclear accented or completely unaccented. However, target words that were primed by words with early nuclear accents were named somewhat more slowly than those with prenuclear accents, suggesting that there may be something 'not normal' about early nuclear accent placement. The difference in reaction time is perhaps best explained by the listener's placing greater attention on the early nuclear accented word when it occurs, which subsequently slows down the naming task.

In the phoneme monitoring experiments, phonemes were detected most quickly in nuclear accented words. However, phonemes were detected less quickly in downstepped nuclear accented words than in regular and emphatic nuclear accented words. This suggests that downstepped accents have less acoustic prominence than the other two types of nuclear accents. Also, phonemes in prenuclear accented words of sentences with downstepped and emphatic nuclear accents were detected less quickly than those in sentences with regular nuclear accents, which is yet to be explained.

Nuclear accent type and location do influence sentence processing, and nuclear accent is not a completely uniform category in terms of prominence.

REFERENCES

[1] Horne, M. (1991), "Why do speakers accent 'given' information?", *Proceedings of Eurospeech 91*, vol. 3, pp. 1279-1282.

[2] Pitrelli, J.F., Beckman, M.E., & Hirschberg, J. (1994), "Evaluation of prosodic transcription labeling reliability in the ToBI framework", *Proceedings, 1994 International Conference on Spoken Language Processing*, vol. 1, pp. 123-126. Yokohama, Japan.

[3] Pierrehumbert, J. (1979), "The perception of fundamental frequency declination", *J. of Acoustical Society of America*, vol. 66, pp. 363-369.

[4] Foss, D.J. (1969), "Decision processes during sentence comprehension", *J. of Verbal Learning & Verbal Behavior*, vol. 8, pp. 457-462.

[5] Forster, K.I. (1981) "Priming and the effects of sentence and lexical contexts on naming times", *Quarterly J. of Experimental Psychology*, vol. 33A, pp. 465-495.

[6] Ayers, G.M. (to appear), *Nuclear accent types and prominence: Some psycholinguistic experiments*, Ph.D dissertation, Ohio State University.

[7] Cutler, A. (1976), "Phoneme-monitoring reaction time as a function of preceding intonation contour", *Perception & Psychophysics*, vol. 20, pp. 55-60.

INFLUENCE OF FOCUS STRUCTURES ON TONAL TARGETS OF PITCH PEAKS

H. H. Rump
Institute for Perception Research/IPO, Eindhoven, The Netherlands

ABSTRACT

The purpose of the present study was to find out how 'equal prominence' and the peak heights of two pitch accents are related to the focus structure of an utterance. Subjects adjusted the height of one of two pitch peaks, matching the pitch contour to four different focus structures. The results suggest the existence of target values for the pitch peaks for each of the different focus conditions.

INTRODUCTION

In previous experiments involving utterances with two pitch accents (e.g. [6], [7], [8]) it appeared that the height of the second pitch peak was somewhat less than that of the first when the peaks lent equal prominence. The heights of the first and the second peaks turned out to be linearly related to each other. This was found for both nonsense and meaningful utterances, and having baselines with or without declination.

From these previous experiments, it is not clear, however, how equal prominence is related to the focus structure of the utterance in which the pitch accents occur. It may be assumed that equal prominence occurs either in the pragmatic context of the broad focus structure, in which the whole utterance is in focus, or in that of the double focus structure, in which the accented syllables are in focus separately. It may be assumed that the double focus structure is prosodically marked by higher pitch peaks, although this is not clear from the literature ([1], [2]). Neither is this clear for the previous experiments, however, since the focus structure of the test utterances was not made explicit. In the present experiments, different focus conditions were used in order to test how the heights of two pitch peaks are related to the focus structure of an utterance.

METHOD

The utterance used in the present experiments was "A'manda gaat naar 'Malta" (Amanda goes to Malta), which was spoken by a male person. It contained two accented syllables, /man/ and /mal/, with associated pitch peaks P1 and P2, respectively. The pitch accents had rising-falling pitch contours while the pitch in the unaccented syllables declined along a baseline which was a straight line in the ERB-rate frequency domain (unit: E; [4]). The starting frequency was 137 Hz, the end frequency was 100 Hz. The duration of the utterance was 1.45 s. The rate of declination was 0.7 E/s.

Two experiments were performed. In the first one, the height of P1 was fixed while the subjects adjusted the height of P2. In the second one, the height of P2 was fixed while the height of P1 was adjusted. Pitch manipulations were performed using the PSOLA method ([3]). Listeners selected the appropriate pitch contours from a prepared set of stimuli.

The task of the subjects was to adjust the height of a given pitch peak so that the resulting pitch contour would fit as close as possible one of four given focus structures. Broad focus was meant to give a neutral reading to the utterance. Single-focus conditions were elicited by asking questions which would result in contrastive readings of the target utterance so that only one of the pitch accents was in focus and the other one was not. In the double-focus condition, both accents were meant to be contrastive at the same time. The four instructions were:

- Wat zei je dat er gaat gebeuren?
(intended focus structure: broad focus)
- Gaat Jan naar Cyprus? Nee,...
(double focus)
- Gaat Jan naar Malta? Nee,...
(single focus on P1)
- Gaat Amanda naar Cyprus? Nee,...
(single focus on P2)

The instructions were printed to the computer screen while the test utterance 'Amanda gaat naar Malta' was made audible through headphones. The broad-focus and the double-focus conditions were expected to represent the 'equal-prominence' conditions. The single-focus conditions were included because they were expected to represent explicit 'different-prominence' conditions, thus providing some kind of boundaries for the equal-prominence conditions.

ADJUSTMENTS OF P2

The first group of subjects adjusted the height of P2 so that the utterance with the resulting pitch contour would be an adequate answer to the question/instruction which was written on the screen. During each trial, the height of P1 was fixed at one of three different values: 165, 183, or 202 Hz. The adjustments started at both extremes of the peak height continuum of P2 which ranged from 110 to 214 Hz, corresponding to excursion sizes of zero to 2.5 E. The range was divided in 10 steps of 0.25 E (about 1.5 st). Each adjustment was repeated twice, so that a subject completed four trials per instruction per P1 height. The order of presentation was completely random.

The ten subjects were students and research staff of the institute. They were all native speakers of Dutch, and they were not working on speech.

Results

The results averaged across all subjects are presented in Table 1. Every subject adjusted the height of P2 to be almost maximal when P2 was in single focus, and to be almost minimal, i.e. having excursion size zero, when P1 was in single focus.

The effect of Instruction on the adjusted height of P2 was highly significant ($F_{(3,27)} = 70.8$, $p < 0.001$). It is remarkable to find that the effect of P1 height on the adjusted height of P2, however, was not significant ($F_{(2,18)} = 2.94$, $p < 0.08$). The difference between the adjusted heights under the broad-focus and double-focus conditions turned out not to be significant ($F_{(1,27)} = 0.69$, $p > 0.05$), although the height of P2 tended to be greater under the double-focus than under the broad-focus condition.

Table 1. Adjusted P2 heights (Hz) under four different focus conditions and for three fixed heights of P1 (Hz).

	focus			
	P2	broad	double	P1
P1				
165	116	157	170	201
183	111	162	167	199
202	111	164	178	206

Discussion

The results for the single-focus conditions were as expected. If P2 was in focus, its height was made almost as large as possible. If P1 was in focus, the height of P2 was adjusted to be as small as possible. The latter is also in line with the theory that the last accented word in an utterance contains the nuclear accent. Focusing on the first accent, making it the nuclear accent, implies that the second accent should be deaccented, i.e. its excursion size should be zero.

For the broad-focus and double-focus conditions there were some individual differences. Some subjects adjusted P2 to be higher under the double-focus than under the broad-focus condition, while others adjusted P2 to be lower under the double-focus than under the broad-focus condition. This may explain why the difference between the two conditions was not significant. This may have been due to the fact that it is very difficult to interpret a neutral reading if you hear the same utterance again and again. The resulting annoyance may then have resulted in a non-neutral reading with a relatively high P2.

ADJUSTMENTS OF P1

The same utterance was tested with a second group of subjects. Again ten subjects participated, meeting the same selection criteria as above. They now adjusted the height of the first pitch accent, P1. The different values of P2 height were 143, 160, and 179 Hz. The P1 height continuum ranged from 131 to 267 Hz, corresponding to excursion sizes of zero to 3 E in twelve steps which were equidistant in E (0.25 E or about 1.5 st). The instructions were the same as the ones in experiment I. The order of presentation was again completely random.

Results

The results averaged across the subjects are listed in Table 2. The effect of Instruction was again highly significant ($F_{(3,27)} = 58.0$, $p < 0.01$). Again, unexpectedly the height of the fixed peak (P2) did not systematically influence the adjusted height of the other peak (P1). P1 height was adjusted to about the maximum value when P1 was in focus. If P2 was in focus, however, the adjusted height of P1 was on average more than 15 Hz above the minimum, resulting in an excursion size of about 2 st.

The double-focus condition resulted in significantly higher adjusted P1 values than the broad-focus condition ($F_{(1,27)} = 5.03$, $p < 0.05$).

Table 2. Adjusted P1 heights (Hz) under four different focus conditions and for three fixed heights of P2 (Hz).

	focus			
P2	broad	double	P1	
P2				
143	149	175	210	250
160	141	174	209	250
179	150	174	219	249

Discussion

If P1 was in focus, its height was adjusted to be as large as possible. Some of the subjects complained that they could not manipulate the height of P2, so that the resulting pitch contour was not optimal, P1 still sounding as a prenuclear accent.

If P2 was in focus, the excursion size of P1 was still about 2 st, so that the resulting peak height was only slightly below the average peak height of P2. This means that the excursion size of the pitch accent on the prenuclear accent may be larger than zero although it is deaccented. This is sometimes called a thematic or rhythmical accent, not lending much prominence to the word containing the accented syllable.

In the broad-focus condition, the average excursion sizes of P1 and P2 were about equal. In the double-focus condition, P1 was much higher than P2. This was true for the results of almost every single subject.

GENERAL DISCUSSION

The results show that the focus structure was crucial for the finally adjusted peak heights, and that the height of the other, fixed pitch peak had hardly any influence. In other words, it was mainly the focus structures that determined the resulting overall pitch contours. The results for the single-focus conditions were as expected in both experiments. The peak heights were large when the accented syllable was in the focused word, and they were small when the word containing the target syllable was explicitly out of focus (deaccented).

The difference between the broad-focus and double-focus conditions was most marked for P1. It was adjusted to significantly higher values for the double-foucs than for the broad-focus condition. For P2, this was found only as a tendency.

Unlike the previous experiments on prominence, the height of the fixed peak of one pitch peak had no systematic influence on the adjusted height of the other. This may indicate that when the pragmatics of prominence are involved, just one peak height represents a target value, which should be reached in order to obtain the appropriate pitch contour. This conclusion is supported by the findings reported in [5], where it was found that the pitch measured at certain points in the pitch contour is quite constant not only for a given speaker but also for a given instruction.

When we combine the results of the two experiments into only one pitch contour per focus condition, it is found that under the broad-focus condition the topline, connecting P1 and P2, and the baseline turn out to be about parallel. Under the double-focus condition, however, the topline turns out to be much steeper than the baseline.

It is not clear yet, however, whether listeners will be able to recognize an intended focus structure when they hear the pitch contour which is created using the pitch values obtained from the present experiments. This will be tested in a follow-up experiment.

REFERENCES

[1] Bartels, C., and Kingston, J. (1994), "Salient pitch cues in the perception of contrastive focus", in: Focus & Natural language processing, Proc. of a conference in celebration of the 10th anniversary of the Journal of Semantics, Meinhard-Schwebda, Germany, vol. 1, Intonation and Syntax, pp. 1-10.

[2] Batliner, A. (1994), "Prosody, Focus, and Focal Structure: Some Remarks on Methodology", in: Focus & Natural language processing, Proc. of a conference in celebration of the 10th anniversary of the Journal of Semantics, Meinhard-Schwebda, Germany, vol. 1, Intonation and Syntax, pp. 11-28.

[3] Hamon, C., Moulines, E., and Charpentier, F. (1989), "A diphone synthesis system based on time domain prosodic modifications of speech", *Proc. ICASSP-89*, pp. 238-241.

[4] Hermes, D. J., and Van Gestel, J. (1991), "The frequency scale of speech intonation", *J. Acoust. Soc. Am.*, vol. 90, 97-102.

[5] Ladd, D. R., and Terken, J. M. B. (1995), "Modelling intra- and inter-speaker pitch range variation", *Proc. ICPhS-95*, Stockholm, Sweden.

[6] Pierrehumbert, J. (1979), "The perception of fundamental frequency declination", *J. Acoust. Soc. Am.*, vol. 66, 363-369.

[7] Repp, B. H., Rump, H. H., and Terken, J. M. B. (1993), "Relative perceptual prominence of fundamental frequency peaks in the presence of declination", *IPO Annual Progress Report*, vol. 28, 59-62.

[8] Terken, J. M. B. (1991), "Fundamental frequency and perceived prominence of accented syllables", *J. Acoust. Soc. Am.*, vol. 89, 1768-1776.

THE EFFECT OF EMPHATIC ACCENT ON CONTEXTUAL TONAL VARIATION

Y. Xu
Research Laboratory of Electronics
Massachusetts Institute of Technology, Cambridge, USA

ABSTRACT

The present study examines the effect of emphatic accent on different syllables in disyllabic words in order to learn more about the nature of contextual tonal variation. Comparisons among the three accent conditions suggest that the tone that receives the emphatic accent probably exerts stronger influence on the less emphasized tone than the other way around. A new carryover effect was also found with which a Tone 3 raises the final portion of a tone that follows it.

INTRODUCTION

Recently, Xu [1][2] found that F_0 contour of a tone in Mandarin is perturbed differently by the tone preceding it than by the tone following it. The influence of the preceding tone (carryover effect) is assimilatory: the beginning of the F_0 contour of a tone becomes similar to the ending pitch value of the preceding tone. The influence of the following tone (anticipatory effect) is mostly dissimilatory: the F_0 maximum of a tone dissimilates from the F_0 minimum of the following tone. It was also found that evidence of this kind of asymmetry could be seen in data reported in other studies on Mandarin [3] and on Thai [4]. In order to learn more about the nature of these contextual tonal variations, the present study examines how contextual tonal variation patterns may change under different emphatic accents in disyllabic words. The carryover assimilation effect is expected to be stronger when the first syllable in a disyllabic sequence receives emphatic accent. The anticipatory dissimilation effect, however, is expected to be either strengthened, kept the same, or reduced when the second syllable receives emphatic accent, depending on the nature of the mechanism that causes the dissimilation.

MATERIAL

Mandarin has five lexically stressed tones — tones 1 through 4 — whose typical F_0 contours are high-level, mid-rising, low-dipping, and high-falling. There is also a lexically unstressed tone — the neutral tone or Tone 5 — whose actual F_0 contour has much greater dependence on the adjacent tones than the stressed tones. In connected speech, however, even the lexically stressed tones show extensive variation due to influence from adjacent tones (Shih [5], Xu [6]).

Following Xu [1][2], disyllabic sequence /mama/ with all sixteen possible combinations of the four lexically stressed tones were used as production material. Four male native speakers of Mandarin produced all those sequences in isolation. The sequences, all but one are nonwords, were printed in Chinese characters on the reading list. Subjects were requested to produce the sequences with emphasis on the first or the last syllable, or with no emphasis on either syllable. A prerecorded pacing tape was used to control the speaking rate. On the tape were groups of six beeps with intervals of three seconds. The speakers thus repeated each sequence six times, each repetition following a pacing beep.

F_0 CONTOUR EXTRACTION

The utterances were digitized at a sampling rate of 16 kHz. A program in a commercial signal analysis package (ESPS by Entropic Inc.) was used to label each vocal pulse in the utterances. The labeled signal files were then hand-edited to correct spurious labeling and to mark segment boundaries between /m/ and /a/. The editted files were further processed by locally developed computer programs to transform the distances between neighboring labels into F_0 values. The F_0 curves thus obtained were smoothed using a simple window function that eliminates any bumps or sharp edges greater than two Hertz. Average segmental duration across the repetitions was also computed. In order to visually examine the F_0 variations, the F_0 contours were time-normalized within each of the two nasal and two vocalic segments, and then plotted in separate groups. Example plots are shown in Figure 1.

the middle panel shows them produced with emphasis on the second syllable; and the bottom panel shows them produced with emphasis on the first syllable.

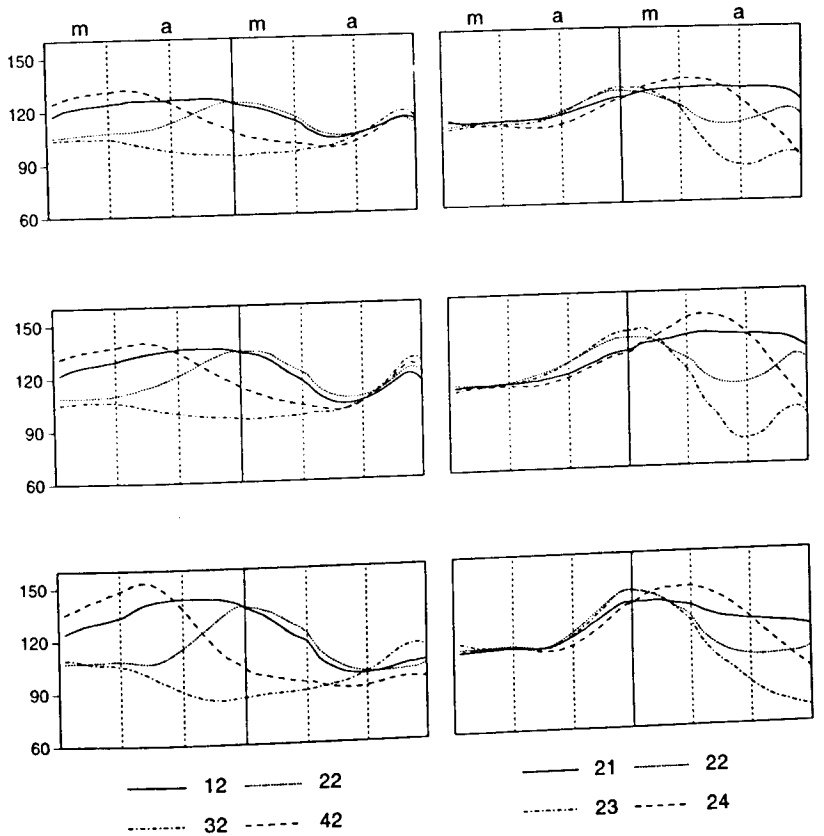


Figure 1. Time-normalized F_0 contours of Tone 2 when preceded (left) or followed by different tones. Those tonal combinations were produced with no particular emphasis (top), emphasis on the second syllable (middle), or emphasis on the first syllable (bottom).

ANALYSIS

In Figure 1, the carryover and anticipatory effects on a target tone are illustrated by the case of Tone 2. In the left column, Tone 2 is shown to be *preceded* by four different tones; the right column shows Tone 2 *followed* by four different tones. In each column, the top panel shows the ditonal sequences produced with no particular emphasis;

Carryover Effects

The most obvious carryover effect, as can be seen in Figure 1, is the assimilation of the starting F_0 value of the second syllable to the ending F_0 value of the first syllable. The tones in the first syllable end with distinct F_0 values, whereas the starting F_0 values of the same tone, here Tone 2, closely follow

the ending F_0 of the previous tone, resulting in a wide range of starting F_0 contours for the same tone. The differences caused by the preceding tone remain until about a quarter of the way into the vowel of the second syllable. A three factor (tone of syllable 1, tone of syllable 2, and emphasis pattern) ANOVA found the overall difference in F_0 caused by the tones of the first syllable to be highly significant at the beginning of the vocalic segment, $F(3, 9) = 29.8$, $p < .001$, and still significant by the end of the first quarter of the vocalic segment, $F(3, 9) = 5.9$, $p < .05$.

The effect of different emphasis conditions on the tone of the second syllable also can be seen in Figure 1. In general, the two conditions in which there is emphasis on one of the syllables show stronger influence on the initial portion of the F_0 contour of the second syllable. The effect is the strongest when the first syllable receives emphatic accent. An ANOVA test found interaction between the effect of the tone of the first syllable and the effect of emphasis on the beginning ($F(6, 18) = 9.52$, $p < .001$) and first quarter of the following tone ($F(6, 18) = 3.98$, $p < .05$).

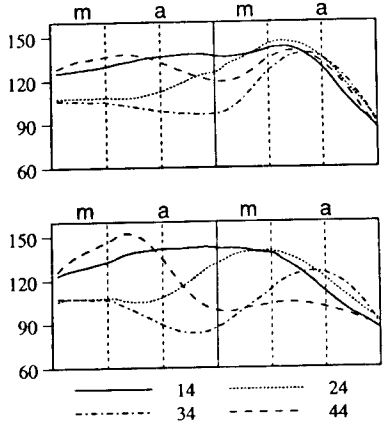


Figure 2. F_0 contours of Tone 4 preceded by four different tones. In the upper panel, the second syllable was emphasized; in the lower panel, the first syllable was emphasized.

There is a second kind of carryover effect which has not been discussed before: Tone 3, the low tone, raises the F_0 value in certain portions of the following tone. This is a seemingly dissimilatory effect, because the ending F_0 value of Tone 3 is the lowest among all the tones. This raising effect is rather peculiar in that it is the strongest near the end of the following tone. In Figure 1, the final portion of Tone 2 is the highest when preceded by Tone 3 than by the other tones. A similar effect was also found on Tone 1. As shown in Figure 2, Tone 4, which has low ending F0, does not show a higher ending F_0 when preceded by Tone 3. Instead, there seems to be a delay effect so that the point of maximum F_0 in Tone 4 appears later when preceded by Tone 3 than by other tones. This phenomenon is the most obvious when the emphasis is on the first syllable. Considering the fact that the canonical form of Tone 3 has a final rise, it is likely that this final boost in the tones following Tone 3 is actually an alternative way of manifesting the canonical final rise in Tone 3, and would explain why an emphasis on the first syllable would produce the biggest raising effect. Interestingly, a re-examination of the F_0 data for one speaker reported by Shih [3] also found that Tone 2 had a higher ending F_0 when following Tone 3 than when following other tones.

Anticipatory Effect

The plots on the right in Figure 1 demonstrate the influence of a following tone on the preceding tone. First, there is a much smaller range of variation in a given tone when it is followed by different tones than when it is preceded by different tones. The only visible anticipatory effect in Figure 1 is dissimilatory: when the tone of the second syllable is Tone 2 or 3, both of which have low starting F_0 values, the maximal F_0 of the preceding Tone 2 is higher than in other two cases. A three factor ANOVA found this difference to be significant, $F(3, 9) = 7.84$, $p < .01$.

Comparison among the three emphasis conditions suggests that this kind of anticipatory dissimilation is strongest when the second syllable is emphasized. However, an overall ANOVA including the measurements from all four speakers did not show a significant interaction between emphasis condition and the effect of preceding tone, whereas analysis of data on individual speakers did show significant interaction for each speaker. It seems that data from more speakers are needed to reach a more definitive conclusion on this effect.

DISCUSSION

While it is not surprising to see that a tone under emphatic accent exerts a stronger assimilation effect on the tone that follows it, it is quite interesting to see that the anticipatory dissimilation reported before was found to be well preserved and probably even boosted when the tone of the second syllable receives emphatic accent. The finding that the tone of the second syllable did not "spread" its initial pitch value leftward when it is under emphatic accent suggests either a) that the ending portion of the pitch contour of a tone is so important that it is not altered even when the emphasis is on the tone that follows it; or b) that the pitch value of the initial portion of a tone is totally undefined so that even an emphasis on that tone would not help it to impose any particular assimilatory influence on the preceding tone.

More interestingly, the raising of F_0 before a low tonal target does not seem to be unique to the East Asian tone languages. A similar phenomenon was also found in Yoruba, an African tone language [7][8]. Further investigation on the generality of this finding in other languages will be needed.

SUMMARY

Comparisons among the three accent conditions confirms that carryover assimilation is strongest when the first syllable is emphasized. However, a new effect is also observed when the first syllable is emphasized: Tone 3, which has the lowest minimum F0, exerts a dissimilation effect on the following tone, raising the final portion of its F_0 contour. As for the anticipatory effect, dissimilation is found in all three accent conditions, but the magnitude of the effect seems to be strongest when the second syllable receives accent. In short, for both carryover and anticipatory effects, the accented syllables seem to exert greater influence on the unaccented syllables than the other way around, regardless of the nature of the influence.

ACKNOWLEDGMENT

This study was supported by NIH Grant T32DC00038.

REFERENCES

[1] Xu, Y. (1993), *Contextual tonal variation in Mandarin Chinese*. Ph. D. dissertation, The University of Connecticut.

[2] Xu, Y. (1994), "Asymmetry in contextual tonal variation in Mandarin", In H.-W. Chang, J.-T. Huang, C.-W. Hue, & O.J.L. Tzeng (Eds.), *Advances in the study of Chinese language processing Volume 1*. (pp. 383-396). Taipei: Department of Psychology, National Taiwan University.

[3] Shih, C. (1986), "The phonetics of the Chinese tonal system", *AT & T Bell Laboratories technical memorandum.*

[4] Gandour, J. (1992), "Anticipatory tonal coarticulation in Thai noun compounds", in *Linguistics of the Tibeto-Burman Area* (pp. 111-124).

[5] Shih, C. (1992), "Variations of the Mandarin rising tone", in *Proceedings of the IRCS Workshop on Prosody in Natural Speech No. 92-37*, The Institute for Research in Cognitive Science, University of Pennsylvania.

[6] Xu, Y. (1994), "Production and perception of coarticulated tones", *Journal of the Acoustical Society of America*. vol. 95, pp. 2240-2253.

[7] Connell, B., & Ladd, D. R. (1990), "Aspects of pitch realization in Yoruba" *Phonology*, vol. 7, pp. 1-29.

[8] Laniran, Y. (1992), *Intonation in Tone Languages: The phonetic Implementation of Tones in Yorùbá*,. Ph. D. dissertation, Cornell University.

STRATEGIES FOR FOCAL ACCENT DETECTION IN SPONTANEOUS SPEECH

Anja Petzold
Institut für Kommunikationsforschung und Phonetik, University of Bonn
Poppelsdorfer Allee 47, 53115 Bonn, Germany
email: ape@asl1.ikp.uni-bonn.de

ABSTRACT

In this paper a new method for detection of focus is developed. Speech data consists of German spontaneous speech from several speakers. At present the algorithm uses only the fundamental frequency values. By computing a nonlinear reference line through significant anchor points in the F_0 course, points of highest prominence are determined. The global recognition rate is 78.5 % and the mean recognition rate is 66.6 %.

INTRODUCTION

In the last years the use of prosodic information for support of automatic speech recognition systems has been widely extended. Prosodic features can be determined independently of the segmental level and therefore can provide recognition modules on higher levels (e. g. morphology, syntax, semantics) with additional help for decision. In this study prosody shall give help to a semantic recognition module by detecting the focus.

Focus is defined here as the semantically most important part of an utterance, which is in general marked by prosodic means. If the focus is marked otherwise (for instance by word order), prosody will no longer provide a salient contribution; in this case the focus has to be derived from the linguistic context. On the other hand, there are also prominent parts of an utterance, which carry information of less importance, for example exclamations and greeting stereotypes.

DATA

The speech material consists of dialogues of German spontaneous speech, containing meeting arrangements supplied within the research project VERBMOBIL. Focused areas in these dialogues contain information about time and place, like "thursday afternoon", "in my office", and also judgements like "that is ok for me", "fine" and so on.

Focus accents were labelled for 7 dialogues (154 turns with one or more phrases, 247 focal accents) with 6 different speakers (2 female, 4 male) by a phonetician (i. e. the present author) through acoustic perception. The size of the focus areas was left variable, there was no restriction to word or syllable boundaries.

METHOD

Already in earlier investigations [1] the prosodic features of focus were examined for German. A corpus of read speech with isolated sentences (containing 2 grammatical objects) was used. A statistical classification procedure (discriminant analysis) was implemented to decide which of the 2 objects was the focused one. F_0-maxima and minima of the object phrases and the difference of their positions on the time

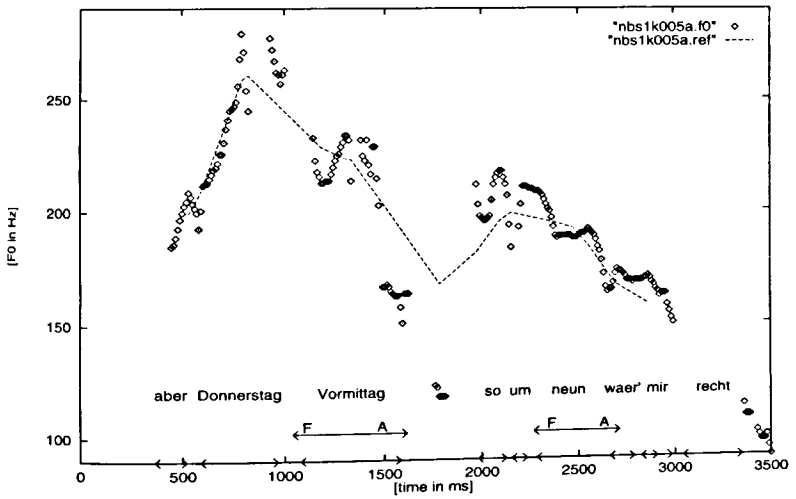


Figure 1. Utterance of a dialogue with reference line and labelled focus (FA). ("But thursday morning at about nine [o' clock] would be ok for me")

axis were found as the most significant feature variables. Duration and intensity were not so important for the decision.

This paper will try to solve focus recognition by global description of the utterance contour. At first we will just look at the fundamental frequency F_0. How can we now find the most prominent parts in the F_0 contour? There is no hope that we just take the absolute maxima, we have always to take in account declination, i. e., the fall of fundamental frequency toward the end of the utterance.

Investigations of Swedish spontaneous speech [2] have shown that declination can be controlled by the focal accent: It was found that in prefocal position there is no downstepping, but after a focal accent downstepping is significant and characteristic. We can suppose a physiological correlate for this effect: The physical effort for producing an utterance seems to be not equally distributed. The effort remains high until the focus is reached, after the focus the effort sinks to a significant lower level.

To examine this feature in German spontaneous speech, several possibilities for computing a reference line were tested. A good description of these problems is found in [3]. For our work we cannot use a linear declination line: for detecting a downfall after a focus, we have to look especially at the extrema of the F_0 course.

The reference line was computed as follows: First the F_0 contour was postprocessed by a special smoothing algorithm described in [4]. (Without smoothing results get worse by about 7 %.) In a second step significant maxima and minima in a window of 90 ms size were detected. The average values between the maximum and minimum lines yield the global reference line (see Figure 1).

FOCUS RECOGNITION

According to the already mentioned Swedish investigations the focus must be in the area of the steepest fall in the F_0 course. Therefore the points with the highest negative gradient were determined first in each utterance. There was no limitation for the number of focal accents in a sentence or phrase. Phrase boundaries were not considered. Minimum length for a fall was set to 200 ms.

Starting from the points of steepest fall, how can we now get to the position of focus? For the time being, we assumed as simplest solution the nearest maximum in this region to be the focus. In further experiments we will also consider the relative height and intensity of the maxima, perhaps also a kind of duration measure.

RESULTS

In our data only about 20 % of the frames pertain to focused segments. To take account of this, two recognition rates will be displayed: first, the global recognition rate which denotes the percentage of correct classification regardless of focus or not and second, the mean recognition rate with equal weighting of focused and non-focused segments. This is illustrated in table 1.

As is shown in table 1, there are far more deletions than insertions, i. e., the recognition rate for focus areas is significantly worse than for nonfocus areas. But we have to bear in mind that in a collaboration of a prosody and a semantic recognition module it would be worse to have insertions of focal accents than to have deletions. Hints to focused areas shall only be an additional help for the semantics - without this help it can do its work as well. But false alarms could divert semantic analysis.

The different recognition rates for the dialogues reflect the degree of "liveliness". In a boring and monotone discussion even 'human recognizers' have problems to pick up the most important part of a message. So, the more engaged the discussion is, the clearer marked are the focal accents. No significant differences between male and female voices could be found.

DISCUSSION

Results are still not too satisfactory but in no way disappointing. The phenomenon of significant downfall after a focus in the F_0 contour appears to be strong enough to be useful for automatic focus recognition in German spontaneous speech. Moreover, there are a lot of possibilities left to optimize

Table 1. Focused parts and recognition rates in percent.

No. of Dialogue	Focused part	Total recognition		Recognition for	
		global	mean	focus areas	nonfocus areas
n001k	23.22	74.91	59.12	29.66	88.57
n002ka	21.57	76.17	66.23	47.13	85.33
n002kb	23.72	88.23	80.02	63.00	97.05
n002kc	17.59	77.60	55.79	20.53	91.05
n003k	16.15	76.92	66.95	51.00	82.91
n008k	7.52	74.52	67.42	56.03	78.82
n009k	16.69	81.24	71.10	53.45	88.74
Total	18.43	78.51	66.66	45.82	87.49

the results.

First, there is the computation of the reference line. Most important is a correct smoothing of the F_0 values. Likewise there are a lot of ways to determine the points with the steepest fall and to detect the focus starting from these points.

Second, we have to think about the problem of labelling the focus. To which extent the acoustic perception is influenced by semantic knowledge? Do we get the same results when labelling delexicalisized speech without semantic information but with intact prosodic structure? It is necessary to make further investigations in this direction; comparisons between different human labellers should be done as well.

Another open question is how to fix the size of the focus regions. As mentioned earlier, the size of the focus areas was left variable when labelling the focus accents. Therefore distinction between broad and narrow focus has not been made till now. As defined in [5], narrow focus is used for contrastive accents (just one syllable is in focus) and broad focus represents the 'normal' focused constituent (the whole word is put in focus), both expressed by a pitch accent on a syllable. At least for Dutch Sluijter & van Heuven [5] found that there are no acoustic differences in duration and pitch between a broad and a narrow focus accent. It seems that the distinction for these two kinds of focus has to be made rather at the linguistic than at the acoustic level.

Until now we did not take into consideration syntactic information like phrase boundaries or sentence modality. Phrase boundaries could help us to restrict focus determination to single phrases and therefore to divide the recognition task.

Sentence modality is another important fact. Already in [1] it is shown that in questions with a final rising contour the focus cannot be determined in the same way as in declarative sentences. We could expect another increase in recognition rate by separating questions and nonquestions.

ACKNOWLEDGEMENT

This work was funded by the German Federal Ministry of Education, Science, Research and Technology (BMBF) in the framework of the Verbmobil Project under Grant 01 IV 101 G. The responsibility for the contents of this study lies with the author.

REFERENCES

[1] Batliner A. (1989): Fokus, Modus und die große Zahl. Zur intonatorischen Indizierung des Fokus im Deutschen. In H. Altmann, A. Batliner, W. Oppenrieder, Zur Intonation von Modus und Fokus im Deutschen, Niemeyer

[2] Bruce G., Touati P. (1990): On the Analysis of Prosody in Spontaneous Dialogue, Working Papers, Lund University 36, 37 - 55

[3] Gussenhoven C., Rietveld T. (1994): Intonation contours and the prominence of F_0-Peaks, Proceedings of the ICSLP 1994 in Yokohama, 339 - 342

[4] Petzold A. (1994): Nachverarbeitung bei der Grundfrequenzbestimmung von Sprachsignalen zur Erfassung von Intonationskonturen, Fortschritte der Akustik - DAGA '94, 1345 - 1348

[5] Sluijter A., van Heuven V. J. (1995): Effects of Focus Distribution, Pitch Accent and Lexical Stress on the Temporal Organization of Syllables in Dutch, Phonetica 52, 71 - 89

LOUDNESS, SPECTRAL TILT, AND PERCEIVED PROMINENCE IN DIALOGUES

W. N. Campbell

ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan.

ABSTRACT

This study explores the correlation between spectral tilt and perceived prominence in the continuous speech of simulated conference-registration dialogues. It builds on previous work showing that syllable prominence and focus marking can be detected automatically, using differences in normalised segmental duration and energy, by introducing spectral information that compensates when the prosodic clues are weak or absent.

INTRODUCTION

There has been continuing debate about the relation between loudness and stress (see Beckman [1] for a summary). Early theories presented stress as having fixed phonetic levels (*c.f.*, Bloomfield, Trager & Smith, Chomsky & Halle), as being related to *force of utterance*, as tonetic (Kingdon), or as dependent on *pitch accents* (Bolinger). Beckman highlights the role of pragmatics in determining the accentual organisation of an utterance. In examining the phonetic correlates of stress and non-stress accent, she shows that syntagmatic accentual contrasts divide an utterance into a succession of shorter phrases in larger groupings, defining *stress* as a phonologically delimitable type of accent in which the pitch shape of the accentual pattern cannot be determined from the lexicon.

Early experimental evidence (from Fry and others) shows energy to be the weakest clue to stress, and fundamental frequency the strongest. Beckman, on the other hand, found metrical stress (at the level of the *prosodic word*), to be best explained by relative loudness (*i.e.*, temporal summation of waveform energy through the syllable nucleus, expressed relative to its duration), and she emphasises the trading relation between energy and duration in the perception of prominence.

Duration and energy

Previous work has confirmed that segmental duration and energy are both reliable cues for the automatic detection of prominence in read speech [2], and the present paper extends that work to show that spectral information is also present in the marking of prominence, and that it exhibits a trading relation with duration in interactive speech.

For the analysis presented in [2], a set of sentences extracted from a corpus of conference-registration dialogues was marked (by capitalising certain words) to show shift of focus, resulting in different stress patterns on the same sequence of words (as in *"Please take the SUBWAY to Kyoto station."*). Each sentence (30 in all) was given three or four different patterns, and a set of 100 of these was produced in three different utterance contexts. To study the way focus is marked in speech, we first asked the speaker to read the sentences in sequential order, and to "emphasise" the capitalised words. Here, with each set of interpretations grouped, the shift of focus was clearly contrastive. For the second reading, we asked her to read the same sentences in randomised order.

Finally, we recorded an interactive dialogue, where the same focus marking was produced by eliciting emphatic corrections of feigned misinterpretations.

The corpus of 300 focus-shifting utterances was then stress labelled to indicate perceived prominence. In order to remain somewhat theory neutral, I had the corpus labelled for accent type and for prominence location in three ways, by different labellers: (1) by simply marking the syllables perceived as prominent (an either-or decision), (2) using an O'Connor & Arnold variant of tonetic stress marks, and (3) more recently, using the ToBI system of tones and break indices. I then took the common subset of these three labellings as defining 'stressed syllables' for the purpose of this study. (However there was a high correlation between all three, and the different labellers seem to be identifying the same feature.)

DETECTION OF PROMINENCE

Because of the use of fundamental frequency in signalling more complex relations than simple prominence, this was not included as a factor for analysis (though it certainly plays a significant part in marking prominence). Instead correlations were examined between stressed syllables and measures of energy and duration normalised by segment type. Viewing the two acoustic measures independently, rather than combined as an energy integral over time, allowed better understanding of their individual contributions, and of the trade-off between them. Absolute values were not examined, but rather, for each phone class (as defined by label type in the segmentation), durations were normalised by expressing deviation from the class mean in terms of standard deviations of the distribution of that class. Similar segment-type normalisation was applied to the waveform energy, measured as average rms amplitude across the duration of each segment. Because these unit-less normalised scores have a zero mean (and a typical range of ± 3) a combined measure of their joint effect was derived by simply adding them. Taken separately, durational lengthening information detected 54% of the prominent syllables, and energy information detected 55%. Combined by summation, this detection improves to the average of 76% across all three speaking styles [2].

Table 1: Stress and focus detection

	A	B	C
stress detection:	92%	78%	72%
focus detection:	79%	78%	74%

Key: A: read grouped, B: read in randomised order, C: interactive

Between the three utterance styles, there was no significant difference in the detection of marked focus (*i.e.*, in identifying the syllable carrying the intended prominence) from amongst the set of syllables detected as stressed, but the initial detection of stressed syllables did vary as a function of speaking style. Table 1 shows that stressed syllables in read speech of grouped sentences were more easily detected than the equivalent syllables in randomly presented sentences or in interactive speech. Although also perceived as prominent, the latter were less easily discriminated by measures of duration and energy. Error analysis confirmed that in the more conversational interactive speaking style the prominences were still easy to discriminate by ear, but the acoustically-derived measures of stress were weaker. This paper attempts to explain why this may be so.

SPECTRAL FEATURES

Of the prominences not detected, 26% were clearly prominent to the ear but showed no significant excursion from the mean in duration, energy, or fun-

damental frequency. This implies that there are also phonation-style differences which can serve as clues to prominence and which may also be of use to automatic detection. This would be particularly useful since although durational information is robust, raw waveform envelope magnitude is *not* a robust measure, as it can vary considerably with distance from the microphone, or more globally reflecting changes in environmental noise.

Lindblom, in sketching the H&H theory [3], suggests a notion of sufficient discriminability to explain the continuum of hyper- and hypospeech observed in interactive dialogues, by which speakers tune their production to communicative and situational demands. This might account for the differences in results relating to speaking style, since in the interactive dialogues the speaker knows the extent of common knowledge with the hearer, and in the grouped presentation of contrasting pairs of utterances, she is more aware of the need to stress the contrast. Lindblom refers to Sundberg's work, on the long term average spectra of singers, in explaining possible mechanisms for the range of clarity of phonation. More recently, Sluijter & van Heuven [4], also citing such work on overall "vocal effort" as Gauffin & Sundberg [5], showed that, in Dutch, stressed sounds are produced with greater local vocal effort and hence with differentially increased energy at frequencies well above the fundamental.

We can measure such spectral tilt in several ways. At the lower end of the spectrum, as the difference in energy between the first and second harmonics, or at the upper end of the spectrum as a general increase of overall energy. A pilot study examining energy across 26 ERB-scaled spectral bands [6] confirmed that at least for read lab speech of English, spectral tilt significantly correlates with linguistic prominence under both high and low tones, for three different vowel types, and confirmed Sluijter's findings of increased energy in the higher spectral regions. This paper shows that for dialogue speech too, spectral information can be very helpful in discriminating prominences.

Table 2: Analysis of variance detection

df=(1,10048)	mean sqr	F
spectral tilt	113.35	603.5,
harmonic ratio	11.63	60.8469
energy (fund)	1.41	7.3856

Extraction of spectral data

Because segment labelling was done for all the dialogues, acoustic measures derived from the waveform can be related directly to individual syllables. To estimate spectral tilt, the fundamental frequency was extracted and then used to index into an fft of the speech waveform for each utterance so that a) the energy at the fundamental, and b) the harmonic ratio could be calculated. As a further measure, the average energy in the top third of the ERB-scaled spectrum (between 2kHz and 8kHz) was measured relative to the overall energy of each spectral slice as a measure of energy-normalised tilt.

These three indicators, normalised by phone type as for duration and energy above, were computed for the sonorant peak of each syllable and compared with the labelled prominences.

RESULTS

Of the 10,049 syllables in the 300 sentences, 2,951 were marked as prominent. There were 16 classes of vowel, none with less than 110 tokens. All had a representative number of prominent variants. Analysis of variance from a linear discriminant analysis predicting prominence as a binary feature on the basis of the three spectral factors showed all to make a contribution (significant at $p < 0.001$, see Table 2.). There were great differences though in the amount of the contribution of each, and energy in the upper areas of the spectrum was by far the clearest predictor of stress.

Interestingly, further factorisation of spectral tilt (as measured by the ratio of high-frequency energy to overall energy in the spectrum) according to speaking style, revealed that the greatest distinction between prominent and non-prominent syllables could be made for the spontaneous speech. See Table 3.

DISCUSSION

The above results confirm the correlation between spectral and prosodic information, and suggest that speakers also change their phonation according to the discourse context and type of information they impart. In style A (the grouped-presentation read-speech), the distinction between prominent and non-prominent syllables was clearly marked to accord with the capitalisation of the focussed word in the text. In the interactive case, when an interlocutor elicited the focus shift by misunderstanding selectively, the speaker was more personally involved in clarifying the meaning. This, too, resulted in a clearer articulation. However, for the intermediate case, where the focal shift was less markedly obvious, the distinction was less clear.

In all speaking styles, relative energy in the higher spectral regions proved the best correlate of prominence, and loudness (as measured by energy at the fundamental) the weakest. It is interesting that although the prosodically-based measures of duration and waveform envelope magnitude (amplitude) were weakend by the greater variation found in the more spontaneous rendition of the dialogues, the spectral measure was apparently strengthened. We can suppose that this trade-off is not coincidental, and in future work, include the spectral measures as well as the prosodic ones in the detection of prominence.

Table 3: ± 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.

ACKNOWLEDGEMENTS

I am particularly grateful to Mary Beckman and Osamu Fujimura for their helpful discussion and advice, and apologies to the many authors whose relevant work would have been cited given more space.

REFERENCES

[1] Beckman, M. (1986) **Stress & Non-Stress Accent**, Floris Publications.

[2] Campbell, W.N. (1992) "Prosodic encoding of English speech", *Proc ICSLP 92*, pp 663-666, Banff, Canada.

[3] 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.

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

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

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

PITCH STEREOTYPES IN THE NETHERLANDS AND JAPAN

R. van Bezooijen, T. de Graaf**, and T. Otake****
** University of Nijmegen, Nijmegen, The Netherlands*
*** University of Groningen, Groningen, The Netherlands*
**** Dokkyo University, Saitama, Japan*

ABSTRACT

Dutch and Japanese female speakers were presented at three pitch levels (low, original, high) to Dutch and Japanese male and female listeners in order to make a cross-cultural comparison of pitch stereotypes. Low pitch was cross-culturally associated with *large*, *strong*, *male-like*, *adult*, *independent*, and *arrogant*. High pitch was associated by the Dutch listeners, but not by the Japanese listeners, with *low prestige*. Finally, original pitches were found more *attractive* than either high or low pitches.

INTRODUCTION

Pitch is universally related to sex, women having higher pitched voices than men, and to age, children having higher pitched voices than adults. These relationships can to a large extent be explained in terms of concomitant differences in body size. Body size affects the dimensions, mass, and shape of the larynx, which in turn determine the ensuing pitch: pitch tends to be higher as the vocal cords are shorter and thinner. The relationship between body size and pitch is found not only among (categories of) humans but among other species as well: bears growl, mice squeak. According to Ohala's "frequency code" [1] there would exist a cross-cultural, cross-species form-meaning correspondence, by which low pitch is associated with physical size (tall and strong), and, by extension, with personality attributes such as aggressive, assertive, self-confident, dominant, self-sufficient, etc., and by which high pitch is associated with the opposite attributes. This hypothesis has not been tested. The present study examined the cross-cultural consistency of pitch stereotypes related to the frequency code in the Netherlands and Japan.

The Dutch and Japanese cultures were chosen because of the higher pitch which has been found to characterize young Japanese women as compared to Caucasian (American, Western European, Australian) women [2,3]. This difference in pitch has been tentatively explained by assuming the influence of sociocultural factors on pitch. The underlying assumption is that speakers adapt their pitch setting, probably largely unconsciously, so as to approximate vocal images reflecting socio-culturally desired personality attributes. Personality characteristics such as modesty, innocence, domesticity, subservience, and helplessness, according to the frequency code associated with high pitch, are traditionally more highly valued in women in Japanese culture than in Western culture [4]. Young Japanese girls and women might raise their pitch to conform to cultural stereotypes and to project the desired attributes.

In our study eight Dutch and eight Japanese female speakers were presented each at three pitch levels (low, original, high) to Dutch and Japanese male and female listeners to be rated on seven scales derived from Ohala's ideas, namely *short - tall*, *weak - strong*, *female-like - male-like*, *childlike - adult*, *dependent - independent*, *modest - arrogant*, and *low prestige - high prestige*. Assuming a universal basis of the frequency code, we hypothesized that both the Dutch and Japanese listeners would associate high pitch with the attributes named first and low pitch with the attributes named second (weak hypothesis). More stringently, we expected significant contrasts between all three pitches in the same direction (strong hypothesis). No interactions were expected between pitch and culture of speaker, nor between pitch and culture/sex of listener.

In addition, the scale *attractive - unattractive* was included to examine the subjective evaluations of different pitches in the Netherlands and Japan. We predicted an interaction between pitch and culture of listener in the sense that Japanese listeners would find high pitch more attractive than Dutch listeners. This outcome would fit in with the differences in sex roles in Japan and the Netherlands described above.

METHOD

Eight Dutch and eight Japanese women were selected as speakers; all were highly educated. The mean ages (ranges in parentheses) for the two groups were 33 years (20-48) and 29 years (21-42), respectively. The mean heights were 166 cm (161-171) and 163 cm (155-174). The differences in age and height, tested by means of *t*-tests for independent samples, were not significant at the 5% level.

All speakers read out the same neutral narrative text, the Dutch speakers in Dutch and the Japanese speakers in Japanese. Of each recorded speech sample three pitch versions were made: low, original, and high. The three versions were identical in all respects (tempo, intonation, pronunciation, etc.), except for the mean fundamental frequency. The pitch manipulations were carried out using the PSOLA (Pitch Synchronous Overlap and Add) technique [5]. To obtain the high and low pitch versions, the original pitches of the speakers were uniformly raised and lowered by .65 ERB [6]. The average pitches in the three pitch versions were 150 Hz (low), 180 Hz (original), and 212 Hz (high) for the Dutch speakers, and 155 Hz (low), 185 Hz (original), and 218 Hz (high) for the Japanese speakers.

The 8 (speakers) x 2 (cultures) x 3 (pitch versions) = 48 speech samples were presented to 30 Dutch subjects, 15 male and 15 female students at the University of Nijmegen, and 30 Japanese subjects, 15 male and 15 female students at Dokkyo University. Subjects rated each speech sample on eight seven-point scales, either in Dutch or Japanese: *short - tall*, *weak - strong*, *female-like - male-like*, *childlike - adult*, *dependent - independent*, *modest - arrogant*, *low prestige - high prestige*, *attractive - unattractive* (from now on the scales will be referred to by the attribute named second).

RESULTS AND DISCUSSION

The interrater reliability was assessed, separately for the Japanese and Dutch listeners and the Japanese and Dutch speakers, using Cronbach's alpha [7]. 28 out of the 32 coefficients exceeded .80. This means that the listeners agreed well on the distribution of the ratings over the stimuli, not only for in-group speakers but also for out-group speakers. The existence of vocal stereotypes for listeners and speakers speaking the same language has been evident since the 1930's [8]. However, evidence for listeners and speakers speaking different languages is still scarce. Scherer [9] found fair reliabilities for Germans rating American speakers but low reliabilities for Americans rating German speakers. Van Bezooijen [10], presenting Dutch speakers to British, Kenyan, Mexican, and Japanese listeners, obtained high reliabilities in all cultures for attributes similar to the ones examined in the

present study. However, attributes such as *reliability, sense of humor, openness, fairness*, and *attractiveness*, were rated less reliably in some or all of the cultures.

The factor pitch had a significant ($p=.0025$, namely .01/8 (the number of analyses)) main effect on all eight scales. There were no significant interactions of pitch with sex of listener, two of pitch with culture of speaker, and one of pitch with culture of listener. The two interactions with culture of speaker, pertaining to *male-like* and *unattractive*, were due to small deviations from parallelism and will be ignored. The interaction of pitch with culture of listener pertained to *prestige* and will be dealt with separately below.

In Table 1 the mean ratings for the three pitch levels and results of post hoc comparisons (Tukey's HSD) are given. For all seven scales derived from the frequency code the weak hypothesis (hw) was confirmed and for three the strong hypothesis (hs). It thus appears that the Dutch and Japanese listeners have identical associations of different pitch levels with speaker attributes in accordance with the frequency code. As expected, when speaking at a high pitch, speakers are cross-culturally perceived as less tall, less strong, less male-like, less adult, less independent, and less arrogant than when speaking at a low pitch. The perception of pitch is not obscured by listeners and speakers speaking different languages.

The only interaction between pitch and culture of listener, shown in Figure 1, pertains to prestige. The findings for the Dutch listeners are as expected: high pitch is associated with less prestige than low pitch. However, the expected effect is not found for the Japanese listeners, where high pitch seems to even raise prestige. The latter finding probably has to be placed within a more general framework of the

Table 1. Mean ratings for the three pitch levels. In the last two columns it is indicated whether the strong hypothesis (hs) and/or the weak hypothesis (hw) (see text) was confirmed (+) or rejected (-). These hypotheses were not formulated for unattractive.

	lo	ori	hi	hw	hs
tall	4.2	4.1	3.3	+	-
stron	4.8	4.5	3.7	+	-
male	2.7	2.1	2.1	+	-
adult	5.4	4.9	3.8	+	+
indep	4.8	4.5	3.6	+	+
arrog	4.2	4.0	3.4	+	+
prest	4.5	4.6	4.1	+	-
unatt	4.2	3.8	4.0		

role social prestige plays in the Japanese culture. The Japanese social structure is hierarchically structured to a high degree [11]. Pitch is one of the ways in which social differences can be signalled. Thus, the lack of prestige of Japanese women as compared to men has traditionally been reflected in high pitch. Although some changes have taken place in Japan in the direction of more western egalitarian principles, the pressure to conform to the traditional norms still seems to be high. It is not unlikely that in Japanese culture, with its emphasis on group behavior, conformation to norms may convey esteem and prestige. So, although high pitch may symbolize low status in a direct sense, it may in this case indirectly be associated with high status.

The scale *unattractive* was included to assess the subjective evaluation of pitch differences. There was an overall effect, with the original pitch of the speakers judged as the most attractive (see Table 1). If this effect is not due no artifacts of the pitch manipulations,

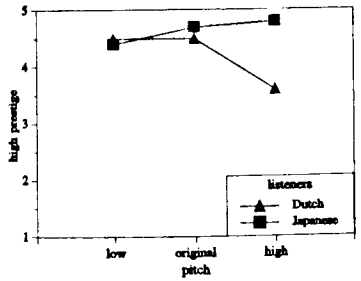


Figure 1. Interaction between pitch and culture of listener for prestige

it has to be concluded that both Dutch and Japanese people, either male of female, are content with the pitch Dutch and Japanese women have. This is a noticeable finding especially for the Japanese speakers included in the present study, as their original pitch was lower than generally reported in the literature. In the introduction we had hypothesized the Japanese listeners to find high pitch more attractive than the Dutch listeners. Of course, the Japanese people serving as listeners in this experiment were young and highly educated. Their ideas may be more oriented towards western egalitarian principles. Further research, with different subjects, may throw light upon this question.

ACKNOWLEDGMENT

This research has been made possible by a Fellowship of the Royal Netherlands Academy of Arts and Sciences.

REFERENCES

[1] Ohala, J.J. (1983), "Cross-language use of pitch: an ethological view", *Phonetica*, vol.40, pp.1-18.

[2] Yamazawa, H. & Hollien, H. (1992), "Speaking fundamental frequency of Japanese women", *Phonetica*, vol.49, pp.128-140.

[3] Loveday, L. (1981), "Pitch, politeness, and sexual role: an exploratory investigation into the pitch correlates of English and Japanese politeness formulae", *Language and Speech*, vol.24, pp.97-107.

[4] Smith, J.S. (1992), "Women in charge: Politeness and directives in the speech of Japanese women", *Language in Society*, vol.21, pp.59-82.

[5] Charpentier, F. & Moulines, E. (1989), "Pitch synchronous waveform processing technique for text to speech synthesis using diphones", *Proceedings Eurospeech Paris*, pp.13-19.

[6] Hermes, D.J. & Van Gestel, J.C. (1991), "The frequency scale of speech intonation", *Journal of the Acoustical Society of America*, vol.90, pp.97-102.

[7] Rietveld, T. & Van Hout, R. (1993), *Statistical techniques for the study of language and language behaviour*, Berlin and New York: Mouton de Gruyter.

[8] Addington, D.W. (1968), "The relationship of selected vocal characteristics to personality perception", *Speech Monographs*, vol.35, pp.492-503.

[9] Scherer, K.R. (1972), "Judging personality from voice: A cross-cultural approach to an old issue in interpersonal perception", *Journal of Personality*, vol.40, pp.191-210.

[10] Van Bezooijen, R. (1988), "The relative importance of pronunciation, prosody, and voice quality for the attribution of social status and personality characteristics", In R. van Hout and U. Knops (eds.), *Language attitudes in the Dutch language area*, Dordrecht and Cinnaminson: Foris Publications, pp. 85-103.

[11] Condon, J.C. (1984), *With respect to the Japanese*, Yarmouth: Intercultural Press.

THE USE OF A CATEGORY-PERCEPTION TEST IN THE STUDY OF ONGOING SOUND CHANGE

Isabelle Malderez

UFR Linguistique, Université Paris 7 - Denis Diderot, France

ABSTRACT

In this paper, I study the relationship between the category-test perception of subjects (dependent variable) and the sociolinguistic independent variables. I show that, for minimal pairs with a merging tendency of acoustical realizations or spelling mistakes by pupils, category perception of two successive generations is different [1].

PURPOSES

Janson [2] showed that category perception can reveal the ongoing changes between two phonems by comparing two populations which represent two successive generations. According to him, when there is an ongoing change, the two populations present different results as to the categorization of the same stimuli: the phonematic frontier in the continuum joining up the two phonotypes is not the same for the two groups of ages.

To emphasize a tendancy of merging of the /Ø/-/O/ opposition, I chose to study all the pairs of oral mid round vowels of French, to which I added the /a/-/O/ opposition. I thaugh that this pair might be subjected to an ongoing sound change as the /ã/-/ɔ̃/ one [3, 4]. The category perception test consists of five identical tests on the five following oppositions of French vowels: /œ/-/ɔ/, /ø/-/o/, /y/-/u/, /a/-/o/ and /a/-/ɔ/.

METHODS FOR THE STIMULI'S SYNTHESIS

The stimuli's synthesis was realized by Gérard Bailly at the Institut de la Communication Parlée (Grenoble). For each test, we used two typical vowels, synthetized by the 24 parameters defined in COMPOST [5, 6]. An interpolation program was built to produce 19 intermediate stimuli. It is a linear interpolation of the 24 parameters between the typical values of the two phonotypic vowels. The 21 stimuli obtained were tripled. The 63 stimuli were randomized. The computer file with all these sounds was reproduced on a Sony HS60 cassette with a Marantz CP162 portable recorder.

METHODS FOR THE PERCEPTION TEST

a) Medium: The test consists of series of 63 stimuli separated by 3 seconds of silence. For each pair, the test lasts about 3 minutes 15 seconds, that is to say that the global test lasts about 16 minutes. A musical signal is played after each 9 stimuli to facilitate the test's progress. I chose not to propose a number before each stimulus: it could have influenced the subject's choice. On the five answer sheets given to the subjects the responses are presented in blocks of 9.

b) Orders: The subject has imperatively to circle one of the two solutions which are proposed for each stimulus. I make him alive to the fact that he will not necessarily hear the two types of vowels the same number of times. He cannot come back on his choice after he heard the following stimulus.

c) Material: The test is presented to all subjects on a Marantz CP162 portable recorder and Philips SBC3155 headphones.

d) Subjects: The 29 subjects who passed this test are described in [1] (groups A: 5-15 years old; B: 16-25; C: 26-35 and D:36-45, table 1).

Table 1. Number of subjects according to age group and gender

	males	females
group A	5	2
group B	6	2
child génération	11	4
group C	2	3
group D	3	3
parent génération	5	6

QUANTIFICATION

Front's index of the perception for each opposition (F)

A value (N) is given to each stimulus, that is the number of times that it was perceived as the front vowel of the pair ($0 \leq N \leq 3$). Front's index (F) for each pair fit, for each subject, with the sum of the 23 (N). Hence, the bigger is the index, the nearer of the back vowel is the category cut.

Recoding of rough data (F')

In certain subjects, the zone of variability of the perception is very large, and it is difficult to determine the category cut. Then, I considered that 0 and 1 value of (N) corresponded to a back perception, recoded 0, and 2 and 3 values of (N) corresponded to a front perception, recoded 3. So, (F') is the recoded index (F).

RESULTS

Mid/height vowels: global result

The structuralist construct of correlation provides the same treatment of the three vowel oppositions /ø/-/o/, /œ/-/ɔ/ and /y/-/u/ about a possible backing of the articulation. One may suppose that, in the perception test, the three pairs will present similar results. The statistical analysis, carried out on these three pairs considered two by two, shows for the front's index the same treatment for the pairs /ø/-/o/ and /œ/-/ɔ/, and a difference in the treatment for each one of these two oppositions compared with /y/-/u/.

The gender variable

The variationist theory of sound change [7] predicts a difference in the treatment of the dependent variables related to gender. Nevertheless, in this category-perception experiment, this factor is not significant. The front's index is not related to the independent variable 'gender of the subject', neither in the global population nor in each generation of children and parents. In other words, none of the 5 vowel oppositions I studied presents genderly differentiated treatments (figure 1).

The age variable

Most theories on linguistic change appeal to the concept of successive generations to explain or describe changes. The variationist model of change predicts a significant difference in the production of a vowel wich is subjected to an ongoing change between two successive generations. Janson [2] also showed this for perception. In this study, I considered groups A and B as the child generation, and C and D as the parent generation.

The generation factor is statistically significant in term of the front's index (F and F') for /ø/-/o/ and /œ/-/ɔ/ oppositions. More precisely, these dependent variables are higher in the parent generation. In the same way, the space taken up by /ø/ and /œ/ is larger in the parent generation than in the child generation. For /œ/, this difference is significant only with recoded values.

This study brings to the fore a difference in the category cut of the [ø_o] and [œ_ɔ] continuums in two successive generations. The category cut is backer in the oldest (figure 2).

If we look at a finer stratification of age groups - A (5-15), B (16-25), C (26-35) and D (36-45) - the age factor is significant for the /ø/-/o/ and /œ/-/ɔ/ oppositions for front's index. particularly, for this index, we see an increase in the means of ranks for groups A, B and C followed by a decrease in group D, where by the generation factor is linked to a signifiant superiority of the values in the oldest (figure 3). A simple linear regression analysis for this independent variable shows that it can explaine itself, between 37% to 52% of the front index's variation (rough and recoded) when we consider groups A, B and C. These coefficients drop to 19 (minimum) or 31 (maximum) when we add group D in the treatment.

Unlike /ø/-/o/ and /œ/-/ɔ/ oppositions, the /y/-/u/ one presents different features in terms of the statistical significance of the rough values of front index in regard to the recoded ones.

The categorization of the [y_u] continuum is statistically linked to the generation variable only for the rough values. I can add that this generation factor presents the

same categorization features than the two precedent oppositions: a larger area of the front category perception in the parent generation. This factor is not statistically significant for the recoded dependent variables. The recoding of rough data aims to limit the weight of the intra-individual variation. One could set out the hypothesis that the uncertainty area of this categorization is linked to the generation factor. But this is not checked by the statistical analysis. At last, the front index is not linked to the age variable.

The weight of the other independent variables in the category perception

Geographical origin of the subjects, family membership and level of education do not act upon phonemic cut.

Indubitably, individual stategies are implemented in the perception of sound continuums. However, statistically, the individual factor is not significant: the subjects in this study, taken as a whole, do not present any differential behaviour for the rough index (F). Nevertheless, the examination of the category-perception curves - (F) index - of each subject brings to the fore behavioural differences: subject D6, for example, has a much more regular perception than D5. Likewise, the variation area for each of the oppositions can stretch upon few (B8) or many stimuli (B5) [1].

PERCEPTION AND PRODUCTION

This category-perception test confirms the existence of an ongoing change in French in the /Ø/-/O/ opposition, and that, in its two dimensions /ø/-/o/ and /œ/-/ɔ/. Moreover, the results of these tests allow me to present a hypothesis about the production of this opposition's vowels. If the oldest subjects have a backer phonemic cut, one can suppose that the production of the /Ø/ vowel will be backer also. If the eighth stimulus, in the two continuums, is perceived /o/ or /ɔ/ by the children but /ø/ or /œ/ by the parents, it is because the /Ø/ producted by the oldest partly merges with the /O/ producted by the youngest, and vice versa, that the /O/ producted by the youngest partly merges with the /Ø/ producted by the oldest. Indeed, if the phonemic cut is different, the direction of the evolution of change in this opposition is not elucidated by this perception's study. The fact that the youngest present the frontest cut would tip the scale to the tendency described by Martinet [8]. I cannot compare this test with another one realized ten years ago because such a study -based on the same kind of stimuli - does not indeed exist.

ACKNOWLEDGEMENT

This study was supported by a research benefit from the Ministère de la Recherche et de l'Espace, France.

REFERENCES

[1] Malderez, I. (1995) *Contribution à la synchronie dynamique du français contemporain : le cas de voyelles orales arrondies*, unpublished Doctorat, Paris: Paris 7 - Denis Diderot University .

[2] Janson, T. (1986), Sound change in perception, *Experimental phonology*, J.J. Ohala & J.J Jaeger eds., Orlando: Academic Press, 253-260.

[3] Fónagy, I. (1989), Le français change de visage ?, *Revue Romane, 24(2)*, 225-254.

[4] Malderez, I. (1991), "Tendance de neutralisation des oppositions entre voyelles nasales dans la parole des jeunes gens d'Ile-de-France", *12th International Congress of Phonetic Sciences, vol. 2*, Aix-en-Provence; Provence University Press, 174-177.

[5] Bailly, G. & Guerti, M. (1991), "Synthesis-by-rules for French", *12th International Congress of Phonetic Sciences*, , Aix-en-Provence; Provence University Press, 506-509.

[6] Guerti, M. & Bailly, G. (1992), "Synthesis-by-rules using Compost: modelling resonances trajectories", *Eurospeech, 1*, 43-46.

[7] Labov, W. (1992), La transmission des changements linguistiques, *Langages, 108*, 16-33.

[8] Martinet, A. (1958), C'est joli le Mareuc, *Romance Philology, 11*, 345-355. [also in Martinet, 1969, *Le français sans fard*, P.U.F., 191-208].

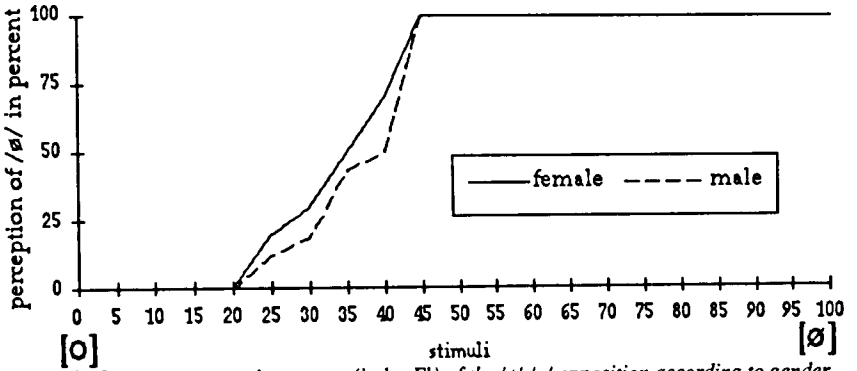


Figure 1: Category-perception curves (index F') of the /ø/-/o/ opposition according to gender

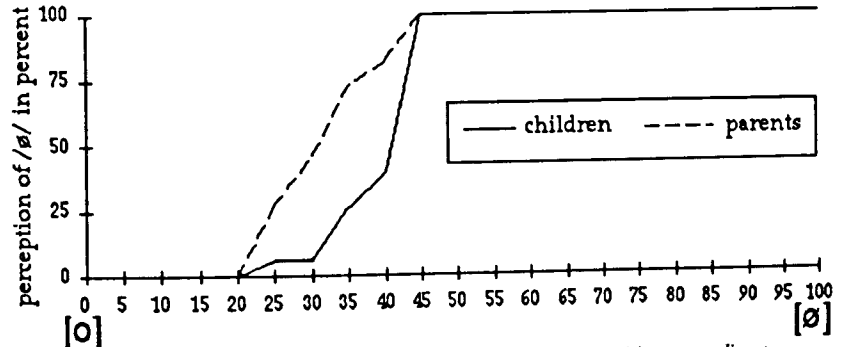


Figure 2 : Category-perception curves (index F') of the /ø/-/o/ opposition according to generation

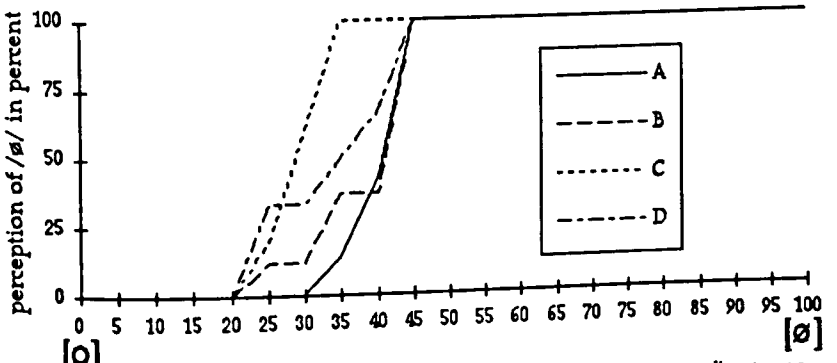


Figure 3 : Category-perception curves (index F') of the /ø/-/o/ opposition according to age group

BACK-CHANNEL SIGNALS IN QUEBEC FRENCH: PHONETIC DESCRIPTION AND FREQUENCY OF USE

Marty Laforest, Julie Nicole and Claude Paradis
C.I.R.A.L., Université Laval, Québec, Canada

ABSTRACT

This paper deals first with the prosodic characteristics of the main channel at the point of insertion of a BC signal. Although no categorical pattern emerges, it is clear that the presence of certain prosodic cues favours the utterance of a back channel signal by the addressee. Based on these results, modified versions, in terms of the proportion of BC signals, of three interview excerpts were submitted to judges. It appears that the best interviews are the ones with a proportion of BC signals in the range of 25% to 50%.

PURPOSE

We call *back-channel* signals (*BC signals* hereafter) all the gestural (smiles, nods of the head, etc.) and the vocal and verbal (*ok, h'm, yeah*, repeats, etc.) signals that convey to a speaker that an addressee is manifestly listening. We claim that these signals are not randomly distributed with regard to speech that is produced on the main channel. We also believe that listening strategies, in the same way as whole sets of conversational strategies, are closely linked to a given culture [1], in particular with regards to the voicing frequency. The production of vocal and verbal BC signals is thus beyond the idiosyncrasies of a given speaker. More specifically, we posit that there is a range of adequate proportions of BC signals in a conversation, without and beyond which the functioning of an interaction may be at risk.

METHOD

From an analysis of nine excerpts of spontaneous discourse extracted from the same number of sociolinguistic interviews carried out in French in Montréal, we first identified the prosodic characteristics of 100 BC signals. Fifty of these were selected quasi-randomly in that the first ten tokens or so of BC signals were analyzed. The remaining 50 tokens were chosen because they met a set of definite criteria; namely, they belonged to the *h'm* family of BC signals; their sound quality was good and they did not overlap, even partially, on the signal produced on the main channel. The quasi-random (QR) and the non-random (NR) subsets of BC signals will be treated separately.

PHONETIC DESCRIPTION

Since BC signals function as marks of acknowledgment and means of supporting and backing up a speaker, it appears reasonable to believe that they are not inserted randomly in the speech chain, but, on the contrary, appear in specific positions. For this study, three prosodic parameters –stress, pause and intonation patterns– were checked for their effect on the utterance of BC signals.

Stress

Table 1 shows that, from the analysis of the 100 BC signals selected, 81 are perceived as following a stressed syllable, whereas 19 are judged as following an unstressed one. It is also worth noting that for the isolated tokens of BC signals in the QR sub-sample, the number of preceding stressed syllables is 32 (91.4%) whereas this figure drops to 8 (53,3%) when they are superimposed on the main channel.

The same table also shows that for the QR sub-sample, the number of isolated signals –35 (70%)– is significantly more important than the number of overlapping ones –15 (30%). The overlapping of the back and main channels in the QR sub-corpus assumes different forms: in 3 cases the speaker completes the preceding sequence even if uttered with a final low tone; in 3 other cases, the speaker resumes

Table 1: Number of BC' signals following stressed or unstressed syllables according to the type of sample and the relative position of the signals.

	Isolated		Superimposed	
	+ stress	- stress	+ stress	- stress
QR	32	3	8	7
NR	41	9	–	–

speaking while the interviewer is uttering a BC signal; 5 times the interviewer either utters a BC signal over the lengthening of a syllable caused by a hesitation, or appears to wait too long to utter its BC signal. The 4 last cases can only be explained by the idiosyncrasies of the interviewers or by the context of the interview.

Pause

Out of the 50 tokens in the QR sub-sample, 40 tokens (80%) are preceded by a silent pause of 100ms or more, whereas for 3 tokens there is virtually no pause: 0, 33 and 59 ms. The remaining 7 BC signals, for which there were no pauses, are the ones that are completely superimposed on the main channel. The average duration of the pause in the QR sub-sample is 369ms (σ245ms). Since the BC signals in the NR sub-sample had to be completely detached from the main channel to be part of the sample, there is a pause for every one of the 50 tokens, the average length of which is 339ms (σ143ms).

Intonation pattern

Using the INTSINT transcription system of intonation [2], for each of the 100 tokens, we identified the pitch movement on the last syllable before a BC signal. Table 2 gives the distribution of the various types of pitch targets in the corpus under analysis. It was determined empirically [3] that a pitch point is higher (H) or lower (L) than the preceding one when there is a variation of 3 semi-tones or more compared to the preceding target. When a variation of less than 3 semi-tones was found, the pitch target was said to be the same (S). For 7 tokens (?), it has been impossible to extract a reliable F_0 contour. It is quite clear from Table 2 that, even if most syllables before a BC signal (55) are Higher than the preceding one(s), there are quite a few tokens that bear a Lower tone than the preceding syllable(s) (24) or the Same tone as the syllable(s) before the signal (14).

Table 2: Number of syllables with a particular tonal configuration according to the type of BC samples (H higher L=lower, S= same, ? indeterminate).

	H	L	S	?	Total
QR	24	11	9	6	50
NR	31	13	5	1	50
Total	55	24	14	7	100

Since there is a change of tone for 79% (55% + 24%) of the syllables next to a BC signal, it is clear that the BC signals tend to be uttered right after syllables that bear a boundary tone or a pitch accent tone.

From these results on BC signals and stress, pause and intonation patterns, we conclude that even if there is no single prosodic cue that may explain the voicing of a BC signal by an addressee, the probability that a BC signal is to be voiced is greater after a pause following a stressed syllable bearing a pitch movement.

PERCEPTION TESTS

However, the number of BC signals found in real speech is always smaller than the number of locations, on the prosodic level where it can be inserted. This is the reason why, based on the results of the prosodic analysis, we built perception tests in order to determine the proportions of BC signals that are acceptable in spontaneous Quebec French speech.

Method

We first built a pretest based on a one minute long excerpt from a sociolinguistic interview, consisting of a brief statement by the interviewer followed by a long reply from the interviewee. All of the possible locations where a BC signal could be inserted in the interviewee's

speech were then identified empirically. Using CSL from *Kay Elemetrics*, the digitized recording of the excerpt was modified by introducing a certain number of *h'm*, the most frequent BC signal. Five versions of the original recording were built, with the proportion of BC signals introduced ranging from 0% to 100% (in 25% increments) of the possible locations. The exact position of these BC signals was randomly determined. Each of the 5 versions obtained was submitted to 6 subjects. After listening to one version of the modified excerpt twice, the subjects were asked to evaluate the significance or the interest of the excerpt by answering on graduated scales (ranging from 1 to 5) 17 questions on the interview itself and on the interviewer and the interviewee.

For the real test, the same procedure was followed, except that two excerpts, whose duration and structure were comparable to those of the pretest, were used instead of one. Contrary to the pretest, where only *h'm*'s were inserted in the excerpt, *oui* «yes» was introduced instead of *h'm* following a frequently attested proportion of one to four [4]. Using two forms instead of one has been found necessary in order to increase the naturalness of the speech sample. The respective positions of the *oui* and the *h'm* forms were also fixed randomly. Finally, in order to determine if the subjects reacted to the quantity of BC signals more than to the mere repetition of the two forms, we built a 6th version from one of the two excerpts. This last version was identical to the most saturated one (100%) of this excerpt, except that, instead of two forms of BC signals, 7 different forms were inserted. Each of the 11 recordings obtained were submitted to 20 to 35 subjects. No group listened to more than one version.

In all three excerpts used, the BC signals inserted in the recorded interaction had been uttered by the interviewer herself at one point or the other in the interview. For each of the 10 modified versions of the original recordings, only one *h'm* and one *oui* were used, which admittedly reduced the naturalness of the modified excerpts. However, for the *improved* saturated 6th version of one of the 2 excerpts, all of the tokens of BC signals inserted were different, except for a few that were less common.

Results

Among the questions asked to the judges, it is evidently those pertaining to the interviewer that are the most relevant for this study. Therefore, only the results for these questions will be discussed. For the pretest, as shown in Figure 1, the versions of the interaction receiving the better scores are the ones that include a BC signal in 25 to 50% of the possible locations.

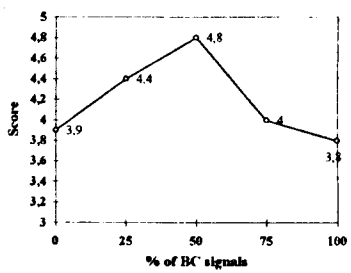


Figure 1: Rating scores by 6 judges for 5 modified versions of an interaction.

By vocalizing his listening to the conversation in these proportions, the interviewer is judged to be more polite, more cooperative, more likeable and brighter. These proportions of BC signals conform to those globally found in all the analyses carried out on other speech samples [1,4]. Conversely, the interviewer is not evaluated favourably when there are no (0%) BC signals or when there is more than 75% saturation.

In spite of the improvements that were thought to have been made on the tests for the larger-scale part of the study, the results that were obtained, as is shown in Figure 2, are less clear than the ones for the pretest. In accordance with what was found in the pretest, the versions with the 25% proportion of BC signals were rated the best, for both excerpts. The most

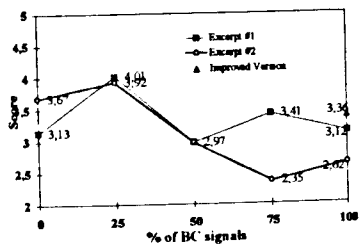


Figure 2: Rating scores by judges for 11 modified versions of two interactions.

negative judgments tend also to involve the extreme proportions. However, for the first excerpt, the score for the 0% version is not far removed from the score for the 25% one, and it goes up slightly from the 75% version to the 100% one.

But it is the results for the second excerpt that are the least consistent with the others. For this excerpt, it is the 50% version, as well as the 0% and 100% versions, that get the worst scores. A careful examination of this version may explain why it is so. The random distribution of the two forms of BC signals and of their respective locations in that version results in a sequence of 8 straight *h'm*'s, seven of which occur one after the other at the end of the excerpt, which seems to have been perceived as quite annoying. In fact, in the last part of this version, it is as if the saturation of the interaction was close to 100%. For the 75% version, the random distribution has introduced a *oui* right in the middle of the *h'm* sequence, which decreases to 5 the number of *h'm*'s that follow each other in quick succession, and, thereby, makes this version more acceptable than the preceding one. Even in the 100% version, the number of consecutive tokens of *h'm* does not exceed 6.

The results obtained for the 6th version of excerpt #2, which contained a maximum number and variety of BC signals, show that the subjects rate it higher than the 100% version containing only repetitions of the same *h'm* and *oui*; however the score for this *improved* version remains lower than the one for the 75% one.

CONCLUSION

Although it is difficult to measure the influence on the subjects of the increasing proximity of the BC signals, which is a result of their increase in number in a short span of time, and of the repetition of a single form, the results from the perception tests indicate that the variation in the number of tokens of BC signals uttered is well perceived, and that the proportions lower than 25% and higher than 75% are considered less favourably. We can predict that a large quantity of BC signals in a short span of time would upset the speaker, to the point of disturbing the communication, to the same extent a silent listening would do. These results thus support the hypothesis that there is a cultural determination of the adequate number of BC signals in a verbal interaction.

ACKNOWLEDGEMENT

The authors wish to thank H. Tétreault for her help with the acoustic analysis, and T. Heisler and A. Manning for revising this text. This research was supported by grants from the *Conseil de recherches en sciences humaines du Canada*.

REFERENCES

[1] Laforest, M. 1994, "Listening Strategies in Sociolinguistic Interviews. Convergence and Divergence", *Culture*, XIV (2):51-61.

[2] Hirst, D. & A. Di Cristo (to be published), "A survey of Intonation systems", in *Intonation Systems. A Survey of Twenty Languages.*

[3] Thibault, L. (1994), *Étude exploratoire du rythme en français québécois*, Unpublished M.A. dissertation, Université Laval, Québec.

[4] Laforest, M. 1992. "L'influence de la loquacité de l'informateur sur la production de signaux *back-channel* par l'intervieweur en situation d'entrevue sociolinguistique", *Language Variation and Change*, 4: 163-177.

CONVERSATIONAL AND PHONOLOGICAL FACTORS GOVERNING THE 'FINAL RELEASE RULE' IN TYNESIDE ENGLISH

G.J. Docherty, P. Foulkes, J. Milroy, L. Milroy and P. Oxley
Department of Speech, University of Newcastle upon Tyne

ABSTRACT

Although glottalisation, including glottal replacement, is phonologically more widely distributed in Tyneside than many other varieties of English, it appears much *less* widely distributed in certain sites. This paper reports findings from a study of stop realisation by 32 Tyneside English speakers. With few exceptions glottals do not occur in pre-pausal or turn-final position. It seems that whilst phonological factors may contribute to a complex structure of constraints on variation, the operation of the 'final release rule' is also dependent on aspects of conversational and/or utterance structure.

INTRODUCTION

Glottal and glottally reinforced stops have wide phonologically-conditioned distributions in Tyneside English. However, previous work on Tyneside and neighbouring Durham English [1, 2] has noted that these forms appear blocked in turn-final and pre-pausal positions. Instead, fully released, non-glottalised stops appear almost categorically in these sites. Thus, although glottal forms may appear sentence internally in words like sheet, bite, [t] is fully released when these items occur before a pause or at the end of a speaker's turn. It seems, then, that a major role is played by conversational and/or utterance constraints, which may perhaps even supersede phonological constraints in governing the operation of the 'final release rule'.

The 'final release rule' (FRR) has been investigated mainly via auditory analysis as part of our current study which focuses on phonological variation and change in contemporary spoken British English. In the near future we intend to supplement the data presented in this paper with detailed analysis of the phonetic correlates of the stops under discussion. Our aim is to combine sociolinguistic research and experimental phonetic analysis with the main goal of assessing the adequacy of different phonological frameworks with respect to these data.

METHOD

Fieldwork in Tyneside has produced recordings of 32 speakers (2 social groups [WC/MC] * 2 sexes [m/f] * 2 age groups [y:16-25/o:45-65] * 4 speakers per cell). Informants were recorded first in a (usually single sex) dyad conversational exchange for around 50 minutes. One young WC female, K, was recorded twice (see Discussion). Informants were then asked to read a word-list constructed to elicit citation forms, including some single items (e.g. sheet, boat), sequences including the same word-final variable (e.g. I beat it, drat it), and disyllabic forms with medial stops (e.g. better, carter).

Analysis of the word-list items was supplemented by examination of similar tokens in the conversational data. Here the aim was to identify 30 tokens per speaker of both pre-pausal and turn-final /t/. This was achievable in most cases in pre-pausal position, but proved more difficult in turn-final position, partly due to the fact that in several cases it was unclear precisely into which category a particular token fell. Generally, token counting was executed after the first 10 minutes or so of the tape had elapsed, in order to ensure that speakers had relaxed into a more natural mode of speech (although where it proved impossible to identify 30 tokens for a speaker, the whole tape was analysed).

RESULTS

Where /t/ appears in medial or intervocalic position, glottalised variants are common, particularly in the speech of males. We do not have space to discuss the phonetic or distributional characteristics of these cases here, but see [3, 4].

As regards items elicited via the word-list, the FRR is applied categorically by 31 of the 32 speakers in monosyllabic /t/-final items. That is, fully released, non-glottalised stops are produced 100% of the time. The exceptional informant is the young WC female K, who produces glottal stops on 2 out of 30 items (print and salt; i.e. 7% failure of the FRR).

Similar but somewhat more complex patterns are exhibited in the conversational data. Tables 1 and 2 present results for this data, showing for each speaker group the number and percentage of glottal variants - i.e. violations of the FRR.

Table 1: conversational data, pre-pausal position - number (N) and percentage of glottal or glottalised tokens

class	*group*	*tokens*	*N*	*%*
WC	OF	120	2	2
	OM	111	2	2
	YF	101	30	30
	YM	120	6	5
MC	OF	120	7	6
	OM	116	2	2
	YF	120	5	4
	YM	120	8	7

In pre-pausal position (Table 1), for 7 of the 8 groups only a few violations of the FRR are found: between 2% and 7% for the groups as a whole, with 11 individual speakers producing 100% fully released stops. In stark contrast to this near categorical pattern, the young WC females (K's group) have 30% glottal tokens. K herself produces glottals in 15 out of 30 tokens (50%).

In turn-final position (Table 2), a comparable distribution is found, although the small number of tokens identified for several groups means that some caution should be exercised when interpreting these figures. No less than 19 speakers deploy the FRR categorically (but in some cases only two or three tokens of turn-final /t/ were identified). It appears, however, that no significant differences exist between the pre-pausal figures and the corresponding turn-final figures. The issue of whether the phenomenon should be considered pre-pausal or turn-final is important in that if it is the latter an interactional explanation should be sought. If, on the other hand, the FRR is triggered pre-pausally, then a linguistic (phonological) explanation is better.

Table 2: conversational data, turn-final position - number (N) and percentage of glottal or glottalised tokens

class	*group*	*tokens*	*N*	*%*
WC	OF	15	0	0
	OM	9	0	0
	YF	36	14	39
	YM	9	1	11
MC	OF	31	2	6
	OM	33	1	3
	YF	70	4	6
	YM	66	3	5

DISCUSSION

The patterns in the Tyneside data can be explained partly in phonological terms, but also require reference to the type of explanation offered by conversational analysis [5].

In careful speech, illustrated by the word-list data, glottal variants in final position are almost categorically

prohibited, whilst the FRR is concomitantly found to apply almost without exception. Contrariwise, analysis of casual speech (represented by the conversational data) shows that violations of the FRR can occur. The exceptions seem overwhelmingly to occur in short vowel items. In long vowel items such as great, meet, the FRR is effectively applied categorically. Amongst the short vowel class, certain items (e.g. that, get, it) occur very frequently. Glottalised forms appear very commonly in these items, such that we might hypothesise that FRR violations are in the main restricted to them. Glottalisation clearly seems to be spreading into pre-pausal and turn-final environments, where it had previously been blocked. It still is blocked in careful speech styles, as well as by and large in the speech of older informants. In traditional phonological terms, then, this process might be well described as operating via lexical diffusion, with frequently occurring items in the vanguard of the change.

The most remarkable difference in FRR application in comparison to other groups is displayed by the young WC females. This is true in both pre-pausal and turn-final context. The high glottalisation scores for this group in the main conform to the lexical patterns already described. However, the conversational behaviour of three of the four speakers in the group, H, K and L, is markedly different from that of the other subjects in the study.

Speaker H (who speaks much less than her partner on the tape) produces glottal variants in 7 of 13 (54%) pre-pausal tokens, and 3 of 8 (38%) turn-final tokens. The glottals tend to occur in common items (got, that, it, out, about) and in items which are clearly turn-final (all right, but - used in Tyneside as a conjunction in the sense of 'though').

Speaker K's behaviour is particularly striking, and gives a strong indication that the FRR is governed principally by conversational constraints. K, as noted, produces 50% glottal forms in pre-pausal position, and 60% (9 of 15 tokens) in turn-final context. Recall, though, that K was recorded twice, first with her brother, and later with a female friend, L. The figures just described, with high use of glottal forms, occur in the conversation with L. However, K's pattern of FRR application in conversation with her brother is comparable to that of other informants: just 4 out of 30 (13%) pre-pausal tokens are glottalised, whilst the FRR applies in all 4 turn-final tokens.

K's violations of the FRR occur overwhelmingly on the sentence tag and that (e.g. you just miss your friends and that). This tag occurs 11 times during the whole tape of K's conversation with her brother, and in 6 of these cases (54%) a glottal form is also used.

Speaker L (K's female dyad partner) produces 6 pre-pausal glottal tokens, 4 of which occur on the tag and that. This tag is much rarer in the speech of other informants, but other tags which terminate in /t/ such as isn't it do appear occasionally to attract glottalised forms, especially in the speech of younger people. For example, the young MC male P produces 4 pre-pausal glottals, 2 of which fall on the tag isn't it. Similarly, another young MC speaker R produces his only pre-pausal violation of the FRR on the tag wasn't it.

The association of FRR violations with tags suggests support for the account in [1]: interactants are oriented to a fully released variant of [t] in a dialect with heavy use of glottals as a signal that a speaker is yielding the floor. In addition to phonetic cues such as fully released [t], grammatical turn-yielding cues such as tags are also available to speakers. Since tags already function as turn-handover cues, this may account for why the phonetic cue often fails to apply in them. In addition, most /t/-final tags involve frequently occurring words such as it and that, which as we have already noted are the items most susceptible in general to attracting glottal variants.

The use of tags has been identified as a feature predominantly of female speech [6]. Our data support this to an extent, with few tags used by males (but also few by the young MC females). Younger WC females use by far the highest number of tags, which partly explains why they have much the highest rate of failure of the FRR.

Thus, violation of the FRR must be accounted for with reference to conversational/utterance constraints, in this particular case identification with sentence tags. It remains to be investigated whether the FRR is *best* explained in terms of conversational structure, or whether e.g. stress and/or intonation patterns play an important role as well. In addition, it should also be noted that the FRR is usually applied before mid-sentence pauses, even when it seems clear that the speaker's turn is not over. Examples include the fact tha[t] # the kids are a lot more streetwise, the da[t]le # was that day. These instances may be regarded as examples of speakers tailing off in mid-sentence and leaving an opening for a turn handover. However, the alternative account that it is the phonological (pre-pausal) context which triggers the FRR must also be borne in mind. If the latter explanation is indeed the better, it may indicate that the constraint on the FRR in general is best viewed as pre-pausal, supporting the suggestion made in [2] rather than that in [1]. We intend to investigate this issue further, although so far we have experienced difficulty in identifying unambiguously whether some particular tokens are pre-pausal or absolutely turn-final.

Our findings have wider implications, particularly with regard to the function of phonological units [7]. Whilst variation in speech sounds has traditionally been regarded as primarily lexical-contrastive in function, what we can clearly see in the case of the FRR is variation being employed in stylistic and demarcative functions. Such variation is certainly systematic, but it is clear that it cannot be governed purely by phonology, given the goals traditionally assumed by phonologists. The relationship between these various constraints and functions has scarcely yet been investigated, but would certainly serve to enhance our understanding of what makes a native speaker a native speaker.

ACKNOWLEDGEMENT

Research supported by the UK Economic and Social Research Council (Grant no. R000 234892 "Phonological variation and change in contemporary spoken British English").

REFERENCES

[1] Local, J.K., Kelly, J. & Wells, W.H.G. (1986) "Towards a phonology of conversation: turntaking in Tyneside." *Journal of Linguistics*, 22, pp. 411-437.

[2] Kerswill, P. (1987) "Levels of linguistic variation in Durham." *Journal of Linguistics*, 23, pp. 25-49.

[3] Docherty, G.J. & Foulkes, P. (1995) "Acoustic profiling of glottal and glottalised variants of English stops." *Proceedings of the ICPhS 1995*.

[4] Milroy, J., Milroy, L. & Hartley S. (1994) "Local and supralocal change in British English: the case of glottalisation." *English World-Wide*, 15, pp 1-32.

[5] Atkinson, M. & Heritage, J. (1984) *The Structures of Social Action*. Cambridge: CUP.

[6] Coates, J. (1986) *Women, Men and Language*. London: Longman.

[7] Ohala, J.J. (1992) "What is the input to the speech production mechanism?" *Speech Communication*, 11, pp. 369-378.

AGE GRADING IN ENGLISH PRONUNCIATION PREFERENCES

J.C. Wells
Department of Phonetics and Linguistics, University College London

ABSTRACT

The results are presented of a survey into speakers' preferences regarding certain words of contentious pronunciation. In some cases sharp age grading was revealed. E.g. for *usage* older people preferred /ˈjuːzɪdʒ/, but the majority /ˈjuːsɪdʒ/. Other sharply age-graded words include *nephew, suit, issue, ate, deity, salt, poor, patriotic, inherent, delirious, applicable, controversy, formidable, harass, kilometre* and *primarily*.

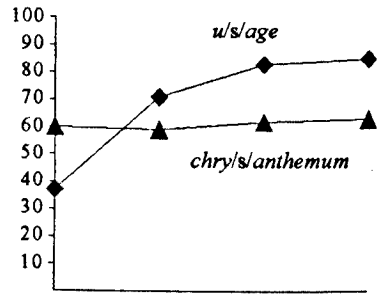
INTRODUCTION

In compiling the *Longman Pronunciation Dictionary* [1] I sought to supplement my own intuitions regarding the prevalence of rival variants by carrying out an opinion poll of speakers' preferences for some ninety words known to be subject to fluctuating or contentious pronunciation. This survey revealed, for instance, that for the word *zebra* 83% of the respondents preferred the /e/ pronunciation, only 17% preferring /iː/. The poll was based on a postal questionnaire submitted to a panel of native speakers of British English (BrE). The respondents numbered 275, and were drawn in equal numbers from the north and the south of England, with small numbers of Welsh and Scots. Most were professionally concerned with speech, being academic phoneticians and linguists, teachers, university students, radio announcers or speech scientists and engineers; but over a quarter were volunteers from the general public recruited by personal contact or by an invitation in a Sunday newspaper. All might therefore be termed 'speech-conscious'.

The polling preferences presented in *LPD* were mostly given as overall percentages. In the present paper I examine their possible correlation with respondents' ages, which ranged from 15 to over 80 years.

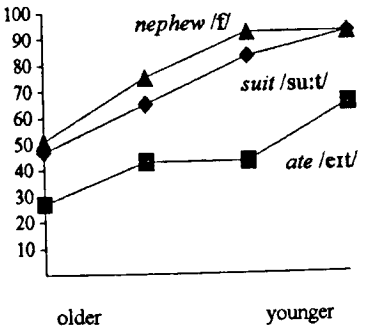
STEADY STATE VS. CHANGE

There are various words which some speakers pronounce with /s/ and some with /z/. *LPD* records, for example, that in *chrysanthemum* 61% of respondents preferred /s/, and in *usage* 72%. This conceals the fact that in *chrysanthemum* the proportion preferring /s/ is virtually unchanged across all age groups, while in *usage* it ranges from 37% among the over-65's to 85% among the under-26's. Hence we infer that in *usage*, but not in *chrysanthemum*, the language is in a state of change, with /s/ increasingly preferred over /z/. It seems that /ˈjuːsɪdʒ/ displaced /ˈjuːzɪdʒ/ as the majority form during the forties, when those born between 1923 and 1947 were growing up.



AMERICANIZATION?

Sharp age grading was also revealed in *nephew, suit,* and *ate*. In each of these words BrE preferences are shown as moving in the direction of the established American (AmE) pronunciation: from /ˈnɛvjuː/ to /ˈnɛfjuː/, /sjuːt/ to /suːt/, and /ɛt/ to /eɪt/.

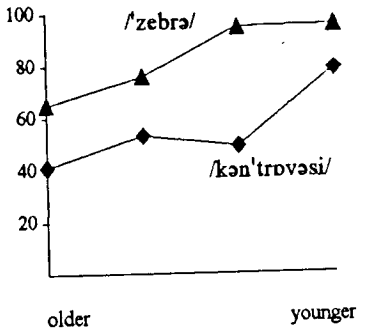


This might lead one to conclude that the most important influence on BrE pronunciation is AmE, and that all or most changes in BrE are to be attributed to American influence. Yet the evidence of other data makes clear that this is not so.

DIVERGENCE FROM AMERICAN

Indeed there are other words in which the opposite trend appears: movement away from AmE. In *zebra* AmE consistently has /iː/, but it was known that in BrE both /iː/ and /e/ are used; *LPD* showed /e/ as preferred by 83% to 17%. When we compare age groups, we see that in BrE /ziːbrə/ is almost exclusively an older people's pronunciation. They young have settled on /ˈzɛbrə/, thus striking away from the AmE norm.

In *controversy* the initial stress pattern universal in AmE is progressively being replaced in BrE by antepenultimate /-ˈtrɒv-/, a British innovation unparalleled in AmE.



INFLUENCE OF PROFESSION

The dip in the graph — the unexpectedly low vote for the new pattern among the younger middle-aged — may be explained by the fact that this group of respondents contained many BBC radio announcers, for whom the stressing /ˈkɒntrəvɜːsi/ is something of a shibboleth. There was a clear correlation between being a BBC announcer and preferring initial stress in this word. It is one of the items on which the BBC Pronunciation Unit has in the past given firm guidance, and announcers have evidently made this preference their own.

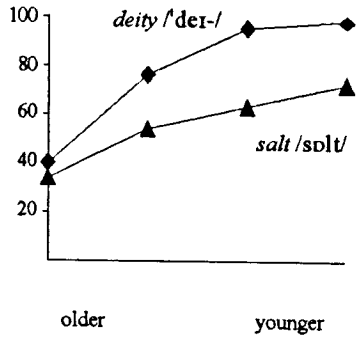
The only other word in the survey to reveal correlation with respondent's profession or occupation was *research*, where a preference for final stress (although common to all groups) is associated particularly with being an academic.

OTHER VOWEL PREFERENCES

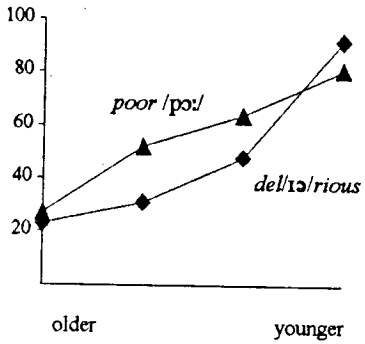
In *deity* the traditional form with /ˈdiː-/ is almost entirely confined to those born before the nineteen fifties, having been displaced, for reasons that are not clear, by /ˈdeɪ-/. Among those born since 1962 preference for the latter form reached 98 percent.

In *salt* it has long been known that there is a form /sɒlt/ in competition with the /sɔːlt/ shown in most dictionaries. It

came as a surprise to me to find that in this word there is fairly sharp age grading, with the proportion preferring the short vowel rising from 34 percent among the oldest group to 72 percent among the youngest.



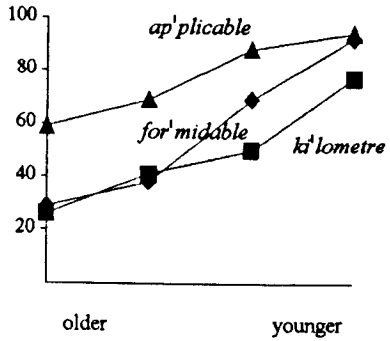
It is well known that in an increasing number of words the diphthong /ʊə/ is being replaced by /ɔ:/, probably through a process of lexical transfer. In *poor* we find that the proportion preferring a pronunciation identical to that of *pour*, i.e. /pɔ:/, rises from about a quarter among the oldest group to over four-fifths among the youngest.



The sharpest age grading found in the survey related to the word *delirious*. The traditional pronunciation has the stressed vowel /ɪ/, as expected on the basis of etymology and orthography. However, for reasons again unknown and apparently peculiar to this word, a new form with /-'lɪər-/ has arisen, and is overwhelmingly preferred by younger respondents.

STRESS PATTERNS

Not only in *controversy* but also in a number of other words of four syllables there is a tendency for initial stress to be supplanted by antepenultimate. This is true, for example, of *applicable, formidable* and *kilometre*. In all of these there is a steady increase in preference for antepenultimate stress across the age groups; and there is no blip caused by the radio announcers, whose views here coincide with those of the other respondents.



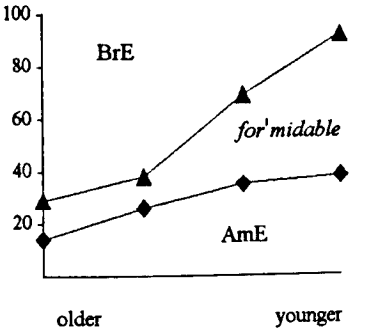
The case of *kilometre* brings us back to the question of possible American influence, since /kɪ'lɑ:məʧə/ is the usual AmE form. As this survey shows, the traditional and logical BrE /'kɪlə,mi:tə/ (cf. *centimetre, millimetre*) is being rapidly displaced by /kɪ'lɒmɪtə/.

AN AMERICAN SURVEY

A comparable survey of AmE pronunciation preferences has recently been carried by my doctoral student Yuko Shitara [2]. Her work reveals, for instance, that in *congratulate* and *February* the forms /kən'grædʒəleɪt/ and /'febjueri/

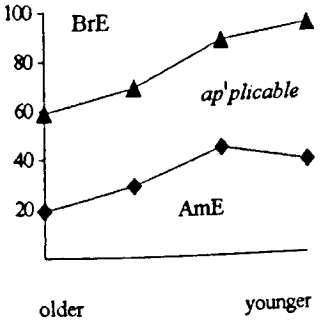
are significantly more favoured by younger than by older Americans; and the younger respondents are significantly more likely to report that *cot* and *caught* are homophonous than their elders.

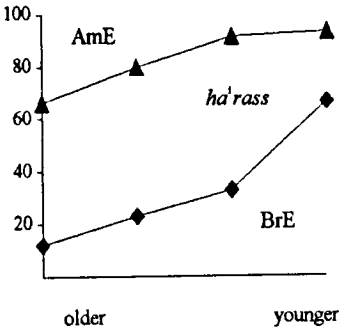
Several stress patterns are changing both in BrE and in AmE. The graph for *formidable* shows BrE to be in the lead in the movement away from initial stress, and to be changing its preference more rapidly.



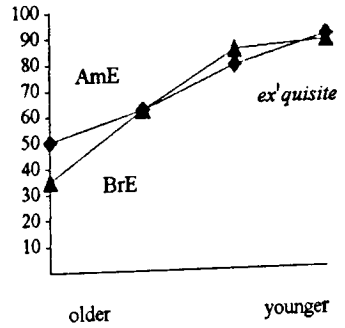
The pattern for *applicable* is similar, though here the change is somewhat further advanced. Here AmE appears to show a recent swing back towards /'æplɪkəbl/.

In *harass* it is the Americans who are in the lead. Their adoption of final stress (at least in some meanings of the word) is increasingly imitated in Britain.





In one word at least, a change in preferred stress placement is taking place at the same time and at about the same rate in AmE as in BrE. This is *exquisite*, where the stress pattern /'ekskwɪzɪt/ is in the later stages of being displaced by the pattern /ɪk'skwɪzɪt/.



REFERENCES

[1] Wells, J.C. (1990),. *Longman Pronunciation Dictionary*. Harlow: Longman.

[2] Shitara, Y. (1993), "Survey of American pronunciation preferences - a preliminary report." *Speech Hearing and Language, Work in Progress*, Phonetics and Linguistics, University College London, 7: 201-232.

OPEN VOWEL BACKING IN CANADIAN ENGLISH

John H. Esling and Henry J. Warkentyne
University of Victoria, Victoria, Canada

ABSTRACT

In Canadian English in Vancouver, the open front vowel /æ/ is acquiring a more retracted quality, beginning with groups with the highest social status, and about one generation earlier for women than for men. The open back vowel /ɑ/ also shows a progression in younger age groups to a preference for a more rounded or retracted [ɒ] variant among men, with women having a slightly higher incidence of [ɒ] vowels in all generations.

INTRODUCTION

Canadian English has ten primarily monophthongal vowels, /i, ɪ, e, ɛ, æ, ɑ, ʌ, o, ʊ, u/, of which [eᴵ, oᵘ, ᵘu] exhibit the most gliding; three diphthongs, /aɪ, aʊ, ɔɪ/, of which the first two have raised allophones [əɪ, əʊ]; and a distinct /ɚ/ with [ɪ, ɛ, a, ɔ, ʊ] also occurring as allophones of the basic set before /r/ [1]. The system's distinguishing characteristic lies in the merger of the two open back vowels /ɔ/ and /ɑ/, which neutralizes oppositions present in other varieties of English, so that 'caught' and 'cot' are homophonous and 'father' and 'bother' also rhyme. The merger is also found in some dialects of American English from New England to the Ohio River [2, 3], and is the likely result of early Scotch-Irish immigration patterns spreading to the southwest as well as north into Canada.

The open front vowel /æ/ functions as in most varieties of English. While American English varieties tend to raise the phonetic quality of /æ/ to [ɛə], [eə], or even as high as [iɛ] in some instances, Canadian English is not reported to participate in that kind of change [3].

Data are drawn from the Survey of Vancouver English, including 240 randomly-selected male and female English speakers native to the region, in three age groups: over 60, 35-60 and 16-34; and four socioeconomic status (SES) categories: middle and upper working class (MWC/UWC), and lower and middle middle class (LMC/MMC) [4].

The findings reported here include auditory evaluations performed for each token of /æ/ and of /ɑ/ for each speaker in the survey, drawn from the survey's reading passage (a conversational narrative with local content) to make the comparison uniform; and consisting of about 50 tokens of each vowel in stressed position for each social/age grouping. Acoustic analysis of these vowels, excluding diphthongs and /ɚ/, began as a sociolinguistic study of long-term voice quality settings [5, 6].

AUDITORY ANALYSIS

Variations in the realization of the open back vowel are grouped into two variants, [ɑ] and [ɒ], unrounded and rounded. The rounded allophones may vary in the degree of rounding or in openness, which ranges between a position just below near-open [ɒ̝] to open [ɒ]. There is also the possibility that the impression of rounding is achieved more by tongue root retraction than by labial protrusion. For a majority of speakers using the unrounded variant, it was judged to be fully back, close to Cardinal Vowel 5. In some cases, the variant [ɑ] was slightly advanced, and only in a few instances was it advanced to a position approaching an open centralized vowel [ɑ̟].

Open back /ɑ/

The distribution of the two variants of /ɒ/ for the Vancouver subjects is presented in Table 1. In the schematic presentation of symbols, where either [ɒ] or [ɑ] appears alone, the frequency of occurrence for that variant is at least 67% for that group. A higher than 33% incidence of a competing variant is marked by a tilde (signifying alternation) and enclosed in parentheses. A still higher alternation of a competing variant (greater than 40%) is indicated by only a tilde, without parentheses. Numerical results are also presented.

The usage of the two variants is divided, with [ɒ] more common at 59%, while [ɑ] occurs 41% of the time. Women have a slightly higher incidence of [ɒ] (60.5%) than do men (57.5%). This can be accounted for by the high frequency of [ɑ] in the over-60 male category (59%). Female subjects show a small increase from oldest to middle-age categories, but then a sharp decrease in [ɒ] usage in the youngest age group. Thus, no clear progressive change is in evidence for Vancouver women, although [ɒ] remains the preferred variant in all generations.

Table 1. Distribution of /ɑ/ variants in the Vancouver Survey -- schematic and numerical representations.

Women Over 60	(ɑ ~)	ɒ	99	157
Women 35-60	(ɑ ~)	ɒ	92	164
Women 16-34	ɑ ~	ɒ	112	144
			(39.5%)	(60.5%)
Men Over 60	ɑ	~ ɒ	148	104
Men 35-60	(ɑ ~)	ɒ	94	161
Men 16-34		ɒ	82	174
			(42.5%)	(57.5%)

Male subjects, on the other hand, show a clear progression, by age group, of an increasing incidence of the [ɒ] variant: old-age, 41%; middle-age, 63%; young-age, 68%. The difference from the female pattern, described in more detail elsewhere [7], suggests that either young women are initiating a reverse trend, or that variation among women is freer and not responding to the same forces for change at this point in the vowel space as with men.

Open front /æ/

The choice of /æ/ variants is between a fronted [æ] and a backed [æ̠]. Choices of a more close or a more open variant occur only occasionally and are not included in the calculation of results. Chi-squared tests with one and three degrees of freedom give a rough indication of the significance attributable to one or another cells of the two-pair and four-pair comparisons. Small differences in the 40-50% range are statistically insignificant.

The distribution of variants of /æ/ is compared and interpreted phonetically for all SES classes combined together in each of the three age groups in Table 2.

Older women's usage favours [æ], except in the MMC group where the backer variant appears to have taken hold. This finding supports the accuracy of spectral analysis of formants which suggests that /æ/ is in fact more backed for MMC women than for MWC women, while most other vowels are more fronted. The middle-aged women demonstrate active variation between [æ] and [æ̠], while the younger women clearly favour [æ̠], especially in the MC.

For the men in the Vancouver survey, the same development -- backing of /æ/ -- appears to be in progress, but lagging behind the women by about one generation. As with the women, the youngest group adopts the retracted variant, with the MC leading the change.

In summary, /æ/ appears to be acquiring a more retracted quality in Vancouver English, beginning with individuals with the highest social status, and about one generation earlier for women than for men. Western Canadian English, to the extent represented by the Vancouver survey, differs from American English in this respect, where the trend often reported in the U.S. is

towards a fronted and more close (front raising) diphthongal variant.

Table 2. Distribution of /æ/ variants in the Vancouver Survey -- schematic and statistical representations.

Women Over 60	æ	(~ æ̝)		
			p < 0.019	
Women 35-60	æ	~ æ̝		p < 0.0001
			p < 0.0001	
Women 16-34		æ̝		
Men Over 60	æ			
Men 35-60	æ			p < 0.0001
			p < 0.0001	
Men 16-34		æ̝		

ACOUSTIC ANALYSIS

Spectral Peak Distributions

Vocalic inventories for each SES by age by gender group are compared using vowel tokens from the identical lexical contexts to those used in the auditory evaluations, taken from the same reading selection. First and second formant frequencies (F_1,F_2) are calculated for 80% of the 50 tokens for each of the ten vowels in the basic vowel system. In the first instance, a linear predictive coding (LPC) routine in the Computerized Speech Lab (CSL) environment is used. F_1 and F_2 are averaged over 20-msec intervals for the duration of the nucleus for each of 40 tokens representing each vowel class (5 tokens x 8 subjects per survey group), with resulting values written to data files for statistical processing. In the second instance, average spectral FFTs are calculated for the vowel nucleus.

Higher second formant frequency ranges for the older women indicate a more fronted quality of /æ/ than for middle-aged women in all groups except the MMC group. Average spectral evidence for the /æ/ vowel indicates a lower F_2 for MMC women than for MWC women in the oldest age group, suggesting that the more retracted [æ̝] variant is accurate. Variation is considerable among middle-aged women, but younger women show a shift towards a retracted target in the MC SES groups.

Spectral peak measurements for the men demonstrate a higher degree of similarity from SES group to SES group than for the women. This corresponds to the relative lack of SES variability in the men's vowels noted in the auditory judgements. Compared to the women, the men switch later but more abruptly in apparent time to the retracted variant.

Significance Measures

Statistical significance of spectral distributions is assessed through group-by-group means-limits comparisons. In the pronunciation of /æ/, middle-aged MMC women are clearly differentiated from all other SES groups of the same age. They are also differentiated from every older group including those of similar social status except the MWC women. Complementarily, the older MMC women use an /æ/ with a quality different from any other group of their same age except the MWC group at the opposite end of the social scale. For the women in general, UWC and MMC SES groups consistently maintain separate qualities of /æ/.

System-shifting Anomalies

A potentially anomalous situation appears in the F_1,F_2 distributions of other vowels in the set. For most of the ten vowels (other than /æ, ʌ, u/), for example, the mean value of F_2 is higher for the older MMC women than the mean value of F_2 for the older MWC women. Since the tokens have been obtained from identical contexts, the F_1,F_2 distribution would imply that the MMC women's vowels are more fronted than the MWC women's vowels. However, this has been shown *not* to be the case for /æ/, i.e., the rule does not apply in the same way. The only contextual variable likely to interfere with the F_1,F_2 locations is the difference in stress or timing with which some subjects may have spoken the target items as they occurred during the reading passage. However, as the items selected are largely in stressed position, and considering the large number of items represented, it is probable that the shift upwards in F_2 for MMC women is not the result of performance anomalies. It is entirely possible that most of the vowels of the system are shifting in one direction for one group relative to another, *except* for certain key vowels which are shifting in the *opposite* direction.

Long-term Spectral Comparisons

In comparing these vocalic results with earlier results of LTAS (long-term spectral averaging) [6], a few parallels are worth noting. The first is that LTAS techniques reveal wider differentiation among female SES groups than among male SES groups of the survey. Secondly, the age and SES distributions of the open vowels and the distribution of long-term settings isolate certain SES groups. MMC women, for example, are consistently differentiated in LTAS from UWC women, and older MMC women show the clearest separation from all other groups except middle-aged MMC women. As with vocalic distributions, older MC women and younger MC women are more distinct in long-term setting than the set of middle-age women, and both move decidedly in favour of a retracted [æ̝].

As a matter of speculation, it may be less accurate to say that the SES groups are "doing" something with their vowels than to say that *we* are *measuring* something that they are doing; for example, we are probably measuring the middle-age women halfway through a change in which vowel quality is "jostling" with voice quality, or shifting its units around to accommodate a new background setting. Clearly, this hypothesis must be subjected to further testing.

ACKNOWLEDGEMENT

This research has been supported by grants 410-87-0334 and 410-89-0191 from the Social Sciences and Humanities Research Council of Canada.

REFERENCES

[1] Esling, J. H. (1994), *University of Victoria Phonetic Database*, version 3.0, Victoria: Speech Technology Research Ltd.

[2] Moulton, W. (1990), "Some vowel systems in American English", in S. Ramsaran (Ed.), *Studies in the pronunciation of English: A commemorative volume in honour of A. C. Gimson*, pp. 119-136, London: Routledge.

[3] Wells, J. C. (1982), *Accents of English*, vol. 3, Cambridge: Cambridge University Press.

[4] Gregg, R. J., Murdoch, M., de Wolf, G. and Hasebe-Ludt, E. (1985), "The Vancouver survey: Analysis and measurement", in H. J. Warkentyne (Ed.), *Papers from the Fifth International Conference on Methods in Dialectology*, pp. 179-200, Victoria: University of Victoria.

[5] Esling, J. H. (1991), "Sociophonetic variation in Vancouver", in J. Cheshire (Ed.), *English around the world: Sociolinguistic perspectives*, pp. 123-133, Cambridge: Cambridge University Press.

[6] Esling, J. H., Harmegnies, B. and Delplancq, V. (1991), "Social distribution of long-term average spectral characteristics in Vancouver English", *Actes du XIIème Congrès International des Sciences Phonétiques*, vol. 2, pp. 182-185, Aix-en-Provence: Université de Provence.

[7] Warkentyne, H. J. and Esling, J. H. (in press), "The low vowels of Vancouver English", in J. Windsor Lewis (Ed.), *Studies in general and English phonetics: Essays in honour of Prof. J. D. O'Connor*, London: Routledge.