AN ACOUSTIC TIMING STUDY OF PHARYNGEAL AND LARYNGEAL FRICATIVES IN ARABIC

A. Djéradi*, P. Perrier & R. Sock

Institut de la Communication Parlée, U.R.A. C.N.R.S. n° 368
INPG/ENSERG Grenoble, France.
***also Institut of Electronics, USTHB, Algiers, Algeria BP n°39.**

ABSTRACT

In this paper a study, on the *acoustic level*, of the temporal control for back fricatives of Arabic is presented. These consonants are examined in different vocalic quality and quantity contexts. Our results show a tendency for a global control of the VCV domain. We thus focus on the *timing* of our fricatives within this temporal span.

1. INTRODUCTION

Knowledge of articulatory coordinations underlying the production of fricatives requires a precise description of the timing of their component *gestures*. This paper deals with the temporal organization of three fricatives in Arabic :

- the pharyngeal fricatives [ʕ] and [ɦ], respectively voiced and unvoiced, produced by a constriction of the low-pharynx ;
- the unvoiced glottal fricative [h], produced by a constriction at the laryngeal level.

The notion of *relative duration* proposed by Lehiste [6] is exploited in this investigation ; so also is the concept of *cycles* or *temporal domains* and *phases* borrowed from the field of motor control and transferred to the study of speech production by Tuller and Kelso [8], among others. On this score, events related to specific articulatory gestures - like onset/offset of vocal fold vibrations or vocal tract closing or opening gestures - are detected on the acoustic signal. Determined temporal coordinations of these events give us our phases and cycles (cf. *infra*).

2. CORPUS

Our corpus is a set of 18 short sentences, each containing a [C1V1C2V2C3V3] item, with C1=[s], V1=[a], C2=[ʕ ɦ h], V2=[a, i, u, a:, i:, u:], C3=[l] and V3 =[a]. The carrier interrogative sentence is of the type [manC1V1C2V2C3V3], for example: [mansaʕala] meaning "Who coughed ?"

3. RECORDINGS

Recordings were carried out in a sound proof room. The speaker, a male Algerian adult, had to repeat each sentence 13 times in front of a directive microphone 'ELECTRET' placed at a distance of 20 cm from his mouth. The signal, digitized by a SONY P.C.M. and sampled at 40 kHz, was finally stored on a BETAMAX videotape. The subject had been instructed to say the sentences in a normal conversational rate, at a regular rhythm, with a slight pause before each sentence. The sentences were presented to the speaker in a random order.

4. MEASUREMENTS

Two vocalic phases, DV1 and DV2, and a consonantal one T were retained. These phases were determined with the help of articulatory-acoustic events proposed by Abry et al. [1]:

- the vocalic phases DV1 and DV2 are specified as the duration between the onset (VVO) and the offset (VVT) of the clear formant structure of the vowels (V1 and V2) that flank the fricative ;
- the consonantal phase T is defined as the duration between the offset of the clear formant structure of vowel V1 and the onset of the clear formant structure of vowel V2.

Phase measurements for vowels V1 and V2 (respectively phase DV1 and DV2) and for the consonantal phase T, are given first in absolute values, and then in relative values with respect to the VCV temporal base.

5. DATA ANALYSES

5.1 Vocalic quantity DV2 (ms)

It is well known that the temporal control of vowel duration in Arabic can be linguistically significant (see for example [3], [4] and [5]). Our results, presented in table 1, confirm this vocalic quantity contrast: a global observation of our data shows that short vowels have a mean duration of 95 ms, and that long vowels have a mean duration of 255 ms with, in both cases, small standard deviations. Vowel duration ratio is thus around 2.5, which is indeed significant. Results also show that vowel lengthening is noticeably influenced by the consonantal context. Finally, it can be observed for the three consonantal contexts, that the most significant vocalic quantity contrast is obtained with vowel [a].

Table.1. Vocalic phases

			[a]	[i]	[u]
[ɦ]	B	m	89	84	87
		σ	6	8	10
	L	m	247	212	241
		σ	20	25	20
[ʕ]	B	m	95	88	95
		σ	20	12	10
	L	m	310	231	264
		σ	30	20	22
[h]	B	m	97	106	109
		σ	13	15	10
	L	m	297	220	-
		σ	30	30	-

5.2 Consonantal phase T

Table 2 shows that average consonantal durations, calculated for the entire data, vary with vocalic quantity contrast: globally, the consonantal duration T has a mean value of 105 ms in short vocalic contexts, whereas the average value for T in long vocalic contexts is 125 ms. This finding seems to support the notion of a preprogrammed attribution of consonantal values with regards to the vocalic linguistic task, short versus long (See [7] for a related discussion on this latter notion). Furthermore, separated analyses of our data for each fricative show comparable influences of the vocalic contexts on the three fricatives.

Table. 2. Consonantal phases (ms)

			[a]	[i]	[u]
[ɦ]	B	m	136	129	119
		σ	10	6	10
	L	m	154	147	133
		σ	15	18	13
[ʕ]	B	m	85	88	66
		σ	15	12	10
	L	m	81	103	87
		σ	10	8	10
[h]	B	m	108	101	91
		σ	13	20	10
	L	m	141	115	-
		σ	20	15	-

5.3 The significant temporal domain

As mentionned above, vocalic quantity contrast is portrayed not only by a difference in intrinsic vowel duration (Table 1), but also by a variation of consonantal durations. Moreover, such durational differences depend on consonant type. We posit therefore that the vocalic quantity contrast is not simply limited to the vocalic phase. To be able to propose hypotheses on the type of sequence which is temporally programmed in this contrast, we looked for the domain that maximizes these differences [2]. We therefore applied the Student test to our data so as to evaluate

Table.3. Student test for absolute durations

	PAR	t		PAR	t
aɦa	VCV	22,6	aʕa	VCV	19,3
	DV2	20,9		DV2	17,8
/	CV	15,8	/	CV	13,3
	T	2,51		T	-
aɦaa	VC	-	aʕaa	VC	-
	DV1	-		DV1	-
aɦi	VCV	13,2	aʕi	VCV	18,6
	DV2	13,7		DV2	17,4
/	CV	9,3	/	CV	13,1
	T	2,99		T	-
aɦii	VC	6,57	aʕii	VC	-
	DV1	-		DV1	-
aɦu	VCV	20,1	aʕu	VCV	29,2
	DV2	18,7		DV2	20,3
/	CV	20,2	/	CV	13,8
	T	2,78		T	2,98
aɦuu	VC	-	aʕuu	VC	-
	DV1	-		DV1	-
aha	VCV	16,6	ahu	VCV	11,8
	DV2	15,6		DV2	9,03
/	CV	10,3	/	CV	8,1
	T	3,61		T	2,98
ahaa	VC	-	ahuu	VC	-
	DV1	-		DV1	-

the distinctive power of the phases described above and also that of combinations of these phases: VC=DV1+T, CV=T+DV2, and VCV=DV1+T+DV2. Only significant values of t ($a<0.05$) are presented in table 3. In general, the VCV domain provides the most significant temporal base for class distinctions; one might therefore think, in the absence of more data on speech rate variation, that the VCV span is the programmed entity for these specific linguistic tasks: actually, vowel phonological differences seem to spread out significantly to the entire VCV sequence.

5.4. Influence of the vocalic context.

Let us now take a look at the influence of vocalic contexts [i], [a], [u], on the fricatives within the VCV domain.

In the case of short vowels, the total duration of the cycle is not affected by vocalic variations; however, consonantal durations show significant differences:

-for both [ʕ] and [ɦ], we observe significant differences in [a] vs. [u] contexts, and in [u] vs. [i] contexts, but not in the [a] vs. [i] ones.

-for [h] we observe only significant differences in [a] vs. [u] surroundings.

Therefore, in the case of short vowels, place of articulation does not seem to have much influence on consonantal duration, but vowel rounding seems to induce modifications in this duration.

As concerns long vowels, it can be noticed that vocalic length (V2) varies, depending on both place of articulation and lip shape characteristics. From a general point of view, all component phases of the VCV cycle are modified as a function of vowel type. However, the total duration of the complete cycle remains more or less stable.

What can be observed therefore, is a temporal restructuring of phases within the VCV cycle for each vowel class; however, the control of the total duration of this cycle evokes an isochronous constraint principle. These results corroborate the hypothesis addressed above regarding the temporal programming of the VCV sequence as whole (cf. *supra*).

Average values for VCV domains are different for the three fricatives and for short versus long vowels:

- [ɦ]=210 ms for short vowel vs. 360 ms for long vowels;
- [ʕ]=180 ms for short vowel vs. 330 ms for long vowels ;
- [h]=200 ms for short vowel vs. 400 ms for long vowels.

But one must be cautious in generalizing such results concerning this temporal restructurings, as long as speech rate has not been explicitly introduced in to our experimental paradigm.

6. TIMING OF THE FRICATIVES IN THE VCV DOMAIN

How do constituent phases of the VCV domain help to distinguish the different fricative classes in relation to this domain ? We observed these relative differences for the voiced/unvoiced contrast [ʕ] vs. [ɦ] and for differences in place of articulation [ɦ] vs. [h] in the various vocalic contexts. Figures 2a and 2b give the structural types of the VCV sequences for each context.

Phase DV1 discriminates the voiced vs. unvoiced classes [ʕ] vs. [ɦ], but does not distinguish the difference in place of articulation for the unvoiced [ɦ]/[h].

Phase T discriminates the voiced vs. unvoiced classes ([ʕ]/[ɦ]), but the distinction associated with place of articulation observed in short vowel contexts disappears completely with vowel lengthening.

Phase DV2 is weaker than phases DV1 and T in discriminating consonantal classes, for voiced/unvoiced contrasts. When DV2 does exist as a distinctive parameter in opposing place of articulation, its t values are comparable with those obtained for phase T, and better than those obtained for phase DV1.

7. CONCLUSION

In the study of the acoustic timing of fricatives in Arabic, the analysis of absolute durations shows a global control of the VCV temporal base. Within this cycle, the voiced vs. unvoiced distinctions are made essentially by a temporal reorganization of the VC domain, which corresponds to the combination of phases DV1 and T.

The distinction of place of articulation is obtained generally by a restructuring of of T and DV2 corresponding to the CV span.

These preliminary results must be consolidated by a study that includes speech rate as the controlled perturbing factor of vocalic quantity and consonantal types.

REFERENCES

[1] ABRY, C., BENOIT, C., BOË, L. J, & SOCK, R. (1985), "Un choix d'événements pour l'organisation temporelle du signal de parole," *Proceedings of the 14th J.E.P.*, GCP-GALF, 111-137.

[2] ABRY, C., ORLIAGUET, J.P., & SOCK, R. (1990), "Pattern of speech phasing," *European bulletin of cognitive psychology, 10, 3*, 269-288.

[3] AL-ANI, S. (1970), *"An acoustic and physiological investigation,"* The Hague, Paris.

[4] GHAZALI, S. (1987), "Elements of Arabic phonetics," *Applied Arabic Linguistics and Signal & Information Processing*, R. DESCOUT (ed), 51-58.

[5] JOMAA, M. & ABRY, C. (1988), *"La résistivité de la quantité vocalique aux variations de la vitesse d'élocution : le cas de l'arabe tunisien," Proceedings. of the 17th J.E.P.*, G.C.P.-GALF, 231-236.

[6] LEHISTE, I. (1970), *"Suprasegmentals,"* M.I.T. Press, Cambridge Mass.

[7]SOCK, R., (1983), *"L'organisation temporelle de l'opposition de la quantité vocalique en Wolof de Gambie "* Doctorat de 3ième cycle, Université des Langues et Lettres de Grenoble, France.

[8]TULLER, B. & KELSO, J.A.S. (1984), "The timing of articulatory gestures: evidence for relational invariants", *J. Acoust. Soc. Am., 76*, 1534-1543.

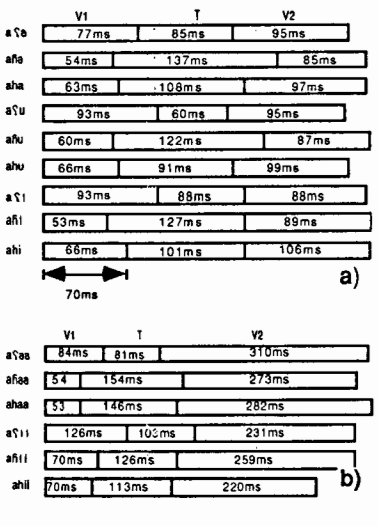


Fig.1.Absolute durations
a) short vowels
b) long vowels

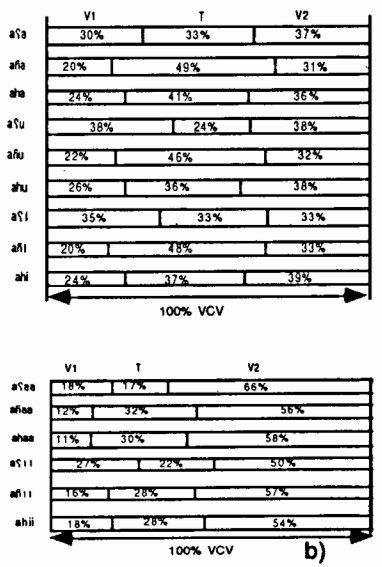


Fig.2 Relative durations in VCV domain
a) short vowels
b) long vowels

VELAR MOVEMENT DURING PRODUCTION OF NASAL AND NON-NASAL VOWELS IN SOUTH MIN DIALECT OF CHINESE

S. Horiguchi[1], R. Iwata[2], S. Niimi[3] and H. Hirose[3]

1) Dept. of Neurotology, Tokyo Met. Neurological Hospital, JAPAN
2) Dept. of Humanities, Shizuoka University, JAPAN
3) R.I.L.P., Faculty of Medicine, University of Tokyo, JAPAN

ABSTRACT

The movement pattern of the velum during the production of words contain nasal and non-nasal vowels in the South Min dialects of Chinese was investigated.

Result showed that when a non-nasal vowel preceded nasals, the velum started falling short after the release for the initial consonant /p/ occurred. This suggests that anticipatory coarticulation for nasals strongly effects the entire part of the vowel in the South Min dialect. Difference which seemed to depend upon the quality and duration of vowels were also observed in the movement pattern of velum. The nature of tones might modify the trajectory of the velar movement.

1. INTRODUCTION

There is a distinction between nasal and non-nasal series of vowels in the South Min dialect of Chinese. Question will arise; whether the domain of nasalization is only the main vowel or the entire part of the vowel in diphthongs of nasal series; whether the vowel in CVN type syllables, like [pan] or [pin], is nasalized or not.

Another question is whether there is any relationship between the nature of tones or quality of vowels and the movement pattern of the velum or not.

2. METHOD

2.1. Subject

The subject was a 31 year old male speaker of the South Min dialect of Chinese from Taiwan.

2.2. Speech Material

An experiment was conducted to investigate whether there was the difference in the movement pattern of velum during the production of nasal and non-nasal vowels in the South Min dialect of Chinese. The list of test words are shown in Table 1. Test words consisted of CV(V)(N), where the initial consonant C was /p/ and V(V)(N) were /ã/, /ĩ/, /iã/, /uã/, /an/, /in/ and /ian/. Some of these words have two different nature of tones for different meanings. These test words were embedded in a carrier sentence [tse^{33} ʃi^{21} ___ dzi^{33}] (This is a character for ___.) and pronounced five times for each by a native speaker of the South Min dialect as naturally as possible.

Table 1 Test Words

[pĩ^{55}]	[pĩ^{21}]
[piã^{55}]	[piã^{21}]
[pin^{55}]	
[pan^{55}]	
[pien^{55}]	[pien^{21}]
[pã^{55}]	
	[puã^{21}]

The numbers following phonetic description correspond to the nature of tones for each test word. For example, [pĩ^{55}] was pronounced in the 1st tone "high level" and [pĩ^{21}] was pronounced in the 4th tone "short low" in the South Min dialects of Chinese. All words in this list are meaning words. However, only [pã^{55}] does not have a corresponding Chinese character, because it is an onomatopoetic word for the sound of horn of automobile.

2.3. Data Recording

Velar movement was monitored and recorded using the Velotrace simultaneously with the acoustic signal. The Velotrace, which had been previously reported, was inserted through the subject's nose with its internal lever resting on the nasal surface of velum so that the movement of the velum could be monitored from outside[1]. The maximum excursion of the velar movement was confirmed using the non-speech gestures such as forced nasal inhalation and dry swallowing. The movement of velum, which was tracked by the internal lever, was converted to the analog electrical signal using an electro-magnetic rotation sensor (see Fig.1). The Velotrace signal was digitized 100 samples per second, and the acoustic signal was digitized at 5000 samples per second. These signals were then stored on a microcomputer for further analysis.

2.4. Data Analysis

Following parameters were extracted from the data (see Fig.2).

1) the acoustical duration of the test word (p-dz)
2) the duration of the downward movement of velum (bV-eV)
3) the interval between the acoustical release of /p/ and the beginning of the downward movement of velum (p-bV)
4) the interval between the acoustical release of /p/ and the end of the downward movement of velum (p-eV)
5) the height of velum at the beginning of the downward movement (hV)
6) the height of velum at the end of the downward movement. (lV)
7) the excursion of the movement of velum (hV-lV)
8) the speed of the downward movement of velum (sV)

3. RESULTS

A part of the results is as follows.

3.1. Single Vowel versus Diphthong

There were no significant differences between [pĩ^{55}] and [piã^{55}], between [pin^{55}] and [pien^{55}] and between [pĩ^{21}] and [piã^{21}] in the p-bV. The p-bV was the second shortest in [piã^{55}]. (The shortest was in [puã^{21}].)

3.2. High Vowel versus Low Vowel

There were significant differences between [pĩ^{55}] and [pã^{55}] in the hV and the lV. The trajectories of the velar movement were always higher in case of [pĩ^{55}] than [pã^{55}]. There was same tendency in the case compared [pin^{55}] with [pan^{55}], however, statistically significant difference was only seen in the lV. There was no significant difference in the hV-lV and the sV for either pairs.

3.3. Influence from Difference of Tones

Although very few parameters were statistically significant, systematic difference was observed in the pairs compare the different natures of tones. For example, the hV's were higher in the cases of [pĩ^{55}], [piã^{55}] and [pien^{55}] than in the cases of [pi^{21}], [pia^{21}] and [pien^{21}]. The sV's were faster in the cases of [pĩ^{55}], [piã^{55}] and [pien^{55}] than in the cases of [pĩ^{21}], [piã^{21}] and [pien^{21}].

The bV-eV's were longer in the cases of [pĩ^{55}],[piã^{55}] and [pien^{55}] than in the cases of [pĩ^{21}], [piã^{21}] and [pien^{21}].

4. DISCUSSION

Result showed that when a non-nasal vowel preceded a nasal consonant, the velum started falling short after the release for the initial consonant /p/ occurred. This suggests that anticipatory coarticulation for nasals effects the entire part of the vowel in the South Min dialect. Differences which seemed to depend upon the quality and duration of vowels were also observed upon the movement pattern of velum. In the case of test word contained low vowel, the trajectory of the velar movement was higher than in the case of word contained high vowel. This suggests that there might be some physical (either aerodynamic or mechanical) interaction between tongue and velum. Also, differences which were observed in the pairs compare the different natures of tones suggest the possible interaction between the position of larynx and velum via the root of tongue.

5. REFERENCE

[1] HORIGUCHI, S. & BELL-BERTI, F. (1987), "The Velotrace: A Device for Monitoring Velar Position", Cleft Palate Journal, 24, 104-111.

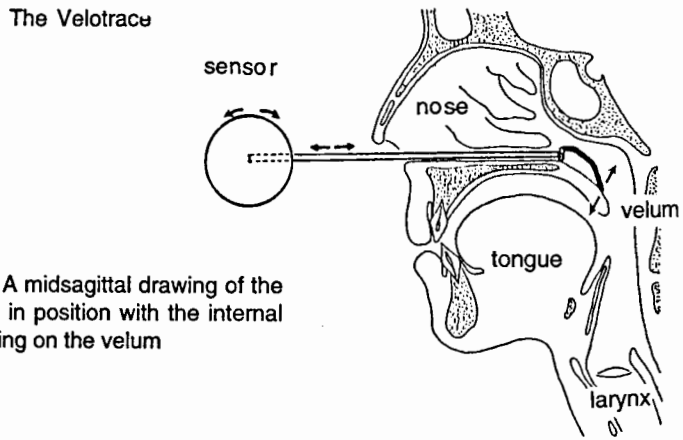


Figure 1 A midsagittal drawing of the Velotrace in position with the internal lever resting on the velum

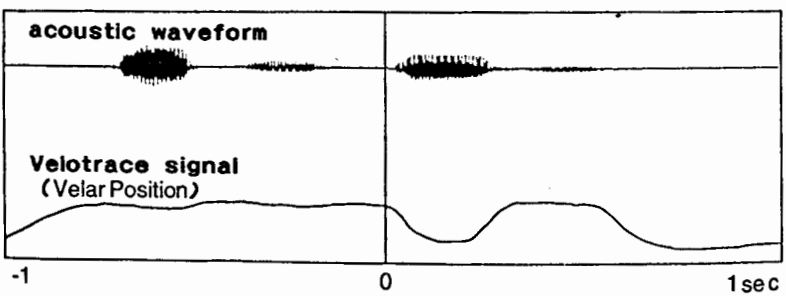


Figure 2 The time function for one token is shown with the acoustic waveform at the top and the Velotrace signal below. Zero on the abscissa indicates the reference point which the release for the initial /p/ occurred.

LARYNX CLOSED QUOTIENT MEASURES FOR THE FEMALE SINGING VOICE

David Howard[1], Geoff Lindsey[2] and Sarah Palmer[3]

(1) Signal Processing: Voice and Hearing Research Group; Electronics Department, York University, Heslington, York YO1 5DD, UK.
(2) Linguistics Department, Edinburgh University, Adam Ferguson Building, George Square, Edinburgh EH8 9LL, UK.
(3) Phonetics and Linguistics Department, University College London, Wolfson House, 4 Stephenson Way, London NW1 2HE, UK.

ABSTRACT

Electrolaryngographic techniques previously used to quantify larynx closed quotient (CQ) change with singing training/experience for adult males are further applied to a group of 21 adult female singers. This data suggests that there is a change in the *patterning* of CQ variation with fundamental frequency which correlates with the number of years singing training/experience.

1. INTRODUCTION

Professional singers often report that when their voices are working well less voice productive effort is required. Acoustic explanations have been offered for this effect, particularly the presence of the *singers' formant* (e.g. [7] and [8]). Electrolaryngographic larynx measures suggest ([4], [5]) that for adult male singers there is a statistically significant positive correlation between the number of years singing training/experience and larynx closed quotient (CQ) -- the percentage of each vocal fold cycle for which the folds remain in contact. This is acoustically plausible since high CQ values mean reduced loss of energy to the sub-glottal cavities. This paper presents CQ data for adult female singers, and discusses it in terms of its variation with singing training/experience. Possible applications of this work include the development of new visual displays for use in research and training of both the singing and speaking voice.

2. SUBJECTS AND DATA

Twenty-one adult female singers (F1 - F21) took part in the experiment. Each subject completed a questionnaire relating to her singing training/ experience and other musical skills, as well as environmental, dietary and general health factors which she felt could affect her voice. In this paper, the subjects have been ordered by the number of years singing training/experience, which is summarised as follows:

F1-F5 at least 5 years formal training, extensive choral and solo experience;
F6-F8 less than 5 years formal training, some choral experience
F9-F10 minimal formal training, some choral experience;
F11-F15 no formal training, some choral experience;
F16-F21 no formal training, no choral experience.

Stereo digital audio tape (DAT) recordings were made in a sound-isolated room at University College London. The speech pressure waveform from an electret microphone was recorded on one channel and the output from the electrolaryngograph (Lx) on the other [1]. The recorded data consisted of:

1) a read prose passage lasting approximately two minutes, and

2) a two octave sung major scale, ascending and descending from G (196 Hz) on the vowel of *palm* with each note lasting approximately a third of a second. (Some of the less experienced singers were unable to sing a two octave G major scale accurately and they were encouraged to produce notes across as wide a range as possible.)

3. DATA ANALYSIS

The Lx was analysed on a MASSCOMP 5600 computer system, using the *Speech Filing System* [6], to give a scattergram for each subject's sung scale. This scattergram, referred to as Qx [1], shows the distribution of CQ values against the logarithm of fundamental frequency (F0).

The technique of cycle-by-cycle CQ measurement is described in [3]. Lx (polarised to ensure increased inter-electrode current flow is represented as a vertical deflection) is time-differentiated and the positive peaks are used to mark the start of the closed phase. The closed phase ends when the negative-going Lx waveform crosses a fixed ratio (7:3) of the current cycle's amplitude. The time between the start of the closed phase in one cycle and the next gives the fundamental period for that cycle. CQ is thus given by: ((closed phase) / (fundamental period) * 100) %.

Figure 1 shows Qx plots for the scales sung by all subjects (F1-F20), ordered by the number of years' singing training/experience.

4. DISCUSSION AND CONCLUSIONS

The Qx plots for the sung scales by our adult female subjects (see figure 1) tend to have CQ values which are confined within a narrow range for a given sung note, but vary as the pitch is altered. The Qx data for our 18 male subjects on the other hand [5], also based on a sung two octave ascending and descending G major scale, tended to exhibit comparatively constant CQ values with fundamental frequency, with a statistically significant correlation between their mean CQ value and the number of years' training/experience.

A survey of the data presented in the figure reveals some patterning in the variation of CQ with F0. Subjects with more training/experience tend to exhibit an upward change in CQ with rising F0 (e.g. F1, F2, F5) whilst those with minimal or no formal training exhibit a downward trend in CQ with rising F0 (e.g. F10-F17 and F18-F21). Amongst these data are subjects who exhibit a mixture of rising and falling CQ with F0, some in a 'V' shape (e.g. F9, F16, F17), some with an inverted 'V' (e.g. F3, F4) and some with more than one change in CQ trend with F0 (e.g. F19). Subjects F11 and F13 have a general downward trend in CQ with rising F0, but they both show some much higher CQ values in their upper F0 range. This suggests a change from a downward CQ trend with rising F0 towards a 'V' shape.

Singing ability can be developed with singing training and to a lesser extent with singing experience, but the number of years' singing training/experience cannot *in itself* be used to quantify a singer's ability, excluding as it does considerations such as those of natural talent. It does, however, provide a useful indicator. Singing ability can be viewed as position along a developmental continuum (c.f. [9] for children's singing), from those with an inability to 'sing a note in tune' at one end, to those acknowledged as having mastered the art of singing. Singing training and singing experience are just two aspects of singing development which encourage and enable movement along the continuum.

When our adult female Qx data is considered in terms of this continuum, the following trend in the Qx patterning can be observed. Singers towards the 'untrained' end of the continuum exhibit a definite falling of CQ values as F0 increases, while those towards the other end have rising CQ values as F0 increases. Thus the *tilt* of CQ values with respect to F0 appears to give some measure of position along the developmental continuum for our adult

female singers. Qx scattergrams for singers with limited training/experience are indicative of a development from a downward tilt towards an upward tilt. Thus there appear 'V' and inverted 'V' shaped scattergrams as CQ in only certain parts of their vocal range is being increased at that time; these may relate to register breakpoints. (Sundberg [8] discusses 'chest', 'middle' and 'head' registers for the female voice.)

One goal of singing training is to 'cover' the register breakpoints to make them imperceptible. It is suggested ([8] and [2]) that more experienced singers keep the larynx in a lowered position to cover the tone around the breakpoints. Untrained singers tend to raise the larynx in order to attain higher pitches, in many cases producing a 'strained' sound. The smoothly rising Qx patterns exhibited by our most trained singers could be the result of keeping the larynx lowered and the pharynx open to help cover the breakpoints. The Qx patterns exhibited by our least trained singers could result from inappropriate larynx/pharynx usage, the 'V' and inverted 'V' shapes being considered as intermediate snapshots along the developmental continuum. For these singers the voice is beginning to work efficiently (rising CQ with F0) over some of the F0 range and elsewhere it is not (falling CQ with F0). The turning points in the Qx pattern thus represent voice register breakpoints. Clearly a true longitudinal study of these effects is a desirable next step in this research.

5. ACKNOWLEDGEMENTS

The authors would like to thank all subjects who took part in the recordings. This work was part supported by SERC research grant number GR/F/30642

6. REFERENCES

[1] ABBERTON, E.R.M., HOWARD, D.M., and FOURCIN, A.J. (1989). 'Laryngographic assessment of normal voice: A tutorial', *Clinical Linguistics and Phonetics, 3,* 281-296.

[2] BUNCH, M.A., and SONNINEN, A. (1977). "Some further observations on covered and open voice qualities", *National Association of Teachers of Singing Bulletin,* 26-30.

[3] DAVIES, P., LINDSEY, G.A., FULLER, H., and FOURCIN, A.J. (1986). 'Variation in glottal open and closed phase for speakers of English', *Proc. Institute of Acoustics, 8,* 539-546.

[4] HOWARD, D.M., and LINDSEY, G.A. (1987), 'New laryngograms of the singing voice', *Proc. 11th International Congress of Phonetic Sciences,5,* USSR: Tallinn, 166-169.

[5] HOWARD, D.M., LINDSEY, G.A., and ALLEN, B. (1990). 'Towards the quantification of vocal efficiency ', *J. Voice, 4,* 205-212

[6] HUCKVALE, M.A., BROOKES, D.M., DWORKIN, L.T., JOHNSON, M.E., PEARCE, D.J., and WHITAKER, L. (1987). "The SPAR Speech Filing System", *Proc. Eur. Conf. on Speech Technology, 1,* 305-317.

[7] LINDSEY, G.A. and HOWARD, D.M. (1989), "Spectral features of renowned tenors in CD recordings", *Proc. Speech Research-89,* Budapest, 17-20.

[8] SUNDBERG, J. (1987), "*The science of the singing voice* ", Northern Illinois University Press, Dekalb, Illinois.

[9] WELCH, G.F. (1986). "A developmental view of children's singing", *British Journal of Music Education. 3,* 295-303.

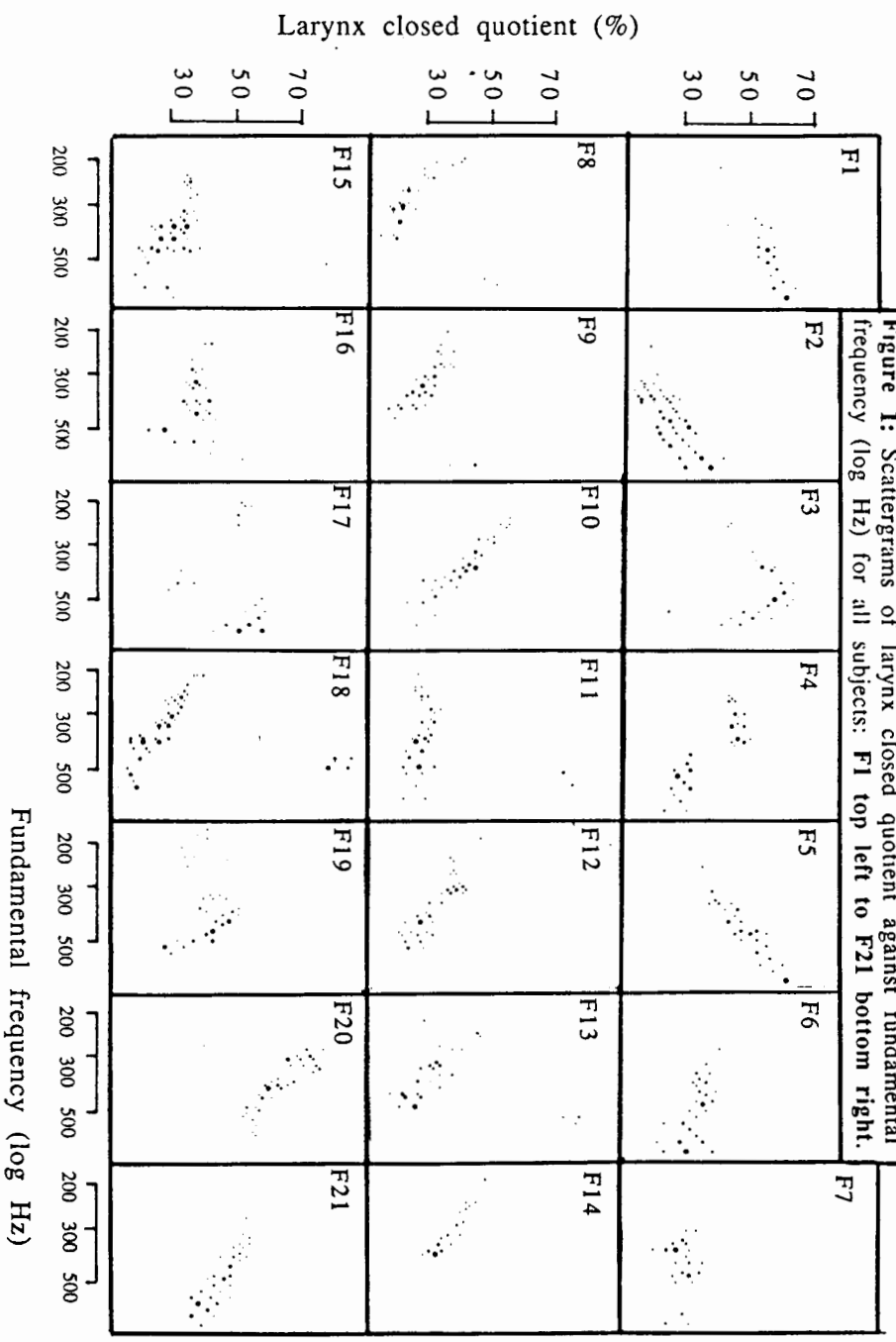


Figure 1: Scattergrams of larynx closed quotient against fundamental frequency (log Hz) for all subjects: **F1** top left to **F21** bottom right.

A PHONETIC STUDY OF OVERTONE SINGING

Gerrit Bloothooft, Eldrid Bringmann, Marieke van Cappellen, Jolanda B. van Luipen, and Koen P. Thomassen

Research Institute for Language and Speech, University of Utrecht Trans 10, 3512 JK Utrecht, The Netherlands

ABSTRACT

We describe the phenomenon of overtone singing in terms of the classical theory of speech production. The overtone sound stems from the second formant or a combination of both the second and third formants, as the result of careful, rounded articulation from /ɔ/, via schwa /ə/ to /y/ and /i/. Strong nasalisation provides, at least for the lower overtones, an acoustic separation between the second and first formants, and can also reduce the amplitude of the first formant. The bandwidth of the overtone peak is remarkably small and suggests a firm and relatively long closure of the glottis during overtone phonation. Perception experiments showed that listeners categorize the overtone sounds differently from normally sung vowels.

1. INTRODUCTION

Overtone singing is a special type of voice production resulting in a very pronounced, high and separate tone which can be heard over a more or less constant base sound. The technique is rarely used in Western music but in Asia (especially Mongolia and Tibet) it is more common and overtone singing can be heard during secular and religious festivities. The high tone follows a characteristic musical scale [for instance, for pitch C3 (130.8 Hz) (- and + indicate a deviation from the exact tone): **C3**, **C4**, G4, **C5**, E5-, G5, A5+, **C6**, D6, E6-, F6+, G6, G#6+, A6+, B6-, **C7**,...], from which it can be concluded that one really hears an overtone of the fundamental.

The literature contains only a few reports on overtone singing [1,5,7,8], which indicate both the importance of formants and register type. In this paper we present both an acoustic analysis of overtone singing and a study to evaluate the perception of the overtone sounds, in relation to normally sung vowels.

2. MATERIAL

We have recorded series of sung overtones from a singer with many years of experience in overtone singing, both as a performer and as a teacher. In this paper we describe the results for an F_0 value of 138 Hz (C#3). In addition, 12 Dutch vowels /a/, /ɑ/, /ɔ/, /o/, /ɛ/, /e/, /ɪ/, /i/, /œ/, /ø/, /u/, and /y/, sung in a normal way at the same F_0, were recorded.

3. ACOUSTIC ANALYSIS

The recordings were digitized at a rate of 10 kHz and stored in a computer. From the middle, stable, part of each recording 300 ms was segmented. Average power spectra were obtained from FFT analyses (1024 points, shift 6.4 ms) over this segment. Formant frequencies were computed on the basis of appropriate LPC or ARMA analysis.

3.1. FFT-Spectra

Figure 1 shows the average FFT spectra of all overtone recordings. Despite the averaging procedure, the width of each individual harmonic is limited, indicating the stability of F_0 over the interval (standard deviation of F_0 was less than 0.1 semitone in all cases). It can be seen from the shifting peak in the spectra that overtone singing seems interpretable as a special use of a formant. Obviously, the singer tries to match a formant with the intended overtone frequency and succeeds very well.

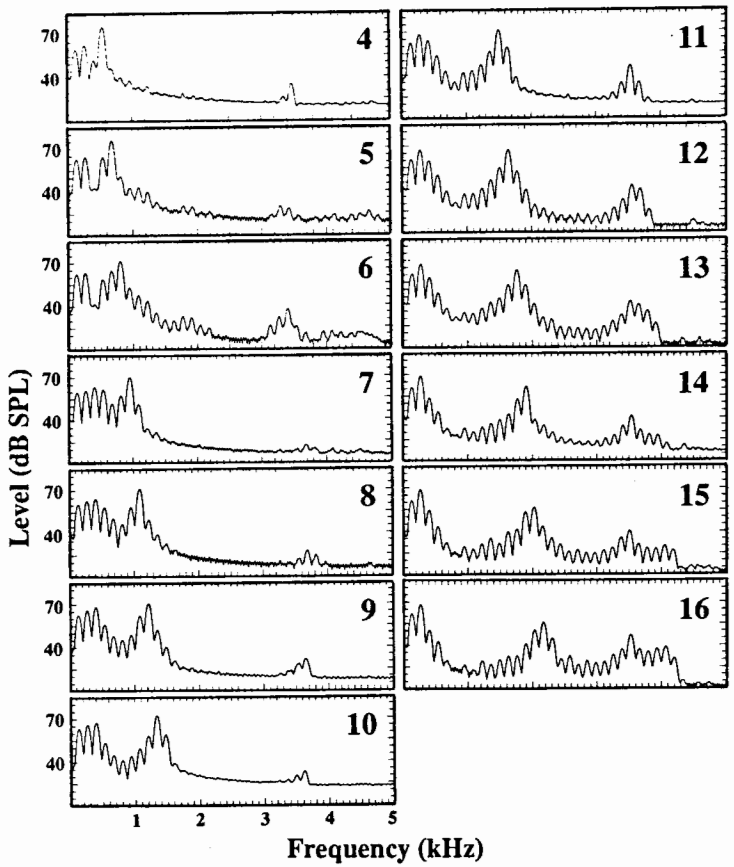


FIG. 1. Average FFT spectra for overtone sounds, sung at F_0 = 138 Hz (C#3). The overtone sounds are numbered according to the main partial involved.

3.2. Formant frequency analysis

In Fig. 2 we present formant frequency results for both the overtone sounds and the sung vowels in the F_1 - F_2 plane. The figure shows two modes in the production: firstly, the overtone sounds 4-6 around /u/, and secondly, the track from /ɔ/ to /i/.

In the first mode, it can be seen from the FFT-spectra that there is energy absorbtion around 400 Hz, indicating a strong nasalisation. The characteristic overtone sound resides in the second formant, as others [1,8] had already suggested. The bandwidth of the second formant is very narrow and, especially for the lower overtones, seldom exceeds 40 Hz. This indicates little acoustic damping in production: firm glottal closure and small losses in the vocal tract. All these characteristics indicate a low, rounded, nasalised, back vowel /u/ or /ɔ/ (low F_1 and F_2, a nasal pole/zero pair, and suppressed F_3 [3]).

The second mode in the production of an overtone sound, applies for overtone frequencies higher than 800 Hz. The main peak of the spectrum still rises in tune with the intended overtone frequency and is interpreted as a combination of F_2 and F_3. It may be of interest that the singer explains this series of overtones with the articulatory variation

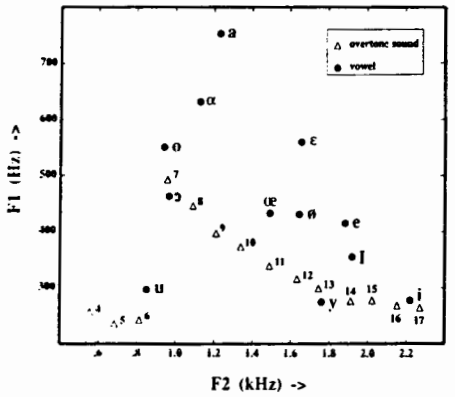


FIG. 2. F_1 - F_2 plane for stimuli sung at F_o = 138 Hz, with positions of the vowels (IPA symbols) and overtone sounds (represented by the number of the corresponding partial).

during the word 'worry'. It is known, already from the Peterson and Barney data, that in a retroflex /r/ the F_3 frequency can be remarkably low and can approach the F_2 frequency. This has also been mentioned by Stevens (1989), especially in combination with liprounding, while Sundberg (1987) mentioned the effect as the acoustic result of a larger cavity directly behind the front teeth.

For the higher overtone sounds, the articulation comes near /y/ and /i/, where continued lip rounding makes it possible to bring F_2 and F_3 together [4], although for the highest overtones a subtle lip spread may be needed to reduce the front cavity to a minimum.

3.3. The glottal factor

The very narrow bandwidth of the "overtone formant" suggests a good and long glottal closure. We believe that the singer used modal register, with a relatively long glottal closure, originating from a firm glottal adduction. This hypothesis does not exclude that performers may use the vocal fry register as well [7]. In all cases, the long glottal closure requires a strong adduction of the vocal folds, which could easily result in general muscular hypertension in the pharyngeal region. This may relate to the prominent role of the buccal cavity, suggested by Hai (1991).

3.4. Intensity analysis

Up to an overtone frequency of 1.5 kHz, the overtone harmonic has a stable relative intensity of -10 dB relative to overall SPL, and dominates the spectrum. For higher frequencies, the relative level of the overtone harmonic sharply drops with a slope of about -18 dB/octave.

4. THE PERCEPTION OF OVERTONE SINGING

4.1. Material, listening experiment, and analysis

As stimuli we used the combined set of 14 overtone sounds and 12 Dutch vowels. From these stimuli we used the same segment (300 ms) as had been used for the acoustical analyses, but we shaped the first and final 25 ms sinusoidally to avoid the perception of clicks. In a computer-controlled experiment, these stimuli were judged by fifteen listeners on ten 7-point bipolar semantic scales. Further details of semantic scales will be presented in a forthcoming paper. The judgements were analyzed by means of multidimensional preference analysis MDPREF [2]. In the technique of MDPREF a stimulus space is constructed in which distance corresponds to perceptual (dis)similarity.

4.2. The perceptual stimulus space

The plane of the first two dimensions of the stimulus space is shown in Fig. 3. 41 % of the total variation in the judgements was

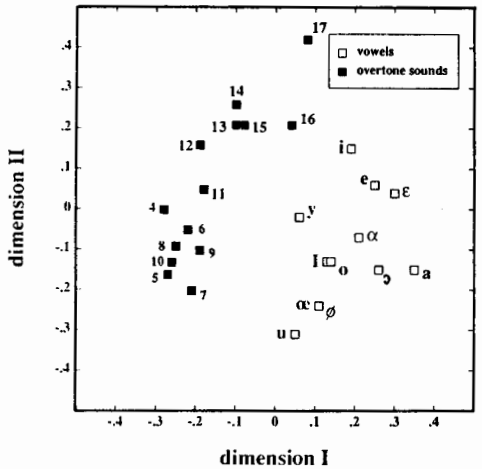


FIG. 3. The perceptual stimulus space. The overtone sounds are given by the number of their corresponding partial, the vowels by their IPA symbol.

explained in this plane, while higher dimensions each explained less than 6.3 %.

The overtone sounds and normally sung vowels are perceptually separated clusters. The vowels are situated roughly in a triangle, with the cardinal vowels /i/, /u/, and /a/ at the angles. The overtone sounds are roughly ordered according to their harmonic number, although the stimuli numbered from 4 to 10 can be described as a cluster. This probably relates to the constant relative energy of the overtone harmonic for this set. The direction of the overtone sounds is, from the lower to the higher numbers, about the same as from /u/ to /i/, as may be expected from the relation between harmonic numbers and F_2 frequency values.

4.3. A physical description of the perceptual stimulus space

We attempted to match the perceptual stimulus space with multidimensional physical descriptions of the stimuli [formant frequency space (see Fig. 2), 1/3-octave bandfilter energy space both by means of the Plomp metric and the Klatt metric [2,6]]. These attempts were not successful (low correlations between coordinate values along dimensions) because of the division into two clusters of the stimulus space, for which these metrics do not present an explanation.

Some additional perceptual sensitivity to the very small bandwidth of the "overtone formant", which clearly physically separates overtone sounds and normally sung vowels, seems necessary to explain the results.

6. REFERENCES

[1] BARNETT, B.M. (1977), "Aspects of vocal multiphonics", *Interface 6*, 117-149.
[2] BLOOTHOOFT, G. and PLOMP, R. (1988), "The timbre of sung vowels", *JASA 84*, 847-860.
[3] FANT, G. (1960), " *Acoustic theory of speech production*" The Hague: Mouton.
[4] FUJIMURA, O., and LINDQUIST, J. (1970), "Sweep-tone measurements of vocal tract characteristics", *JASA 49*, 541-558.
[5] HAI, T.Q. (1991), "New experiments about the Overtone Singing Style", *Proc. Conference 'New ways of the voice'*, Becançon, 61.
[6] KLATT, D.H. (1982), "Prediction of perceived phonetic distance from critical-band spectra: a first step", *Proc. ICASSP*, Paris, 1278-1281.
[7] LARGE, J. and MURRY, T. (1981), "Observations on the nature of Tibetan chant", *J. of Exp. Research in Singing 5*, 22-28.
[8] SMITH, H., STEVENS, K.N., and TOMLINSON, R.S. (1967), "On an unusual mode of chanting by certain tibetan lamas", *JASA 41*, 1262-1264.
[9] STEVENS, K.N. (1989), "On the quantal nature of speech", *J. of Phonetics 17*, 3-45.
[10] SUNDBERG, J. (1987), *"The science of the singing voice"*, Dekalb: Northern Illinois University Press.

SOME CROSS LANGUAGE ASPECTS OF CO-ARTICULATION

Robert McAllister and Olle Engstrand

Institute of Linguistics Stockholm, Sweden

ABSTRACT

The work reported in this paper concerns som temporal aspects of vowel dynamics in English, French and Swedish. The language specific auditory effects of dynamic complexity and direction of tongue movement are starting points for a study of VCV sequences in these three languages using dynamic electropalatography. Tongue movement is compared between the three languages. Results support the assumtion that differences in auditory impressions of vowels in Swedish and English are dependent on differences in the timing of similar articulatory events whereas French seems to employ quite different articulatory strategies.

1. Introduction

This paper is a brief progress report on research activities in connection with the ACCOR project (Articulatory-Acoustic Correlations in Coarticulatory Processes: A Cross-Language Investigation) which is part of ESPRIT's Basic Research Action program. The work being reported on here is focused on articulatory dynamics in VCV utterances and, in particular, vowel dynamics in these sequences. In many dialects of English, high vowels such as /i/ and /u/ are heard to glide from a somewhat centralized towards a more cardinal vowel quality. The corresponding Central Swedish vowels tend to display a more complex dynamic behavior with a final offglide from cardinal to centralized. In French, on the other hand, these vowel colors sound essentially constant. These language specific, auditory effects are quite characteristic. From a cross-linguistic point of view, these dynamic patterns tend to typify a phonetic typology based on two continuous dimensions: 1) dynamic complexity (monophthongal, diphthongal, triphtongal, ...), and 2) direction of movement (offgliding, ongliding). Among the languages mentioned above, French would approximate the dynamically less complex type, whereas English and Swedish would approximate the dynamically more complex type; and English would approximate the ongliding type, whereas Swedish would approximate the offgliding type.

From a motor control point of view, it is of some interest to explore the articulatory means employed to bring about these effects. It might be assumed, in particular, that differences in perceived vowel dynamics between some languages (perhaps English and Swedish) are brought about essentially by means of different relative timing of onsets and offsets of parallel activity in the articulatory and phonatory subsystems, whereas the activity pattern in each particular subsystem varies less between the languages; other languages (perhaps French) might employ a different articu-

latory scheme altogether. In this paper, we present some preliminary electropalatographic (EPG) data relevant to this question.

2. METHODS

We used the EPG system available at Reading to record a set of vowel-consonant-vowel (VCV) utterances, forming all possible combinations of V=/i,a/ and C=/p,b/, spoken by an Australian English, a French, and a Swedish speaker. The English and Swedish vowels belonged to the set of tense vowels; the French vowel inventory has no tense vs. lax distinction. Randomly ordered lists of these combinations were read several times by each speaker.

3. RESULTS

We will limit this report to some results on the sequence /ipi/ as produced by the three speakers. Figure 1 shows number of activated electrodes (out of a total of 62) at various points in time for English and Swedish; from left to right: a) acoustic onset of V1, b) maximum number of activated electrodes during V1, c) acoustic offset of V1, d) minimum number of activated electrodes (for English = acoustic /p/ release, e) (only Swedish) acoustic /p/ release, f) maximum number of activated electrodes during V2, g) acoustic offset of V2. For French, where no clear maxima or minima could be discerned, the triangles correspond to a) acoustic onset of V1, b) acoustic offset of V1, c) acoustic /p/ release, d) acoustic offset of V2. Acoustic segments corresponding to /i/1, /p/ and /i/2 are indicated at the bottom of the figure for each subject. The data represent averages of 5 repetitions of the test utterance. The Swedish data are shown by filled squares, the English data by filled circles, and the French data by triangles. These symbols are connected by straight lines. The data are aligned to the point in time where there is a minimum number of active electrodes for all three subjects. This point also corresponds to the /p/ release for the Australian English and the French subject. When the data are synchronized in this way, the similarity be-

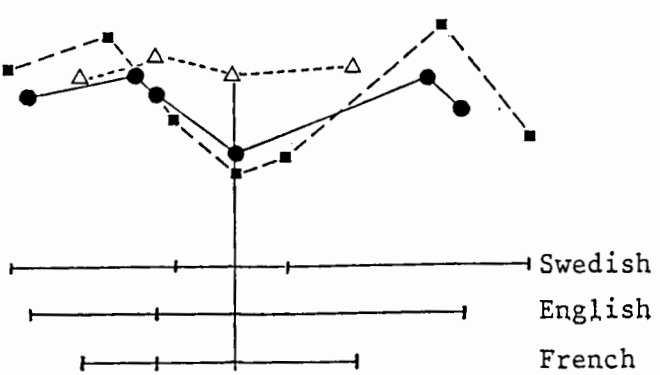


Figure 1. Number of activated EPG electrodes at different points in time during the production of the utterence /ipi/ by an Australian English (circles), a Swedish (squares) and a French speaker (triangles). Below: segment boundaries between the vocalic portion of /i/1, and /i/2.

tween the overall English and Swedish contours, and the difference between these and the French contour, are evident. In particular, the English and Swedish data both display a deep "trough" in the electrode activation pattern, corresponding to a relaxation of the tongue position roughly coinciding with the consonant; the tendency to such a trough in the French pattern is to weak to be statistically significant.

There is, however, a clear difference between the English and the Swedish contours. In the Swedish contour, most of the vowel offglides fall within the vocalic segments, whereas they mostly fall outside the vocalic segments in the English contour. In other words, the troughs in the respective EPG pattern are differently timed relative to the acoustic segment boundaries; the minimum number of activated electrodes occurs at the middle of the consonant segment in the Swedish subject, and at the C/V2 boundary in the Australian-English subject. These differences are thus due to a different relative timing between the tongue articulation underlying the EPG activation patterns and the parallel labial and glottal activities.

4. DISCUSSION

In summary, this limited data set supports the assumption that the difference in perceived vowel dynamics between English and Swedish can be primarily brought about by means of different relative timing of onsets and offsets of activity in the articulatory and phonatory subsystems, whereas French seems to employ a quite different articulatory scheme. In French, the auditory impression of a constant, non-dynamic vowel quality seems to correspond to a constant articulatory position throughout the /ipi/ sequence. This also shows that the presence of a trough in a VCV sequence is language specific rather than universal [4], and that its timing relative to acoustic boundaries is related to characteristic dynamic properties of vowels in the respective languages. A further factor possibly contributing to the presence of troughs in vowel-symmetrical utterances in English and Swedish is related to conditions on aspiration as discussed in [3] and [1] [2]. In particular, the aerodynamic requirements on the production of the stressed, aspirated /p/ release would include a relatively wide vocal tract (cf. [5]), a condition met when the high vowel position is temporarily relaxed. In French, where voiceless stops are not aspirated, or considerably less aspirated, this adjustment would not be necessary.

REFERENCES

[1] Engstrand, O. 1988. Articulatory correlates of stress and speaking rate in Swedish VCV utterances. *JASA* 83, 5 1863—1875

[2] Engstrand, O. 1989. "Towards an electropalatographic specification of consonant articulation in Swedish." *Phonetic Experimental Research, Institute of Linguistics, University of Stockholm (PERILUS)* X, 115—156.

[3] McAllister, R. 1978. "Temporal assymmetry in labial coarticulation." *Papers from the Institute of Linguistics, University of Stockholm (PILUS)* 35.

[4] Perkell, J. 1986. "Coarticulation strategies: preliminary implications of a detailed analysis of lower lip protrusion movements." *Speech Communication* 5, 47—68.

[5] Stevens, K.N. 1971. "Airflow and turbulence noise for fricative and stop consonants: static considerations." *Journal of the Acoustical Society of America,* 50, 1180—1192.

ACKNOWLEDGMENTS

This work was supported by ESPRIT/BRA and by The Swedish Board for Technical Development (STU). We are grateful to our ACCOR partners Fiona Gibbon, Katerina Nicolaidis and Bill Hardcastle for helping us carrying out the above recording at the phonetics lab in Reading.

EFFET DE CONTEXTE INTER-LETTRE SUR LE DÉROULEMENT TEMPOREL DES MOUVEMENTS D'ÉCRITURE : SIMILARITÉS AVEC LA PAROLE

L-J. Boë[1] J-P. Orliaguet[2] & R. Belhaj[1]

[1] Institut de la Communication Parlée URA CNRS n° 368
Université Stendhal, Grenoble, France.
[2] Laboratoire de Psychologie Expérimentale URA CNRS n° 665
Université Mendès France, Grenoble, France.

ABSTRACT

This research aims to analyze the spatial context effects on the timing of handwriting. Results show that the duration of phases of a given letter is influenced by the size and the direction of rotation of the following letter. This finding suggests, in the same manner as for speech, the existence of anticipatory processings.

INTRODUCTION

La parole et l'écriture sont deux activités à finalité sémiotique mobilisant chacune différents niveaux de traitement de l'information.

Les recherches issues notamment de la neuropsychologie et de la psychologie cognitive [1, 2 et 6] ont montré, qu'entre le traitement sémantique et les sorties motrices, intervenaient plusieurs modules de contrôle, organisés hiérarchiquement, ayant pour fonction d'analyser, de transformer et de transmettre sériellement l'information d'un niveau à un autre. Par ailleurs, le déroulement continu des séquences motrices suggère une activation en parallèle de ces différents niveaux, ce processus permettant, en cours de mouvement, une préparation anticipée des séquences motrices restant à réaliser [5]).

Ce fonctionnement modulaire, à la fois sériel et parallèle, a fait l'objet de nombreuses investigations expérimentales certaines d'entre elles plus particulièrement centrées sur les phénomènes d'anticipation motrices observés avec des contextes inter-graphème ou inter-phonème différents [3, 4].

C'est ainsi par exemple que les courbures, la taille et la durée de l'écriture d'une même lettre sont modifiées en fonction de la position de cette lettre dans le mot [7] et de la forme de la lettre suivante [4].

Les travaux de Perkell [3] mettent en évidence des phénomènes similaires. Les phases composant le mouvement de protusion des lèvres destiné à produire une voyelle arrondie [u] varient en fonction de la partie consonantique précédente (CCV ou CV). La durée de la première phase ("phase lente") augmente dans la situation CCV (*look-ahead model*)par contre la deuxième phase se produit à date fixe (*time-locked model*) par rapport au début acoustique de la voyelle, ce qui conduit Perkell à proposer un modèle composite (*hybrid model*).

Notre recherche sur l'écriture se situe dans le cadre de cette discussion théorique. Elle vise à analyser dans le cas d'une coproduction de deux lettres (ℓ ℓ, ℓ e, ℓ n) les effets du changement de taille des lettres (ℓ ℓ – ℓ e) et du changement du sens de rotation des mouvements (ℓ e–ℓ n) sur le déroulement temporel des différentes phases composant la première lettre (ℓ).

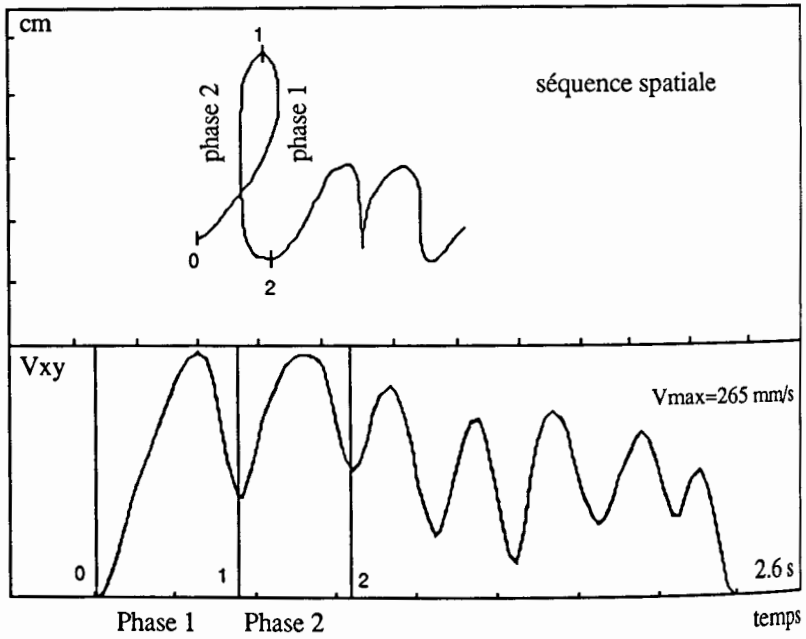


Fig. 1 Phases temporelles du ℓ déterminées à partir des minima de vitesse (exemple du couple ℓ n).

2. EXPÉRIENCE

2.1 Sujets

L'échantillon est constitué de 5 sujets adultes tous droitiers manuels et âgés en moyenne de 24 ans.

2.2 Matériel

Les mouvements d'écriture sont réalisés sur une tablette graphique Numonics 2202 pour la saisie des coordonnées orthogonales de la position d'un stylo électronique (fréquence d'échantillonnage 200 Hz, précision, 0.2 mm).

2.3 Procédure expérimentale

La tâche a consisté à écrire en cursive les paires de lettres "ℓ ℓ ", "ℓ e" et "ℓ n". Les sujets devaient reproduire à vitesse normale les couples de lettres en essayant de suivre le plus précisément possible un modèle transcrit sur la tablette graphique et correspondant à un enregistrement préalable de leur écriture naturelle. La taille des lettres était de 1 cm pour le "e" et le "n" et de 2.5 cm pour le "ℓ ". Le démarrage et l'arrêt du mouvement avaient lieu sur la tablette Chaque couple de lettres a été reproduit 40 fois. L'ordre de passation des séries a été contrebalancé d'un sujet à un autre.

2.4 Recueil des données.

Un programme informatique permet le calcul et le tracé du profil de vitesse (lissé par fonction spline) en correspondance avec le tracé xy (figure 1). Pour chaque couple de lettres on a déterminé sur le profil de vitesse, à partir des limites données par les minima de vitesse, la durée des deux phases temporelles du "ℓ ". La phase 1 correspond au déplacement "ascendant" du stylo (*up-stroke*) et la phase 2 au déplacement "descendant" (*down-stroke*) (cf. Fig. 1). Pour le couple "ℓ ℓ" le relevé des durées a été effectué sur le premier "ℓ ". Sur les 40 essais enregistrés seuls ont été retenus les 20 essais présentant une reproduction précise du modèle d'écriture.

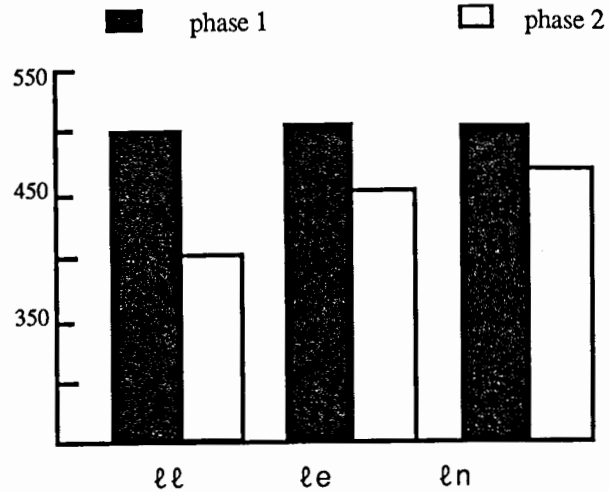


Fig. 2 Durée des phases 1 et 2 (en ms) du ℓ en fonctions de la nature de la lettre suivante : ℓ (identique), e (différence de taille), n (différence de taille et de sens de rotation).

RESULTATS

On a calculé pour chaque couple de lettres et pour chaque sujet la durée moyenne des deux phases du "ℓ ". Les données ont été traitées globalement par analyse de la variance et par le t de Student pour les comparaisons par paires. Les résultats apparaissent sur la figure 2. On constate une relative stabilité de la durée de la première phase du "ℓ " ($F(2,4) = 1.04$ non significatif à .05). Par contre la durée de la deuxième phase augmente en fonction des caractéristiques de la lettre suivante ($F(2,4) = 17.34$ $p < .01$). Cette augmentation estt moins importante pour "ℓ e" c'est à dire lorsque le changement porte seulement sur la taille que pour "ℓ n" qui présente à la fois changement de taille et de rotation.

CONCLUSION

Comme pour la parole [3] on constate que les contraintes contextuelles ont un effet différencié sur le déroulement temporel des phases des mouvements d'écriture.

La stabilité temporelle de la première phase du "ℓ " semble indiquer qu'il s'agit d'une phase préprogrammée du mouvement. Le caractère prédictif du déplacement effectué au cours de cette phase et le faible temps dont dispose le contrôle rétroactif pour traiter et ajuster la trajectoire suggèrent l'existenced'un contrôle proactif de cette partie du mouvement.

Par contre l'augmentation de la durée de la deuxième phase lors du changement de taille et/ou du changement de sens de rotation semble correspondre à une charge attentionnelle dont le système moteur est dispensé lorsqu'il s'agit de reproduire deux lettres identiques. Cette augmentation de temps n'est pas due à un contrôle sensoriel de la réalisation de cette phase qui ne présente aucune difficulté supplémentaire par rapport à celle rencontrée en "ℓ ℓ ". Elle parait plutôt relever d'une anticipation destinée à ajuster la préparation de la lettre suivante.celle-ci étant d'autant plus coûteuse pour le système moteur que le nombre de paramètres spatiaux à traiter est élevé (taille ou taille et sens de rotation).

RÉFÉRENCES

[1] ELLIS, A.W. & YOUNG, A.W. (1988). *Human cognitive neuropsychology*. London : Lawrence Erlbaum.

[2] MARGOLIN, D.I.(1984). The neuropsychology of writing and spelling : semantic, phonological, motor and perceptual processes. *Quatrely Journal of Experimental Psychology*, 36, 459-489.

[3] PERKELL, J.S.(1990). Testing theories of speech production : implications of some detailed analyses of variable articulatory data. In W.J. HARDCASTLE & A., MARCHAL (Eds), *Speech production and speeech modelling*, London : Kluwer Academic Publishers.

[4] THOMASSEN, A.J.W.M. & SCHOMAKER, L.R.B. (1986). Between-letter context effects In handwriting trajectories. In H.S.R. KAO, G.P., VAN GALEN. & R., HOOSAIN (Eds), *Graphonomics : contemporary research in handwriting*. Amsterdam : North-Holland.

[5] VAN GALEN ,G.P., MEULENBROECK, R.G.J. & HYLKEMA, H (1986). On the simultaneous processing of words, letters and strokes in handwriting : evidence for a mixed linear and parallel model. In H.S.R. KAO, G.P., VAN GALEN. & R, HOOSAIN (Eds), *Graphonomics : comtemporary research in handwriting*. Amsterdam : North-Holland.

[6] VAN GALEN, G.P. (1990). Phonological and motoric demands in handwriting : evidence for discrete transmission of information. *Acta Psychologica*, 74, 259-275.

[7] WING, A.M., NIMMO-SMITH, M.I. & ELDRIDGE, M.A. (1983). The consistency of cursive letter formation as a function of position in the word. *Acta Psychologica*, 54, 197-204.

STUDIES OF SOME PHONETIC CHARACTERISTICS OF SPEECH ON STAGE

Gunilla Thunberg

Dept of Linguistics, University of Stockholm, Sweden

ABSTRACT

In order to investigate the special techniques for voice production used by actors on stage, recordings have been made of two actresses. Fundamental frequency distribution analysis has shown: firstly, that there are different individual speech production strategies used by different actors; secondly, that a common feature appears to be the conscious exploitation of variability in fundamental frequency in stage speech styles; thirdly, that narrowing the range of F0 distribution appears to be a useful technique for creating intimacy, e.g. in order to give an illusion of whispered speech on stage.

1. INTRODUCTION

Speech as it is produced by actors on stage, before a larger audience and without amplification, requires an extra level of distinctiveness in order to carry a message all the way to the back row of the theatre. This cannot be done simply by raising your voice because that would unable the actor to work with speech nuances and sensitivity in an optimal manner. The actor must use special techniques for speech production, that must not be detected by the audience. These must also allow a range of expression necessary for creating an illusion of real life.

The purpose of the work reported here is to investigate some of the phonetic aspects of speech on stage as compared to other normal speech styles. Two different modes of stage speech have been studied: that of the normal tone of voice as in ordinary dramatic discourse, and that of the lower tone of voice, i.e. the somewhat retracted voice quality used to give an illusion of whispering. Illusion is a key-word in this case since a "real" whisper would not be possible to perceive in most theatrical situations.

This work is the beginning of a series of studies of stage-speech, aimed at determining the relevant phonetic parameters of this kind of speech production. The present report deals with the use of fundamental frequency. Other analyses have been done, including the measurement of acoustic vowel space utilization and long-time average spectrum (LTAS). Results from these analyses will be reported at the XIIth ICPhS in Aix-en-Provence, in August 1991.

2. METHOD

Since the acoustic characteristics of the theatre-room is essential to the choice of strategy for voice production [1], the speech material was recorded on stage, in a small theatre in Stockholm, during a performance specially arranged for this purpose.

Two actresses were recorded, both with professionally trained voices, each performing the same piece of text three times using different speech styles; in turn: No 1 - normal stage speech style, as in ordinary dramatic discourse, No 2 - so called "stage whisper", and No 3 - normal person-to-person conversational speech style.

3. ANALYSIS

Shorter sequences of text, approximately 20 — 25 sec long, were chosen to be analyzed in order to study the acoustic characteristics ot the three speech styles, respectively. Fundamental frequency was extracted and histograms showing F0 distribution were drawn, by means of the computer program SWELL [2].

4. RESULTS

Since the two actresses are using somewhat different production strategies, the data will be presented "within subjects", with comparisons made between the three speech styles for each subject separately. There are, however, in an inter-subject comparison, some common features concerning the ways in which fundamental frequency is distributed. These will be commented on in the concluding remarks.

4.1 F0 distribution within subjects

4.1.1 Subject BA

For style No 1 (fig B1) the histogram shows a total F0 distribution range of roughly 100—390 Hz. Mean value: 222 Hz. Mode: 169 Hz. Standard deviation: 52.9 Hz. The histogram contour forms a neatly gathered figure where the distribution is rather evenly spread mainly between 150 and 290 Hz, and it has a somewhat flatter slope towards the higher frequencies.

In style No 2 "stage whisper" (fig B2) the F0 range is less extended, mainly covering the area between 120 and 360 Hz. Mean: 208 Hz. Mode: 206 Hz. St.dev: 38.9 Hz. This configuration has a triangular shape similar to that of normal speech, with a slightly flatter slope towards the higher frequencies. The mostly favoured frequencies in this style lie between 160 and 260 Hz.

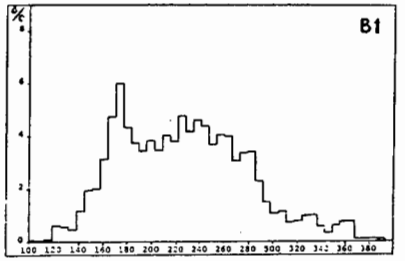


Figure B1. Extracted F0 (in Hz) of normal stage speech (subject BA). Mean: 222 Hz. Mode: 169 Hz. St.dev: 52.9 Hz.

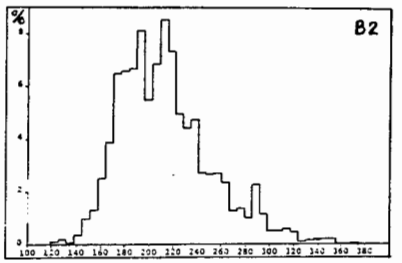


Figure B2. Extracted F0 (in Hz) of stage whisper (subject BA). Mean: 208 Hz. Mode: 206 Hz. St.dev: 38.9 Hz.

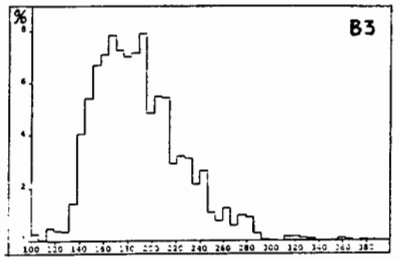


Figure B3. Extracted F0 (in Hz) of normal conversational speech (subject BA). Mean: 185 Hz. Mode: 188 Hz. St.dev: 36.8 Hz.

Style No 3 (fig B3) has a total range of about 100—380 Hz, though it is mainly concentrated to frequencies between 140 and 280 Hz. Mean: 185 Hz. Mode: 188 Hz. St.dev: 36.8 Hz. The pattern of spreading is almost identical to that of the stage whisper (fig B2). The main difference between styles No 2 and 3 lies in the frequency levels being used. This could be described as stage whisper having a frequency downshift of about 20 Hz, compared to normal conversational speech.

4.1.2 Subject GN

For style No 1 (fig G1) the F0 distribution covers a total range of 100—350 Hz but is mainly concentrated to frequencies between 130 and 270 Hz. Mean: 186 Hz. Mode: 155 Hz. St.dev: 45.5 Hz. The histogram displays a tendency towards a bimodal structure where the F0 distribution appears to be divided into two peaks, one around 160 Hz (close to the mode value) and the other around 260 Hz. For this subject, however, there is no evidence of perturbations such as diplophonia or switches between modal and loft registers. The histogram configuration thus presumably demonstrates one impact of an individual speech strategy.

In style No 2 "stage whisper" (fig G2) the F0 range is less extended, roughly covering 100—260 Hz, and with its main distribution concentrated to the area between 130 and 230 Hz. Mean: 173 Hz. Mode: 138 Hz. St.dev: 34 Hz. The contour of this histogram has a very steep slope from around 140 Hz (i.e. about mode value) down to 120 Hz. The slope towards higher frequencies is much flatter.

Style No 3 (fig G3) has a total range of about 100—300 Hz; distribution mainly concentrated between 140 and 260 Hz. Mean: 195 Hz. Mode: 160 Hz. St.dev: 37.3 Hz. In this style, the normal conversational speech, there seems to be a slight tendency towards the same kind of bimodal structure as could be seen in style No 1. This is, however, not as obvious in the normal speech as in the stage version. The appearance of the

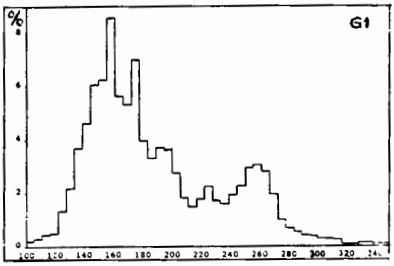


Figure G1. Extracted F0 (in Hz) of normal stage speech (subject GN). Mean: 186 Hz. Mode: 155 Hz. St.dev: 45.5 Hz.

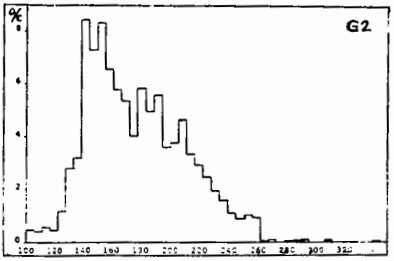


Figure G2. Extracted F0 (in Hz) of stage whisper (subject GN). Mean: 173 Hz. Mode: 138 Hz. St.dev: 34.0 Hz.

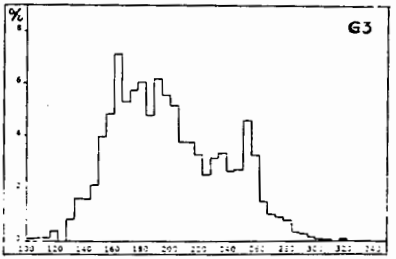


Figure G3. Extracted F0 (in Hz) of normal conversational speech (subject GN). Mean: 195 Hz. Mode: 160 Hz. St.dev: 37.3 Hz.

same distributional pattern in both styles may support the image of the individual production strategy for this subject.

5. DISCUSSION

Subject BA uses a slightly wider total range of fundamental frequency in her normal stage speech but the effectively utilized range in this style is about as wide as that of her normal conversational speech (table BA). For the stage speech, however, F_0 appears to be somewhat higher (+10 Hz) throughout. The mean value here is relatively much higher, and so is the standard deviation which gives evidence of a much greater variability of fundamental frequency in the normal stage speech style.In her whispered style, on the other hand, the effective F_0 range is much narrower. Mean and mode values are almost the same, and standard deviation is much smaller. This also indicates less variability in whispered as opposed to normal stage speech.

Subject GN is consistently using a wider F_0 range in her normal stage speech, totally as well as effectively utilized (table GN). Mean and mode values are somewhat lower in both her stage speech styles than in her normal conversational style. The standard deviation is higher in her normal stage speech, giving evidence of a greater variability in this style. In her whispered style the F_0 range is more compressed, and the mean and mode values are much lower (roughly —20 HZ) compared to her normal conversational speech.

Using a wider range of F_0 usually applies to normal conversational speech as well when the speaker wishes to emphasize or distinguish something in the spoken message. It is therefore not surprising to find that this appears to be systematically used in stage speech.
Decreasing the F_0 range appears to be an effective way of establishing a more intimate speech character, in order to create an illusion of whispered speech. In addition to this, as a recurrent theatrical technique, visual means of conveying the message are being used, such as posture and bodily behaviour, which are crucial elements in stage acting.

Table BA

Style	Mean	Mode	St.dev	Total range	Effect. range
No 1	222	169	52.9	100-390	150-290
No 2	208	206	38.9	120-360	160-260
No 3	185	188	36.8	100-380	140-280

Table GN

Style	Mean	Mode	St.dev	Total range	Effect. range
No 1	186	155	45.5	100-350	130-270
No 2	173	138	34.0	100-260	130-230
No 3	195	160	37.3	100-300	140-260

6. ACKNOWLEDGEMENT

This work has been made possible by the help and support of several of my teachers and colleagues, at the Phonetics dept of the University of Stockholm, and The Royal Institute of Technology, Stockholm, as well as at Folkteatern i Sverige, Södra Teatern, Stockholm. Dr Robert McAllister, University of Stockholm, has been acting as my supervisor. All of them are hereby gratefully acknowledged.

7. REFERENCES

[1] TERNSTRÖM, S. (1989), "Long-time average spectrum characteristics of different choirs in different rooms", *STL-QPSR* 3/89, Dept of Speech Communication and Music Acoustics, Royal Institute of Technology, Stockholm, Sweden.

[2] SWELL, computer program produced by Soundswell Music Acoustics HB, Solna, Sweden.

ARTICULATORY AND ACOUSTIC MEASUREMENTS OF COARTICULATION IN IRISH (GAELIC) STOPS

Ailbhe Ní Chasaide and Geraldine Fealy

Centre for Language and Communication Studies,
Trinity College, Dublin 2, Ireland

ABSTRACT

The rich consonantal system of Irish offers a testing ground for the hypothesis that the phonology of a specific language may constrain otherwise (presumed) universal coarticulatory tendencies. Articulatory and acoustic measures of coarticulation for VCV sequences are presented, where V = each of 5 tense vowels, and C = voiced stops for each of 6 contrasting places of articulation (involving primary and/or secondary places of articulation). Results do in general support the hypothesis; coarticulation of Irish stops is very limited when compared with known data from other languages. Such coarticulation as was found, tended to be carryover rather than anticipatory. Fairly extensive acoustic (but not articulatory) evidence for coarticulation was found for /g/. F2 for articulations in this region may be particularly sensitive to lip rounding.

1. INTRODUCTION

Irish (Gaelic) stops offer a means of testing the hypothesis that a phonological system with a large number of contrasts will constrain the extent to which coarticulation is "allowed". The consonantal system of Irish involves a dichotomy into a palatalised series (phonologically symbolised with /'/) and a velarised series of segments. The opposition of palatalised and velarised pairs may involve simply the secondary articulation (e.g., /b, b'/ = [bˠ, bʲ]), the primary articulation (e.g., /g, g'/ = [ɟ, g]) or a combination of both (e.g., /d, d'/ = [d̪ˠ, d̠ʲ]). Given that Irish has a six way contrast, one would predict that coarticulation for these stops would be much more limited than in a language with, say, a three way contrast.

2. METHODS AND MATERIALS

Recordings were of evenly stressed VCV utterances, where C = one of the stops /b', b, d', d, g', g/ and V = one of the tense vowels /i, e, *a*, o, u/, spoken by a male speaker of Connemara Irish. Two separate recordings were made; one with simultaneous EPG and speech waveform, and a second high quality acoustic recording, on which our spectrographic measurements are based. In the latter, there were five repetitions of each consonant in each vocalic environment, i.e. a total of 750 items. The EPG recording was similar excepting the omission of labial consonants.

From our spectrograms we measured in each instance the frequency of F2 at the following four points: the V1 steady state, the V2 steady state, the endpoint of the transition from V1 to the consonant which we term locus 1 (L1), and the starting point of the transition from consonant to V2, which we term locus 2 (L2). Our procedures were intentionally modelled on those used in Öhman's classic study of coarticulation [3], and our results are compared below with some of his. In the EPG data we measured the contact pattern at the first frame for which there was evidence of full closure (C) and at the last frame for which there was full closure prior to the consonant's release (R). As for the velar stops, the occlusion was further back than the last row on the palate, points C and R were determined from the acoustic recording.

3. RESULTS

3.1 Articulatory measures

The results largely support our hypothesis. Fig. 1 illustrates the frequency of

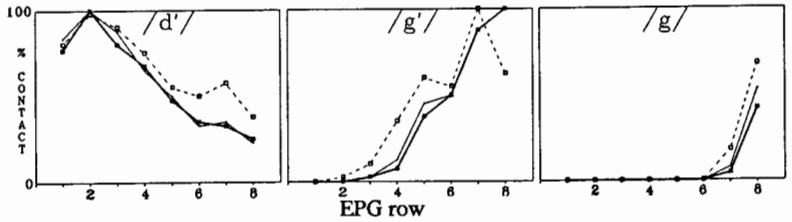


Fig. 1. % contact per EPG row for /d', g', g/ of Irish at time point C. Contexts: □- - -□ = /i-i/, ——— = /*a-a*/ and ✱———✱ = /u-u/.

EPG activation per row for /d', g', g/ at point C in the symmetrical vowel contexts /i-i/, /*a-a*/ and /u-u/, which serve to illustrate the maximum likely range of coarticulation. (Note that row 1 of the EPG palate corresponds to the dental region, and row 8 to the back of the hard palate.) Data for the dental /d/ in the first two of these environments is similarly shown in Fig. 2 along with roughly comparable data for dental stops in French and Italian, taken from [2]. The comparison gives only a general impression of differences, as the data for the latter two languages differed somewhat from the Irish. They involved voiceless consonants measured in /ˈbVtV/ words. Furthermore, the low vowel has a more front quality in these languages.

Some coarticulation does occur for the Irish stops. For /g'/ and for /g/ (insofar as one can determine from the limited contact pattern in the latter) the primary place of articulation appears to be more advanced in /i-i/ as compared to /u-u/ and /*a-a*/ environments. For /g'/ the front of the tongue forward of the constriction is also relatively higher in the /i-i/ context. A greater raising of the tongue front can also be observed for /d'/ in this environment. Of the four stops, /d/ exhibited least coarticulation. For all, the difference is negligible by time point R: the /i-i/ and /u-u/ environments yield almost identical contact patterns, whereas a slightly lower tongue body is in some cases observed in the /*a-a*/ environment.

Fig. 2 illustrates the striking lack of coarticulation in Irish /d/, as compared to the other languages. For Italian and French, the tongue front would appear to be much higher in the /i-i/ than in the /a-a/ environment (as can be deduced from the relative degree of EPG activation in rows 4 to 8). The tongue body behind the primary constriction is free to coarticulate to the vowel's configuration. This contrasts sharply with the Irish pattern, which shows virtually no contextual difference for these rows. In French, the primary place of articulation is also affected by the vocalic context: one would infer a lamino postalveolar articulation in /i-i/, as compared to a more apical dental-alveolar articulation in the /a-a/ context. In Irish, the differences that emerge in the three anterior rows of the

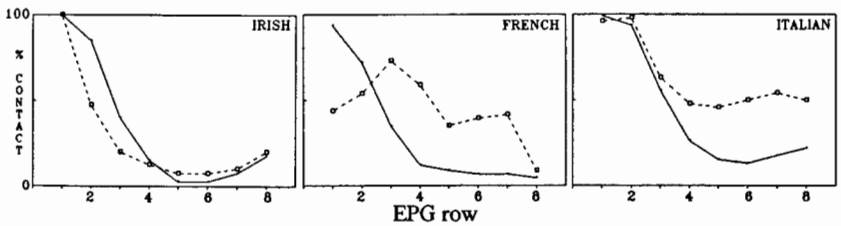


Fig. 2 % contact per EPG row for dental stops: /d/ in Irish; /t/ in French and Italian. Contexts: □- - -□ = /i-i/ and ——— = /*a-a*/ (Irish) and /a-a/ (French and Italian). Further differences in contexts are explained in text. French and Italian data from Farnetani et al. [2].

palate suggest a greater depression of the tongue blade in the /i-i/ context. This is the opposite of what might be expected if coarticulation were to occur, and suggests considerable coarticulatory resistance for this stop. Note that in terms of primary and secondary articulation, the dental phoneme of French occupies about the same phonetic space as the dental *and* postalveolar phonemes of Irish.

The asymmetrical vowel contexts permit inferences on the direction of coarticulation. Carryover coarticulation is clearly dominant. V1 effects are seen at time point C but only marginally at time point R. V2 appears to have little effect at either point. Even in the symmetrical vowel contexts, differences noted at time point C were largely absent at point R.

3.2 Acoustic measures

The acoustic measures also suggest considerable constraint on coarticulation. Fig. 3 shows overall averages for L1 and L2 as a function of V1 and V2 respectively. This shows the correlation between L1 and V1 (or between L2 and V2). As the vowels are arranged in the order of descending F2, coarticulation would be indicated by a negative slope for the line connecting L averages: the steeper the slope, the greater the degree of coarticulation to the "near" vowel. The range of variability in L1 (or L2) due to the transconsonantal vowel is also shown by the vertical lines which run through the average values. For example, a longer line for L1 reflects a greater degree of coarticulation to V2. The same would hold for L2 ranges as a reflection of carryover V1 influence. The right hand panel in this figure shows similar data for Swedish, calculated from Tables II and IV in [3]. The two sets of data differ in that for Swedish, only rounded vowels were used. Öhman estimated that if the near vowel is kept constant, an L1 or L2 range of over 100 Hz can with confidence be attributed to a coarticulatory influence of the transconsonantal vowel.

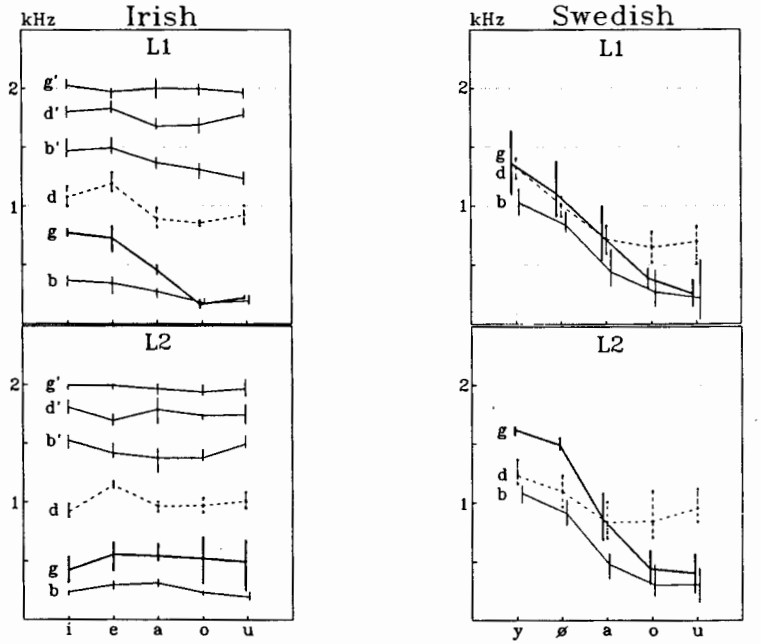


Fig. 3 L1 average values as a function of V1 (upper panels), and L2 averages as a function of V2 (lower panels). Vertical lines through average values show range of variation in any VC- (or -CV) sequence as a function of the transconsonantal vowel. Swedish data calculated from Öhman [3].

Locus variation is very limited in Irish, as comparison with Swedish shows. Considering first the effects of the near vowel, one can see major locus shifts in Swedish (i.e. L1 varies greatly with V1; L2 covaries with V2). In Irish, a much more limited coarticulation is found for L1 as a function of V1. Only with /g/ is the extent striking. There is virtually no evidence of L2 coarticulation to V2.

The effect of the transconsonantal vowel is also generally very limited in Irish, judging by the length of the vertical lines. The velar stop is again the exception. Note also the marked difference in directionality: L2 ranges (for /g/) are large, showing the carryover effect of V1; L2 ranges are typically very restricted (for all stops, including /g/), indicating a general absence of anticipatory coarticulation. The average range of locus variation (L1 and L2) as a function of the transconsonantal vowel is 122 Hz for all the consonants (97 Hz if one omits L2 of /g/). Öhman gives a comparable average of 280 Hz for the Swedish data. This is very close to that obtained in this study for L2 of /g/, which was 290 Hz.

Taking both the effects of the near and the transconsonantal vowel into account one can observe that F2 locus values are much more unique for the Irish stops than for the Swedish. As pointed out by Öhman, for a given VC- (or -CV) sequence, there is typically considerable overlap of locus values, particularly for Swedish /d/ and /g/. The only striking case of overlap in Irish is for /g/ and /b/ when V1 = /o/ or /u/.

4. CONCLUSIONS

Both the acoustic and articulatory measures point to stops in Irish being relatively resistant to coarticulation. This broadly supports the hypothesis that coarticulation is constrained by the phonology of a particular language, and that the propensity to coarticulation can to some extent, be predicted from the size of the phonological inventory. Some limited carryover effects are found: V1 affects C and L1, but generally has little effect on R and L2 (excepting L2 of /g/). There is virtually no evidence of anticipatory coarticulation.

Not everything, however, can be accounted for in terms of phonological constraints. The much greater acoustic evidence of coarticulation for /g/ than for the other consonants would not be predicted on phonological grounds. Nor would it be expected on the basis of the articulatory data. Looking back at Fig. 1, there would appear to be at least as much articulatory evidence of coarticulation for /g'/ as for /g/ (although the articulation of the latter can only be inferred on the basis of EPG data). Yet it is striking how different the acoustic measures are for these two stops, with /g/ showing by far the most variability, and /g'/ almost least.

An explanation of this apparent acoustic/articulatory "mismatch" would probably need to invoke the role of lip rounding as a contributory factor to the acoustic coarticulation in /g/. Fant's [1] nomograms suggest that for constrictions in the velar region, F2 would be very sensitive to lip rounding on the one hand and to tongue advancement/retraction on the other. If this line of explanation is correct, it suggests in turn that palatals may be characterised by a high degree of acoustic stability. We hope to investigate this area more fully in the future, using additional techniques for measuring lip movement.

ACKNOWLEDGEMENTS

This work was funded by ESPRIT II, BRA no. 3279: ACCOR. We are grateful to Francis Nolan for helpful comments.

REFERENCES

[1] FANT, G. (1960), The Acoustic Theory of Speech Production, Mouton, Hague (2nd edition 1970).

[2] FARNETANI, E., HARDCASTLE, W., & MARCHAL, A. (1989), "Cross-language investigation of lingual coarticulatory processes using EPG", NATO Research Report.

[3] ÖHMAN, S. (1966): "Coarticulation in VCV utterances: spectrographic measurements", J.Acoust.Soc.Am. 39, 151-168.

AN ACOUSTIC & PERCEPTUAL STUDY OF UNDERSHOOT IN CLEAR AND CITATION-FORM SPEECH

Seung-Jae Moon

Department of Linguistics
University of Texas at Austin

ABSTRACT

This study investigated vowel reduction (the so-called undershoot phenomenon) in "Clear" (CS) vs. "Citation-Form" (CF) speech. Undershoot was readily observable for all 5 speakers. Furthermore, the results suggested that CS is not merely a louder version of normal speech, but it involves an active reorganization of phonetic gestures. A perception test showed that, in general, CS is more intelligible than CF under identical S/N conditions.

1. INTRODUCTION

In this study we investigated the acoustic characteristics of "clear speech" which was defined in terms of an explicit instruction to subjects to "overarticulate".

The following questions were addressed: Is clear speech merely a louder version of citation-form speech? Or does it also involve an active reorganization of speech gestures? If so, what is the perceptual significance of that reorganization?

The point of departure for the present experiments is unresolved issues of vowel reduction and the so-called undershoot phenomenon. The strong version of duration-dependent undershoot [1] makes vowel duration the only determinant of undershoot. On the other hand, there are findings in the literature[2] that are at variance with that model.

2. ACOUSTIC EXPERIMENT

2.1 Procedure

To induce duration-dependent undershoot, the following test words were used: wheel, wheeling, Wealingham, will, willing Willingham, well, welling Wellingby, wail, wailing and Wailingby. The following three criteria were considered in selecting the test words. First, get a maximum locus-to-target distance. Second, the vowels under analysis must have equal stress. Finally, the duration of the vowels of interest should vary systematically over a considerable range. The first condition was imposed because the larger a given formant movement, the greater the possibility that it will serve as a sensitive indication of articulatory undershoot. It was met by selecting front vowels in a labio-velar context. The second and third criteria were met by using so-called word-length effect.

In addition to the /w/-vowel-/l/ contexts, the same front vowels were also measured in an /h/-vowel-/d/ context. Those measurements were used to provide null-context target values.

For citation-form speech, subjects received no other instructions than to keep their effort and tempo constant and at comfortable levels. For clear-speech, they were explicitly instructed to overarticulate, that is to read the words as clearly as they could. To maintain this performance, during the recording of clear speech, at unpredictable moments, the subject was interrupted through the intercom by the experimenter who would pretend that the token just pronounced had not been understood, and would ask for a repetition. A total of 5 speakers were recorded and measured.

2.2. Results

There is a clear duration dependent undershoot effect: As vowels get shorter, the formant measurements are shifted further and further away from their null-context values and closer and closer to their position in [w].

A closer examination of the raw data indicates that the undershoot effects are vowel-specific. In general, tense vowels are more resistant to undershoot than lax vowels.

Also, the degree of undershoot is talker-specific. Each individual talker exhibits his own pattern of undershoot.

Undershoot is also style-specific: When the clear speech measurements are compared with the data from the other conditions in an F_2-F_1 vowel space diagram, it becomes clear that, for all speakers and all conditions, it is closer to the formant patterns of the null-context vowels. Another way of expressing that observation is to say that clear speech is more peripheral in the vowel space than citation-form speech. It seems as if the vowel space is a flexible object which speakers can adaptively expand or contract according to situational needs.

These findings refute the strong version of the undershoot model: Information on vowel duration alone is not sufficient to predict formant undershoot. Also these results suggest that clear speech is not merely citation-form speech spoken louder and more slowly. Clear speech transforms also involve an active reorganization of phonetic gestures.

A decaying exponential model was fitted to the data from 2 speakers to obtain a more systematical and economical description. The results indicate that the claims made above (vowel-specific, style-specific and talker-specific undershoot pattern) shall be weakened to some extent. It was shown that the dependence of degree of undershoot on identity of vowels and speaking styles is not as strong as it looked based on the raw data. For at

least one speaker, undershoot effects are fairly uniform for all vowels and all speaking styles, provided that appropriate target values are selected for styles and vowels. This modeling also shows that speakers differ in terms of the coefficients used to describe degree of undershoot. However, this fact does not suggest that the speakers need to control these variations directly. These variations are likely to be the results of the articulatory movements themselves and of the non-linear acoustic mapping of the articulatory gestures, not the results of active control over those constraints.

3. PERCEPTUAL EXPERIMENT

What is the perceptual significance of the observed acoustic changes? It is reasonable to assume that, when people speak more clearly, they do so to communicate better and to make their speech more intelligible to the listener. We must ask then, are the clear speech tokens indeed more intelligible than the citation-form speech tokens? To address this question, the following perception experiment was carried out.

3.1. Procedure

The samples for the listening test were chosen from a subset of the words analyzed acoustically. A single representative token was selected for each combination of speaker, vowel, word-length and speaking style. From the various repetitions of the test items, the exemplar which showed a median "acoustic distance" to the null-context was taken as the representative token. For the present purposes, acoustic distance is defined as the Euclidean distance between two points in a three-dimensional formant space calibrated in Mel units.

The representative words were mixed with five different levels of low-frequency weighted Gaussian noise which had a spectral shape of -6dB/oct. One of the noticeable differences between citation-form speech and clear speech was their different intensities. For all five speakers, clear speech was approximately 3-5 dB more intense than citation-form speech. Since our aim was to undertake an intelligibility test based solely on acoustic characteristics other than amplitude, the differences in loudness was normalized by using a special computer program written by Jerry Lame.

Each stimulus was led by a 150ms segment of speech-free noise and was also followed by an interval of noise adjusted so that the duration of the whole stimulus would be the same within a given speaker. There were 120 stimuli per speaker.

These stimuli were presented to normal hearing subjects for identification through headphones. At least 24 responses were collected for each stimulus.

The responses were processed for each speaker and the percentage of correct identification was calculated for each step of S/N ratio.

3.2. Results

In general, the tense vowels show a strong clear-speech advantage while the lax vowels do not. This pattern is consistent for all 5 speakers.

However, let us now consider an alternative measure of S/N ratio. It is the same as before for citation-form speech but does not involve normalizing clear speech. It leaves intensity differences between clear and citation-form speech as they were on the original tape recordings.

When the second definition of S/N ratio is applied to the present data, all test words, when spoken clearly, tend to be more intelligible. With only marginal exceptions that observation is true for all speakers and for all test words.

It can be speculated that the reason for the perceptual advantage of clear speech is multi-dimensional. First, clear speech words tended to be 3-5 dB more intense than citation forms. Second, the formant patterns of the clear speech vowels were found to be closer to their null-context values. And also clear-speech is longer in duration than citation-form speech. In other words, speakers used various strategies to keep undershoot effects down in clear speech.

The intelligibility tests indicate that, in the case of tense vowels, these formant pattern adaptations and systematic duration changes are likely to be responsible for the improved identification scores of clear speech.

Although the style-dependent formant changes and duration changes in lax vowels were entirely analogous to those for the tense vowels, they were not sufficient to make the clear variants more intelligible.

4. CONCLUSION

It has been shown that clear speech is a speech act which involves active reorganization of acoustic patterns and the underlying articulatory gestures, and that it has clear perceptual advantages.

Everyday informal experience suggests that "clear speech" is invoked by a speaker to meet certain communicative and situational demands. And that speakers change and modify their speech according to the needs of their listeners.

The present results indicate that speakers are quite capable of doing so in an experimental situation. They show an ability to successfully adapt to varying demands for explicit signal information.

5. REFERENCES

[1] LINDBLOM, B. (1963), "Spectrographic study of vowel reduction", JASA 35: 1773-1781

[2] GAY, T. (1978), "Effect of speaking rate on vowel formant movements", JASA 63 : 223-230

6. ACKNOWLEDGEMENTS

This research was supported by a grant from the Advanced Research Program of the Texas Board of Coordination and grant No. BNS-9011894 from the NSF.

COARTICULATION AND THE PERCEPTION OF NASALITY

Rena Arens Krakow† and Patrice Speeter Beddor‡

†Temple University, Philadelphia, PA; †Haskins Laboratories, New Haven, CT; and ‡University of Michigan, Ann Arbor, MI

ABSTRACT

Nasality judgments of oral and nasal vowels in nasal, oral, and null contexts were elicited from American English listeners. While nasal vowels were most often perceived as nasal, listeners performed best on vowels in isolation and worst on vowels in a nasal context. The consequences of these results for current approaches to coarticulatory compensation are discussed.

1. INTRODUCTION

A growing body of data indicates that vowel perception is influenced by phonetic context such that listeners adjust for the coarticulatory effects of adjacent consonants. For example, Kawasaki [2] showed that perceived vowel nasality is enhanced as flanking nasal consonants are attenuated; the same vowels in a clearly audible nasal context are more likely to be perceived as oral. One possible interpretation of these results is that, when presented with a nasal vowel in a nasal consonant context, listeners do not integrate the nasal resonance with the vowel itself, but instead hear it as part of the nasal consonant [1].

However, the results of Krakow et al. [3] have been interpreted as suggesting that listeners are able to associate the nasal resonance in a vowel in a nasalizing context with nasal coupling. We found that, for American English listeners, oral and nasal vowels produced with the same oral tract shape were perceived as having the same height given an appropriate coarticulatory context (i.e., CVC vs. CṼNC, where C is an oral consonant and N is a nasal consonant). But when the oral and nasal vowels were embedded in an oral context, the nasal vowels were perceived as shifted in height (CVC vs. CṼC). We suggested that, lacking a context for nasality, listeners interpreted the low-frequency nasal resonance of the nasal vowels in CṼC syllables as reflecting a shift in tongue/jaw height. In contrast, the presence of a nasal consonant in CṼNC syllables enabled listeners to correctly attribute the low-frequency nasal resonance in the nasal vowel to nasal coupling.

The results of Kawasaki [2] and Krakow et al. [3] therefore allow for conflicting interpretations. But the potential conflict cannot be resolved with these two studies alone as there are several methodological differences which may have influenced the findings. First, Kawasaki compared nasal vowels in appropriate coarticulatory contexts (NṼN) and isolation (Ṽ), while we compared nasal vowels in appropriate (CṼNC) and inappropriate (CṼC) consonantal contexts. It is unlikely that the perception of vowels in an inappropriate consonantal context is analogous to the perception of vowels in no context. Second, Kawasaki examined nasality judgments while we examined vowel height judgments. It is possible that, although listeners in our study were able to correctly attribute the effects of nasal coupling on the vowel spectrum to nasality (in CṼNC contexts), they would not have labeled these vowels as "nasal". Third, Kawasaki used edited natural speech while we used synthetic speech.

These differences leave many questions regarding the effects of coarticulatory contexts on perceived vowel nasality unresolved and provide the basis for the present study, which compared listeners' judgments of edited naturally produced tokens of nasal and oral vowels in C_C, N_N, and #_# (null) contexts. Using data obtained from vowel nasality judgments (elicited in a paired comparison test) and vowel identity judgments (matching test), we address the following questions: (1) Can listeners determine whether a vowel in a nasal context is nasalized? Kawasaki's results indicate that listeners will identify the vowel in NṼN as oral, while our interpretation of Krakow et al. suggests that listeners might perceive the vowel as nasal. (2) Are listeners more accurate at determining the nasality of a vowel in isolation than in (an appropriate or inappropriate) context? Previous work by Stevens et al. [4] suggests that within-category information regarding vowel quality is more evident in isolated vowels than in vowels in context. Here we ask whether the same is true of vowel nasality. (3) Are listeners more accurate when judging oral vowels as oral than when judging nasal vowels as nasal? Is nasality per se problematic, irrespective of the context?

2. METHODOLOGY

We recorded a male native speaker of American English producing multiple tokens of bed and men. Two tokens of each were selected so as to yield two bed-men pairings whose members were matched as closely as possible for duration and intensity. Waveform editing techniques were used to create the following 6 syllable types: CVC ([bɛd]), NṼN ([mɛ̃n]), isolated oral V ([ɛ] from [bɛd]), isolated nasal vowel ([ɛ̃] from [mɛ̃n]), cross-spliced CṼC ([bɛ̃d] with consonants from bed and vowel from men), and cross-spliced NVN ([mɛn] with consonants from men and vowel from bed). To control for any effects of splicing, the CVC and NṼN tokens were created by splicing across the two repetitions of each pair type.[1]

Twelve native speakers of American English were asked to respond in two test conditions. The Matching test involved an ABX format. In each trial, listeners heard two consonant-vowel-consonant syllables followed by an isolated vowel. Listeners were asked to determine whether the isolated vowel sounded more like the vowel in the first or second consonant-vowel-consonant syllable. Each AB pair was either CVC-CṼC or NVN-NṼN and X was either V or Ṽ, yielding four ABX condition types.

A Paired Comparison test was presented after the Matching test. This test involved all possible pairings (AB) of the 6 types of syllables, for a total of 21 condition types. Listeners were asked to determine whether the first or second member of each pair sounded "more nasal" or whether they sounded "equally nasal". For both tests, there were 8 randomized repetitions of each condition, with the order of A and B in each condition counter-balanced.[2]

3. RESULTS

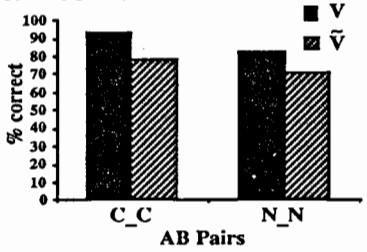


Figure 1. Matching test results. Each column represents responses to one of the ABX conditions (where A and B differ in vowel nasality). A or B (the correct match to X) is represented along the abscissa. and X is represented by column type (solid or hatched).

Figure 1 shows the percent correct responses to the Matching test. Listeners were generally quite accurate at matching vowels in (appropriate or inappropriate) context to vowels in isolation on the basis of nasality. Nonetheless, listeners were more accurate at matching oral vowels than nasal vowels, and more accurate at matching isolated vowels to vowels in an oral context than to vowels in a nasal context. Listeners did least well matching NṼN and Ṽ, making the most common error a match between NVN and Ṽ. Listeners incorrectly matched NVN to Ṽ over 30% of the time; they incorrectly matched NṼN to V less than 20% of the time.

Figures 2-4, which we shall address in turn, show the results of the Paired Comparison test. Figure 2 focuses on the effect of inappropriate consonantal versus null contexts on perceived vowel

nasality. For all types of pairings, the nasality of a nasal vowel was more often correctly judged when in isolation (Ṽ) than when in a NṼN context (Fig. 2a) or in a CṼC context (with one exception; Fig. 2b). Comparison of the perceived nasality of nasal vowels in consonantal contexts shows greater accuracy for vowels in inappropriate CṼC contexts than appropriate NṼN contexts (Fig. 2c). Apparently, an inappropriate consonantal context (CṼC) makes nasality more evident than an appropriate one (NṼN), but a null context (Ṽ) makes nasality most evident.

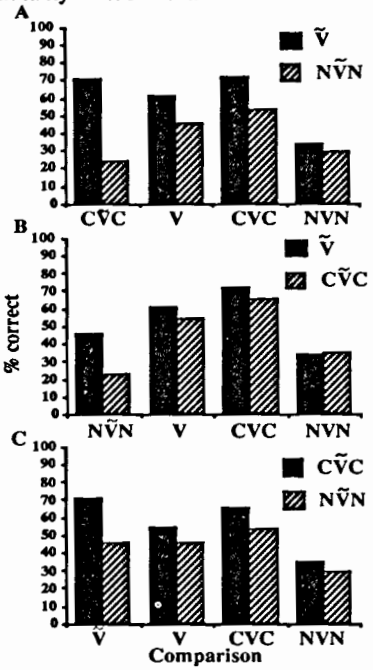


Figure 2. Paired comparison test results showing the effect of context. Each column represents correct responses to the AB comparison indicated. (Some conditions are repeated for reference.)

Figure 3 addresses the question of whether vowel nasality is more difficult to assess on nasal vowels than on oral vowels for American English listeners, independent of context. With one exception, listeners were less accurate at judging two nasal vowels as similar than they were at making the same judgment of two oral vowels.

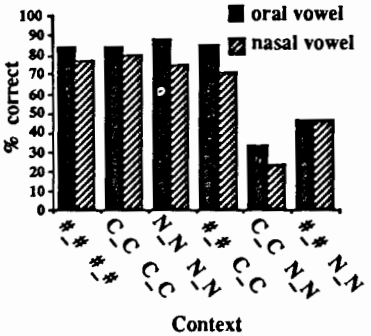


Figure 3. Results of paired comparisons between the contexts shown for two oral or two nasal vowels.

The exceptions to the two generalizations above both involved the N_N context, leading us to ask whether the N_N context is problematic irrespective of vowel nasality. Figure 4 shows listeners' responses to pairs involving one oral and one nasal member. The data represent the percentage of "more nasal" responses to each pair member. Nasal pair members were judged "more nasal" more often than oral ones when the oral vowels were in isolation (Fig. 4a) or in an oral context (Fig. 4b). But, in a nasal context, the *oral* member was more often judged as the more nasal member (Fig. 4c).

4. DISCUSSION

Overall, the results suggest that perception of vowel nasality is influenced by the coarticulatory context in which the vowel occurs. In two types of tests designed to elicit nasality assessments, American English listeners were less accurate at judging a vowel as nasal in appropriate (N_N) than in inappropriate (C_C, #_#) contexts. However, the data exhibit certain patterns not predicted by current approaches to coarticulatory compensation. One such pattern is that listeners were generally more accurate at judging oral vowels as oral than judging nasal vowels as nasal, irrespective of coarticulatory context. This finding may be linked to the non-distinctive status of vowel nasality in English, and points to the importance of extending this research to languages with distinctive vowel nasalization. A second pattern is that the distinction between "appropriate" and "inappropriate" coartic-

ulatory contexts is insufficient to explain listeners' judgments. That is, listeners perform less accurately on C_C than #_# conditions, both of which are inappropriate contexts for vowel nasalization in English.

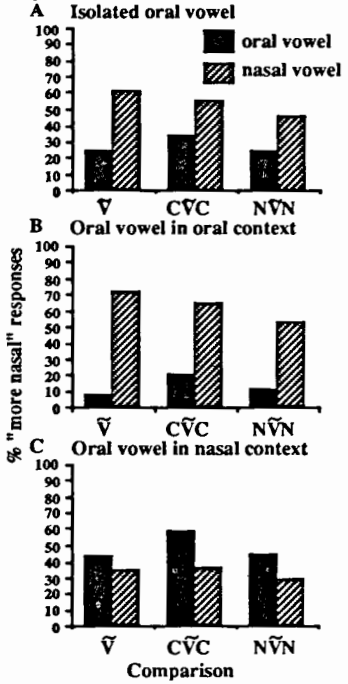


Figure 4. Results of paired comparisons between oral and nasal vowels. The oral member of each pair is represented in each panel (a-c); the nasal member of each pair is represented in the abscissa.

Furthermore, the present data fail to support a strong interpretation of the results of either Kawasaki or Krakow et al. Listener judgments of nasal vowels in appropriate nasal contexts (NṼN) were not as consistently oral or nasal as the former or latter, respectively, seem to predict. While the data suggest that listeners may factor out some of the nasality given an appropriate coarticulatory context, they are still more likely to judge these vowels as nasal than oral. In general, American English listeners demonstrate a lack of certainty as to the nasality of vowels in nasal contexts, an uncertainty which holds for both appropriate (NṼN) and inappropriate (NVN) coarticulatory nasal contexts. (It is unclear from these data whether the unexpectedly large number of "more nasal" responses to NVN stimuli reflect the phenomenon of hyponasality or whether these stimuli simply sounded odd to listeners, with "odd" being encoded as a "more nasal" response.) Listeners appear to be tacitly aware that a nasalizing context alters a phonemically oral vowel. And, in most cases, they will report a nasality difference between a contextually appropriate nasal vowel and a corresponding contextually appropriate oral vowel.

5. REFERENCES

[1] FOWLER, C. A., (1987), "Perceivers as realists, talkers too: commentary on papers by Strange, Diehl et al., Rackerd and Verbrugge.", *Journal of Memory and Language, 26*, 574-587.

[2] KAWASAKI, H., (1986), "Phonetic Explanation for phonological universals: the case of distinctive vowel nasalization", In J.J. Ohala and J.J. Jaeger (Eds.) *Experimental Phonology*, Orlando, FL: Academic, 81-103.

[3] KRAKOW, R.A., BEDDOR, P.S., GOLDSTEIN, L.M., FOWLER, C.A., (1988), "Coarticulatory influences on the perceived height of nasal vowels", *Journal of the Acoustical Society of America, 83*, 1146-1158.

[4] STEVENS, K.S., LIBERMAN, A.L., ÖHMAN, S.E.G., STUDDERT-KENNEDY, M., (1969), "Cross-language study of vowel perception", *Language and Speech, 12*, 1-23.

Work supported by NIH Grants DC-00121 and RR-05596 to Haskins Laboratories, and the University of Michigan.

---

[1] To control for duration differences (the oral vowels in C_C contexts being roughly 50 ms longer than the nasal vowels in N_N contexts), vowel length was manipulated in the isolated vowels and cross-spliced syllables, creating in addition to the normal-length versions, long versions of the nasal vowels and short versions of the oral vowels.

[2] For all test pairs, vowel durations were matched, with the selected duration corresponding to that of the vowel in an appropriate context; in the few pairs where neither pair member was in an appropriate context for English, the duration was that which would normally occur in that context (i.e., longer durations in C_C contexts and shorter durations in N_N contexts).

PERCEPTION OF PREVOICED STOP CONSONANTS BY MONOLINGUALS AND BILINGUALS: EVIDENCE FOR DIFFERENT PERCEPTUAL SENSITIVITIES

L. Coixao and N. Bacri

Laboratoire de Psychologie Expérimentale
Paris V-CNRS-EHESS,Paris, France.

Listeners assimilate foreign speech sounds to their own phonemic categories whenever possible. But what happens for bilinguals when their two languages are closely related? French monolinguals (MF) and Portuguese-French bilinguals (BPF) were tested in identification and AXB discrimination tasks. MFs'fast responses were non random except for the longest prevoicing, whereas BPFs' showed two peaks around Portuguese and French referential values. According to acoustic patterns and task demands, listeners rely either on a phonemic processing strategy or on a goodness of fit strategy which allows MF to build an allophonic space and BPF to keep separate their two languages.

While there is evidence that linguistic experience affects the ability to process phonemic categories as early as the last quarter of the first year [7], there is some disagreement about whether the perceptual analysis bilinguals have to perform is thoroughly determined by phonological constraints or not. When the two languages are closely related, they are not differentiated at the phonological level [2]. But under certain conditions, the effect of phonological constraints can be weakened, and listeners can rely on phonetic cues to keep separate their perceptual representations [3].

Allophonic variants from the point of view of phonemic labelling can be perceived as different. It has been hypothesized that discrepancies between native and nonnative speech sounds receiving an identical label are processed with reference to the acoustic distance between any exemplar and the category center [4] i.e. the acoustic configuration usually produced by native speakers of the two languages.

The present experiments study how listeners process perceptual dissimilarities in two cross-language situations: perception of a Portuguese /da/-/ta/ VOT continuum by French monolinguals (MF) and by Portuguese-French bilinguals (BPF). VOT is generally considered as the most salient cue for voicing when opposing voiced and unvoiced categories if not prevoiced and voiced ones [5]. As for the stop consonant subset, French and Portuguese are closely related. Both languages present a prevoiced-voiced contrast, opposing a long (French) or a very long (Portuguese) voicing lead to a null or a short lagging VOT. According to the assimilation hypothesis [1], allophonic processing for foreign, but neighbouring sounds such as those we study here, is phonemic. MF will assimilate all the prevoiced stimuli to the /da/ category. By contrast, if category goodness plays a role, it could limit allophonic space to certain stimuli.

But what happens for bilinguals when their two languages are closely related? Are their two languages differenciated at the level of perceptual representations, allowing them to detect phonetic differences related to their two languages within a single phonemic category [3]? In this case, they should exhibit a good discrimination accuracy for two distinct areas, corresponding either to the /da/-/ta/ boundary or to a contrast between the French and Portuguese /da/ phonetic categories. On the contrary, if they assimilate the members of the voicing contrast in one of the two languages to those of the other language, due to a partial acoustic overlap, their discrimination should be random, except in the /da/-/ta/ boundary area, common to both languages.

1. EXPERIMENT 1

This experiment was designed first to determine the phonemic /da/-/ta/ boundary values, second to study whether a shift, marking interlanguage interferences [2,3], appeared between BPF and MF responses or not.

1. 1. Method

Subjects. The subjects were 5 MF and 5 BPF students with normal hearing. BPF first language was Portuguese. All bilinguals had been living in France since at least 15 years and acquired French before the age of 5.

Stimuli. A /da/ syllable, produced by a Portuguese monolingual female, was selected (syllable duration: 276 ms, VOT: -96 ms). The test stimuli were digitized at a 16 KHz sampling frequency and VOT reduced by 12 ms steps (from -96 ms to -36 ms) or 6 ms steps (from -36 ms to 0 ms) along the /da/-/ta/ VOT continuum.

Procedure. Subjects listened individually over earphones, in a quiet room, at a comfortable listening level, to 10 blocks, each consisting of one complete randomization of the continuum. The ISI was 3 s and the IBI was 20 s. Listeners' responses were forced choice "Da" or "Ta". All instructions were given in French.

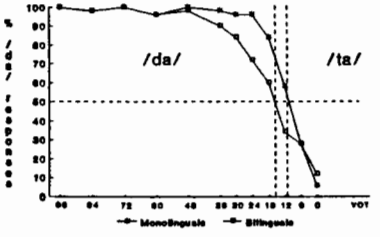


Figure 1. Identification functions for MF (•) and BPF (○) listeners on the /da/-/ta/ continuum (in ms).

1. 2. Results and discussion

The average labeling functions for the two groups are plotted in Figure 1. The /da/-/ta/ boundary fell at -9.2 ms of prevoicing for MF and at -16.6 ms for BPF. An Anova on boundary values showed that this difference between groups was significant ($F(1,8) = 10.4$, $p < .02$). The steep curves suggest that a leading VOT is a strong perceptual cue for BPF as well as for MF. These results differ from those of previous studies on prevoicing [2, 5]. Moreover, there is a clear shift between MF and BPF identification functions. Whatever the case, identification data support the hypothesis of an assimilation of allophonic phonemic variants.

2. EXPERIMENT 2

Even though Experiment 1 suggested an assimilatory process, forced choice labeling could have interfered with perception of differences between stimuli. If allophonic variants have been perceptually assimilated, both MF and BPF should have a good discrimination accuracy just for the stimuli spanning their respective phonemic boundary. Should MF data be non random on the long lead end of VOT continuum and BPF around the medium VOT values, it would undermine assimilation hypothesis and suggest a multi-level processing.

2. 1. Method

Subjects. 10 MF and 10 BPF were tested.

Stimuli and procedure. The same 12 stimuli as in Experiment 1 were used in an AXB discrimination task. A training block of 32 trials preceded 5 blocks of 36 trials, randomized within blocks. ISI was 500 ms, ITI 4 s and IBI 20 s. Subjects had to respond, as quickly and accurately as possible, whether the X stimulus was the same as the first or the third stimulus, by pressing one of two buttons.

2. 2. Results

Mean values of correct responses for the two groups are plotted on Fig. 2. Each data point corresponds to 200 responses per group. The discrimination function for MF exhibited a maximum on the rightward end of the continuum, suggesting an effect of the phonemic /da/-/ta/ boundary. But correct responses are clearly above chance from pair 5 onwards (binomial test, $p < .001$). Results for BPF were less clear-cut, as their discrimination function showed just a

slight peak around a 20ms prevoicing value.

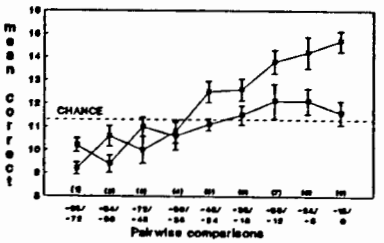


Figure 2. Correct discriminations for MF (•) and BPF (o) as a function of stimulus pairs.

An ANOVA on the correct responses showed an effect of stimulus pair just for MF ($F_{(8,72)}$= 6.9, p<.0001). BPF responses were significantly more correct for the four stimulus pairs presenting at least one short prevoicing (pairs 6 to 9) than for the other ones ($F_{(1,9)}$=8.55, p<.01). Between-group difference was significant for the rightward end of the continuum ($F_{(1,18)}$=6.95, p<.01).

What suggests first a link between discrimination accuracy and phonemic boundary: Discrimination is all the more correct as stimuli pairs span the phonemic boundary. Second, the difference between MF and BPF for the pair enclosing the null VOT value confirms that the slight leftward shift of BPF responses is significant.

An ANOVA on RT data showed a main effect of stimulus pair (MF: $F_{(8,72)}$=6.1, p<.0001; BPF: $F_{(8,72)}$=2.4, p<.03), but not of subject group. Mean RT for MF was 733ms (sd=144ms), for BPF 772ms (sd=202ms).

Following [6], we carried out a threefold partition of RT data to specify the time course of discrimination processes. Each subset/subject contained a third of the data. Between groups range differences were nonsignificant. Proportions of correct responses (Fig. 3a and 3b) were computed for each pair and averaged across subjects (% correct for a specific RT partition * % correct for each partition relatively to the set of correct responses).

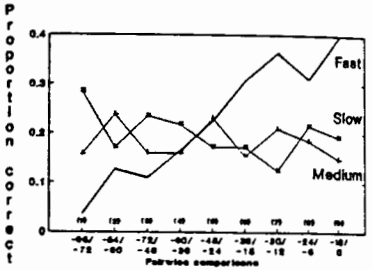


Figure 3a. Proportion of correct discriminations for each RT partition. Monolingual data.

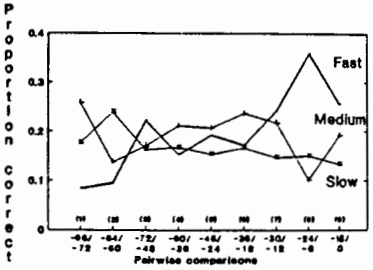


Figure 3b. Proportion of correct discriminations for each RT partition. Bilingual data.

An ANOVA on these proportions showed a main effect of stimulus pair ($F_{8,144}$=4.4, p<.0001). The main difference between groups concerned fast RT (p<.02). Whereas MF discrimination function was quasi-linear, BPF data exhibited 2 peaks, on the pairs 3 and 8. Both groups discriminated better the pair straddling their respective phonemic boundary, but in addition MF showed a good accuracy for "intracategory" pairs and BPF for a pair opposing a long prevoicing to a medium-sized one. It is worth noting that medium and slow responses were random for BPF, but above chance near the rightward end of the continuum (medium and slow RT) and the leftward end (slow RT) for MF.

3. DISCUSSION

A comparison between identification and discrimination data indicates that when processing either a neighbouring language or their two languages, listeners use the phonemic contrast between a long and a short or null prevoicing, but not in all conditions. Specifically, either when labeling stimuli or when responding as fast as possible, they rely on VOT, the critical value of which is common to both languages. Bilinguals and monolinguals gave the same pattern of responses, with a bilingual boundary shift leftward.

However, data do not permit to conclude that both groups have assimilated Portuguese to French stop consonants, as predicted by the assimilation hypothesis [1]. First, MF showed a discrimination accuracy that exceeded widely the /da/-/ta/ boundary area, specially when fast. Second, bilinguals' fast responses were above chance for the contrast between a very long and a medium prevoicing. Thus, listeners of both groups have detected phonetic differences between phonemes receiving the same phonemic label, when corresponding either to values usually produced (MF) or to the contrast between values that respectively characterize bilinguals' two languages. The "goodness of fit" of one of the stimuli in the pair may have facilitated accuracy, if it has been used as a referential point in consideration of French or of French and Portuguese languages.

Another striking result is that, in an AXB test, BPFs' slow responses were never above chance, even though their discrimination accuracy is not significantly poorer than MFs'. In a task having high memory requirement listeners may rely not only on the more salient cue, but also on all the potential cues [8]. Assuming that in this case perceptual analysis takes more time to be processed, its issue depends mainly on the compatibility of cues. Should multiple cues be perceived as lacking coherence, their analysis could not result in a strong discrimination accuracy. It is what happens to bilinguals who, speaking equally well both languages, are plausibly sensitive to the discrepancy between temporal and spectral cues e.g. shorter and shorter French-like prevoicing vs. Portuguese formant values.

Our data thus provide some support to the hypothesis that listeners can use distinct processing strategies when identifying and discriminating speech syllables. According to the acoustic patterns and task demands, they rely on a phonemic processing strategy, specifically in the phonemic boundary area. However, they may take into account the category goodness, when farther from this boundary, in order to differentiate syllables receiving the same label. They can build an allophonic space, and bilinguals can keep separate their two languages. Thus bilinguals and monolinguals appear as perceiving speech according to the same processes, but with different perceptual sensitivities due to linguistic experience.

REFERENCES

[1] BEST, C. (in press), "The emergence of language-specific phonemic influences in infant speech perception", in C. Nusbaum and J. Goodman (eds.), *Development of speech perception*, Cambridge: MIT Press.

[2] CARAMAZZA A., YENI-KOMSHIAN G., ZURIF E., CARBONE M.E. (1973), "The acquisition of a new phonological contrast: The case of stop consonants in French-English bilinguals", *J.A.S.A*, *54*, n°2, 421-428.

[3] FLEGE, J. (1991), "Age of learning affects the authenticity of VOT in stop consonants produced in a second language", *J.A.S.A.*, *89*, 395-411.

[4] GRIESER, D., KUHL, P. (1989), "Categorization of speech by infants: Support for speech-sound prototypes", *Develop. Psychology, 25*, 577-588.

[5] KEATING P., MIKOS M., GANONG III W. (1981), "A cross-language study of VOT in the perception of initial stop voicing", *J.A.S.A*, *70*, 1261-1271.

[6] MILLER J., DEXTER E. (1988), "Effects of speaking rate and lexical status on phonetic perception", *J. of Exp. Psychology: H.P.P, 14*, 369-378.

[7] WERKER J., TEES R. (1984), "Phonemic and phonetic factors in adult cross-language speech perception", *J.A.S.A*, *75*, 1866-1878.

[8] WHALEN, D., SAMUEL, A. (1985), "Phonetic information is integrated across intervening nonlinguistic sounds", *P. & P., 37*, 579-587.

TESTING THE FAIRNESS OF VOICE IDENTITY PARADES: THE SIMILARITY CRITERION

A.C.M. Rietveld* & A.P.A. Broeders**

***Dept. of Language & Speech, University of Nijmegen, Netherlands**
****National Forensic Science Laboratory, Rijswijk, Netherlands**

ABSTRACT

Several factors may adversely affect the reliability of a voice identity parade. This paper concentrates on one of these, the degree of similarity of the voices used in the line-up. It describes two techniques which may be used to measure voice similarity - pairwise comparison and the use of semantic scales - and compares the results with the scores obtained in a recognition experiment using a six-speaker voice line-up. It is suggested that a modified version of the paired comparison technique could usefully be applied either to select voices for inclusion in a line-up or to interpret the results of a voice line-up.

1. INTRODUCTION

In a recent survey of the literature on voice identification research , Deffenbacher et al. [2] observe that voice recognition studies are few in number and widely scattered. Studies dealing with voice identity parades are fewer still. This is somewhat surprising in view of the fact that the voice identity parade is a procedure which, potentially, has a wide range of application in police investigations and legal proceedings. It is frequently the case that the voice of a person involved in the commission of a crime is one of the few clues to the identity of that person. Of course, once a suspect is available, it is possible to ask an earwitness if the voice of the suspect is the same as that heard at the time of the crime. But there is obviously a real danger here that the earwitness may 'recognize' the voice, simply because the suspect's voice sounds similar to that of the criminal. There are many other types of bias which may render the results of an aural confrontation virtually meaningless. Some of these are similar to those that apply to visual identity parades (Clifford [1]). Hammersley and Read [3] briefly discuss some of the precautions that should be taken in conducting a voice line-up. A major criterion for the reliability of the voice line-up is that it should be 'fair'. An important implication of this is that the voices should be similar. It would therefore be useful if some objective measure existed by means of which voice similarity could be determined. Ideally, this would be used in the selection of voices for inclusion in the line-up but if, for some reason, this has not happened, it could also be a useful tool to interpret the results of a voice parade.

2. AIM OF THIS STUDY

The main objective of this study is to explore the possibility of measuring voice similarity for the purposes of a voice identity parade. Two methods, the use of semantic scales and pairwise comparison, are examined, and the results are compared with those of a voice recognition experiment involving a six-speaker line-up.

3. STIMULUS MATERIAL

Various precautions were taken to avoid bias in the stimulus material. Five educated male speakers of Dutch were recruited to produce material similar to that contained in an authentic recording of a 'target' speaker. They were selected on the basis of close similarity to the target speaker in terms of age, educational background and accentedness, and on the basis of their credibility in terms of the role they were asked to play. The six speakers selected were between 35 and 50 years of age and had a mild to very mild The Hague accent. Special care was taken to avoid bias due to speech content and to speech style as a function of lan-

guage use. The five foils all took part in two approximately 5-minute telephone conversations with a third party, as part of a (fictitious) campaign to recruit representatives for a new company. Prior to the recording of the telephone conversations, the five foils were given two sheets, one describing the aims and organisation of the company and the future activities of the representatives, the other containing a list of keywords and phrases to be used as a prompt during the telephone conversations. The third party was provided with similar information to enable him to pose as a prospective representative. Although the material produced by the foils was therefore neither strictly unrehearsed nor unmonitored, none of the listeners turned out to be aware of this. The telephone conversations were edited to remove the voice of the third party. Four stimulus tapes were produced: The first, Tape 1, consisted of 150 sec. samples of nett speech from each of the 6 speakers (i.e. the five foils and the target). The second, Tape 2, consisted of 60 sec. samples of nett speech different from that used for the compilation of the first tape. Tapes 1 and 2 were used in the voice recognition experiment. The third tape, Tape 3, consisted of two sets of 10 sec. samples taken from the 60 sec. samples on Tape 2, randomly arranged to form two series of 15 pairs. Tape 4 consisted of a single listing of the six 10 sec. samples used for Tape 3.

4. EXPERIMENTS CONDUCTED

4.1 Semantic Scales

In the first experiment, 10 listeners were played Tape 2, consisting of 150 sec. samples of each of the six stimulus voices, and asked to rate the speakers on 16 scales, eight of them referring to speech characteristics and eight to speaker personality characteristics.

4.2 Pairwise Comparisons

In the second experiment, two groups of 5 subjects first listened to Tape 3. They were asked to familiarize themselves with the range of voices on this tape before moving on to Tape 4, containing the 15 randomized pairs. For each pair, they were instructed to express the degree of difference between the voices they heard on a scale from 1 to 10, with 1 standing for 'the same' and 10 for 'very different'.

4.3 Voice Recognition Parade

The third experiment was a voice identity parade. The 150 sec. samples of speakers number 1, 3 and 4 respectively, on Tape 1 were played to three groups of second-year Business Communications students. They were told to pay special attention to the voice they heard rather than to what was being said, as they would be questioned about this voice in a week's time. Exactly one week later they listened to Tape 2, containing the 60 sec. fragments. Prior to this, they were told that they were going to hear six speakers, one of whom might be identical to the one they had heard a week earlier. Contrary to the truth, they were also told that none of the speakers might be identical to the target voice, since the experiment involved several groups and that in some of these the target voice was not on the tape. They were asked to circle the number on their answer sheet corresponding to the number preceding the speaker of their choice and to circle 0 if they judged none of the speakers identical. They were also asked to indicate the degree of confidence in their decision on a 5- point scale.

4.4 Mean F0

The 10 sec. samples of Tape 4, used for the Pairwise Comparison task were also used to arrive at a mean F0-value + standard deviation for each of the 6 speakers, using the SIFT algorithm. The following values were found:

Table 1 Mean F0-values and standard deviations

Speaker	Mean F0 (Hz)	S.D. (Hz)
1	103	16
2	139	24
3	82	18
4	123	19
5	110	34
6	104	15

5. RESULTS

5.1 Semantic Scales and Paired Comparisons

The scale values obtained on the 16 scales were used to calculate interspeaker distances by means of the common

squared Euclidean metric. Three sets of distances were calculated: a) distances based on the complete set of 16 scales, b) distances based on the 8 speech scales, and c) distances reflecting the scores on the 8 personality scales. The calculation of distances on the basis of scale values is a somewhat hazardous affair, as the number of scales covering a specific aspect of the object under investigation may have a substantial effect on the distance obtained. For that reason, factor scores should be used, as these scores are not correlated. In actual forensic practice, however, only a limited number of subjects will normally be available, so that the use of factor analysis is not possible, as the number of variables largely exceeds the number of cases.
The following table shows the correlations between *Dall* (distances based on all 16 scales), *Dsp* (distances based on speech scales), *Dper* (personality scales) and *Diss* (the overall dissimilarities, obtained in the paired comparison test), for all 15 pairs of the 6 speakers involved.

Table 2 Correlations between 3 types of distances and the overall dissimilarities; N= 15 (see text). Significant correlations (p = 0.05) are marked *.

	Dall	Dsp	Dper	Diss
Dall				
Dsp	.98*			
Dper	.95*	.97*		
Diss	.45	.46	.51	

The correlations given above show that the scale-based distances and the overall dissimilarities are not equivalent. We also applied cluster analyses to the distances and dissimilarities. The results provided confirmation for the difference found between the two approaches: paired comparisons vs. the use of semantic scales. So we have reason to believe that the two methods of assessing the homogeneity of a group of subjects are not equivalent.

5.2. Voice Recognition Parade

The results of this experiment were as follows:

Table 3 Identification results (Gr= Group; C.I. = correct identification, F.I. = false identification, F.E. = false elimination).

Gr	Target	C.I.	F.I.	F.E.	N
A	1	91.7%	8.3%	-	12
B	3	90.1%	-	9.9%	11
C	4	69.2%	30.8%	-	13

An analysis of the False Identifications reveals that, while one listener mistook Speaker 4 for Speaker 1, 4 listeners wrongly identified Speaker 1 as Speaker 4. The bias towards Speaker 1 may be due to the order of presentation. Group A, whose target speaker was Speaker 1, heard their target before they could be confused by Speaker 4, while Group C, whose target was Speaker 4, first heard the apparently rather similar Speaker 1.

6. FURTHER CONSIDERATIONS

Given the fact that the two methods do not produce equivalent results, it would obviously be desirable to assess their validity by means of some independent test. One way of doing this would be to correlate the results of the two techniques with the confusion scores of a large number of voice recognition tests involving all six speakers in turn as targets, conducted at various time intervals. So far, only the results presented in 5.1 are available. Unfortunately, it appears that with a one-week interval between presentation and recognition sessions, recognition scores are very high so that a ceiling effect is produced. It is expected that longer delays between presentation and recognition sessions will produce the type of scores that are required to calculate correlations. On present information, we would be inclined to prefer the paired comparison test. It has two distinct advantages over the semantic scale test, one theoretical, the other practical. As we have already observed, the use of semantic scales inevitably involves a certain amount of overlap between the scales used, which may seriously affect the distance indices obtained. The distances obtained with the paired comparison test should provide a more accurate reflection of the dissimilarity of the voices. From a prac-

tical point of view too, the paired comparison test is preferable since it is considerably less time-consuming and labour-intensive. However, it should be noted that the dissimilarity measures obtained in the paired comparison test need to be converted to metrical distances by means of a Multidimensional Scaling Technique. This conversion presupposes a small 'Stress-value', associated with a relatively small number of underlying dimensions.

7. A PRACTICAL PROPOSAL

As discussed in the introduction, the voices used in a voice parade should not constitute too heterogeneous a set. The target voice in particular, should not occupy an outlying position in the perceptual space. Presumably, an ideal situation for a voice parade would be for all voices to be located in equidistant positions on a concentric circle around the centroid of the perceptual space. A rule of thumb might be: the target voice should not be situated at a distance from the centroid greater than the average distance + its standard deviation. If this principle is applied to the distances obtained in the paired comparison test, the result is that illustrated in Fig.1

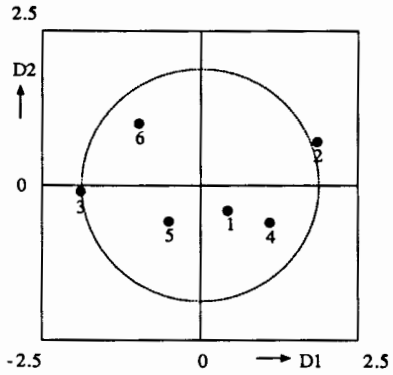


Figure 1 The locations of the 6 speakers in the Perceptual Space

The SPSS-procedure ALSCAL was used; with two dimensions Kruskal's Stress formula 1 was .063 (RSQ = .972). Here, the sum of the average distance plus the s.d. (=1.882) is taken as the radius of a circle around that point. The value of the target voice should not exceed that value. It appears that our target (no 6) is located within the desired distance to the centroid.

8. CONCLUSION

Of the two procedures which may be used to assess the similarity of voices used in a voice line-up, the paired comparison test appears the more promising. Further research to obtain independent support for its validity is in progress.

A final observation concerns the proposed sphere of application of the similarity test. It is clearly not intended as a foolproof procedure which can simply be applied to any random set of voices. It is only when every conceivable effort has been made to avoid bias of any kind in the selection of the voices that the results of the test will be meaningful.

REFERENCES

[1] CLIFFORD, B.R. (1980), "Voice Identification by Human Listeners: On Earwitness Reliability", *Law and Human Behavior*, 4, 373-394.

[2] DEFFENBACHER, K.A. et al. (1989). "Relevance of Voice Identification Research to Criteria for Evaluating Reliability of an Identification". *Journal of Psychology*, 123(2), 109-119.

[3] HAMMERSLEY, R.H. & J.D. READ (1983). "Testing Witnesses' Voice Recognition: Some Practical Recommendations". *Journal of the Forensic Science Society*, 23, 203-208.

THE PERCEPTION OF CONSONANTAL NASALITY IN ITALIAN: CONDITIONING FACTORS

Pietro Maturi

C.I.R.A.S.S. - Università di Napoli
via Porta di Massa,1 - 80133 Napoli Italy

ABSTRACT

A test has been performed with natural and artificial sequences to check the actual role of vowels in the perception of the nasality feature of Italian consonants. The procedure and results are presented and discussed in the present paper.

0. FOREWORD

The Italian phonological system is traditionally said to have a series of nasal consonants but no nasal vowel. Vowels are only said to become nasalized for anticipatory coarticulation in V-C_N or, more often, for carryover in C_N-V contexts. Such vocalic nasalization is given no phonological significance, nor has it ever been investigated whether it has any direct perceptual relevance for the recognition of the V-C or C-V sequence. In a previous research [1] some experiments were made with nasalized and non-nasalized (final) vowels in Italian minimal pairs such as vini - vidi, cane - cade, alma - alba, etc. Such sequences showed a regular nasalization of the final vowel if preceded by a nasal consonant (C_N-V > C_N-V_N), and no nasalization in the opposite case (C_O-V > C_O-V_O). The experiment consisted in an artificial inversion of the final vowels, so to obtain such final sequences as C_N-V_O and C_O-V_N.

The answers given by a group of listeners showed that their perception of the nasality ~ non-nasality feature in the consonants ([n]~[d], [m]~[b]) was very strongly affected by the presence ~ absence of the same feature in the following vowel. This means that in case of inconsistency between the nasality feature present in the consonant and the one present in the vowel, it was surprisingly the latter which used to prevail: so both the [α nasal] vowel and the [-α nasal] consonant were heard as [α nasal], with a somehow asymmetrical behavior, as the effect was general for α = + and apparently underwent some restrictions for α = -.

On the present occasion I will present and describe the results of a new test realized with approximately the same technique as the previous one on a much larger set of Italian minimal pairs, each containing the opposition between an oral and a nasal homorganic consonant, in order to check the results previous-

ly obtained and to relate them to some variables.

1. MATERIALS AND METHODS

The words for the test have been selected according to the following criteria:

a)preceding context: an Italian consonant in pre-final position in a bi- or polysyllabic word can be either preceded by one of the seven vowels [i,e,ɛ,a, ,o,u] or by one of the consonants [z,l,r,m,n]; the vowels [e~ɛ] and [o~] have not been used here because of the very fluctuating use Italian speakers make of such oppositions; with [z] no minimal pair was found; as to nasals, only the homorganic one is accepted, so before alveolars only [n] is used, and only [m] before bilabials; the actual set of preceding contexts used was [a-,i-, u-,l-,r-,N-];

b)following context: in final (unstressed) position only four vowels can be found [-a,-e,-i,-o] as [u] is practically absent in that position and the oppositions [e~ɛ] and [o~] are neutralized;

c)place of articulation: only two pairs of (voiced) oral vs. nasal consonants exist in Italian [b~m] and [d~n], as the pair [g~n] is neither phonological nor can it appear in the examined position.

The product of 6 preceding contexts * 4 following contexts * 2 places of articulation makes 48 potential pairs of sequences. A careful examination of the Italian lexicon showed that only 26 of them are actually employed (not taking into account rare or obsolete words). Here is the list of the theoretical contexts. Those employed in the present research are underlined; only the oral sequences are listed, not their nasal counterparts, which can be easily obtained substituting -n- for -d- and -m- for -b-:

ada	ade	adi	ado
ìda	ìde	ìdi	ìdo
uda	ude	udi	udo
lda	lde	ldi	ldo
rda	rde	rdi	rdo
nda	nde	ndi	ndo
aba	abe	abi	abo
iba	ibe	ibi	ibo
uba	ube	ubi	ubo
lba	lbe	lbi	lbo
rba	rbe	rbi	rbo
mba	mbe	mbi	mbo

A native male Italian speaker of 29 years old uttered the words in our laboratory. The words were recorded on a tape. The splicing of the final vowels from the rest of the word was carried out on the basis of the observation of both the oscillograms and spectrograms of the natural signals, along the conventional segment boarders. The operation of inversion between the final vowels of each pair of words was effected by means of a DSP-Sonagraph 5500 "gating \editing" procedure (such device allows to choose the "cutting" point with an approximation of ±3 ms).

A group of 23 Italian students of foreign languages and literatures between 19 and 27 years old was then asked to listen to both the natural and artificial sequences and to give their judgment about them. The test was organized as follows: each of the 52 words making up the 26 pairs was presented aurally (in headphones) in its natural shape and at the same time

the subjects could read it on a special form prepared for them; the natural stimulus was then followed by two artificial sequences built up with the phonic material of the corresponding minimal pair (e. g. the natural word strada was followed by an artificial stimulus made up with strad- plus the final -a from strana and by one made up with stran- and -a from strada); the order of the natural stimuli was completely random, and so was the order of the two artificial stimuli following the natural one; the listeners were asked to decide which of the two artificial stimuli heard resembled best the natural stimulus previously heard and read.

2.RESULTS AND DISCUSSION

Following strictly the phonological models of Italian, one would expect the substitution of a final nasalized vowel with a non-nasalized one and viceversa to have no effect on the perception of the sequence, as the nasality of the consonant, which is considered the only pertinent manifestation of nasality, is perfectly preserved. So, starting from the natural stimulus strada, one can consider the artificial strad+V_N to be phonologically the "same" as the natural sequence, and stran+V_O a "different" sequence. So, all answers to the test can be classified as "phonological" (+phon) if the "same" stimulus is indicated to resemble best the natural one, and "anti-phonological" (-phon) if the "different" stimulus is chosen.

Globally, the answers given by the students are as follows:

answers		
type	number	percentage
+phon	314	26.2%
-phon	882	73.8%

The results will now be presented and examined according to the variables above listed. A general discussion will follow.

a) Preceding context

	answers		
V-	type	number	percent.
a-	+phon	47	25.5%
	-phon	137	74.5%
i-	+phon	57	24.8%
	-phon	173	75.2%
u-	+phon	83	45.1%
	-phon	101	54.9%
C-	type	number	percent.
l-	+phon	74	40.2%
	-phon	110	59.1%
r-	+phon	36	19.6%
	-phon	148	80.4%
N-	+phon	17	7.4%
	-phon	213	92.6%

As can be seen from the above tables, the number of -phon answers always exceeds the +phon. The most favorable contexts for such effect are the presence of a nasal [N-] and an alveolar vibrant [r-]. Also with [a-] and [i-] the results are pretty good in the direction of an anti-phonological behavior, while after [u-] and [l-] the answers approximate a random distribution of 50%-50%.

b)Following context

	answers		
-V	type	number	percent.
-a	+phon	94	25.5%
	-phon	274	74.5%
-e	+phon	37	13.4%
	-phon	239	86.6%
-i	+phon	158	49.1%
	-phon	164	50.9%
-o	+phon	25	10.9%
	-phon	205	89.1%

In this case, too, some contexts seem to be very favorable to a -phon behavior, such as [-o], [-e] and, to a lesser extent, [-a], and one context, [-i], with a random distribution of the answers.

c)Place of articulation

	answers		
place	type	numb.	percent
alv.	+phon	120	17.4%
	-phon	570	82.6%
bilab	+phon	194	38.2%
	-phon	312	61.8%

The difference is rather large in favour of the alveolar place of articulation, which shows a high percentage of -phon answers; with bilabials the effect is smaller, though still exceeding a casual distribution enough to be considered meaningful.

d)Direction of the effect

	answers		
-C-	type	numb.	percent
oral	+phon	174	29.1%
	-phon	424	70.9%
nasal	+phon	140	23.4%
	-phon	458	76.6%

As indicated above, the results of a previous test had shown an asymmetrical perceptual effect of the inversion of the vowels, so that the sequence C_N-V_O was generally perceived as C_O-V_O, while the perception of C_O-V_N as C_N-V_N was also frequent but not to the same extent. In the present test there is still a slightly higher number of -phon answers for nasal than for oral consonants, but the difference is too little to be meaningful.

The results, on the whole, witness a very strong effect of the vowel's nasality feature on the perception of the preceding consonant. Even the lowest percentages obtained, which still exceed 50%, show that in case of inconsistency between the nasality feature in the consonant and in the vowel, the first does not prevail automatically, as phonologists seem to presume, when they exclude that the nasalization of vowels has any pertinence in Italian. In case of random distribution of the answers, the feature of nasality can be said to have an equal weight in consonants as in vowels. But the general result of the test, and the particular result in most contexts, is that the weight of nasality in consonants and in vowels is not the same, and that the feature of nasality seems to be much more important for vowels than for consonants in Italian.

3.REFERENCE

[1] GIANNINI A., MATURI P., PETTORINO M., "Il ruolo della nasalità nella fonologia dell'italiano", in FUSETTI M.(ed.), Atti del XVIII Convegno Nazionale dell'Associazione Italiana di Acustica, L'Aquila 18-20 aprile 1990, pp.191-6.

PERCEPTION AND PRODUCTION OF A VOICING CONTRAST BY FRENCH-ENGLISH BILINGUALS

V. Hazan and G. Boulakia

Dept of Phonetics & Linguistics, University College London, U.K.
Laboratoire de phonétique du DRL, Université de Paris VII, France

ABSTRACT

The use of F1 onset information, which constitutes a cue to the voicing contrast in English, but not in French, was investigated in French-English bilinguals, classified according language bias. Results show evidence of code-switching in production but not in perception. English-bias bilinguals were more strongly affected by the F1 onset cue than French-bias bilinguals.

1. INTRODUCTION

Perceptual studies with bilinguals investigate whether phonemic categorisation is affected by higher order linguistic information, by presenting a same continuum with different language precursors. The voicing contrast is particularly useful for such investigations, as it is marked differently along the Voice Onset Time (VOT) dimension in languages such as French and English. Results of such studies are contradictory. Some have not found evidence of any language effect on categorisation (eg. [1]) while others have only found evidence of code switching in perception for strong bilinguals [2]. In this study, the effect of language bias was controlled by testing bilinguals both in France and in Great-Britain. Computer-edited natural stimuli were used together with careful test procedures to ensure that subjects were sufficiently induced into a particular language set.

A novel approach was to focus attention on the use by French-English bilingual and monolingual subjects of spectral cues to the voicing contrast. In English, first formant cutback in the vowel following long-lag voiceless plosives contrasts with a rising first formant onset following short-lag voiced plosives while, in French, a rising first formant onset is present after both voiced (lead) and voiceless (short-lag) plosives. F1 onset therefore constitutes an additional cue to the voicing contrast in English but not in French.

2. STIMULI

The /pen/-/ben/ minimal pair was chosen as it is meaningful both in English ("Ben - pen") and in French ("benne - penne"). Test continua were created using digitised natural speech waveforms. In all continua, VOT ranged from -40 ms to +40 ms in 10 ms steps. In the first continuum (Pen/VOT), the [en] portion, burst transient and aspiration were taken from a voiceless [p^{h}en] produced by a male speaker. A "cut and paste" technique was used to create intermediate stimuli. For stimuli with positive VOTs, the aspiration was progressively deleted, in 10 ms slices, following the burst release. For stimuli with negative VOTs, the prevoiced portion was edited out of a voiced [ben], appended to the front of the burst release then cut back in 10 ms steps. In the second continuum (Ben/VOT), the [en] portion from a [ben] token was used. The

same technique as described above was used to obtain the VOT continuum. The two ranges therefore varied in the spectral characteristics of the vowel. In order to create French and English test conditions, each of the stimuli described above was preceded by a precursor: "répète" in the French condition and "repeat" in the English condition. For each condition, an identification test tape was prepared by randomizing and recording ten tokens of each of the nine stimuli.

3. SUBJECTS

Four groups of listeners were tested: 8 bilinguals living in London, 13 bilinguals living in Paris, 11 British monolinguals and 13 French monolinguals. All subjects reported normal hearing.

4. PROCEDURE

Testing was carried out over two one-hour sessions on separate days (one session only for monolinguals). At each session, only one language was used. The session started with a speech recording of "accent-revealing" sentences and of minimal pairs. The Pen/VOT and Ben/VOT stimuli were then presented in two-alternative forced-choice identification tests. Stimuli were presented free-field at a comfortable listening level.

5. RESULTS

5.1 Classification of subjects

Bilingual subjects were classified according to strength of bilingualism and language bias. Strength of bilingualism classification was based on judgments by phonetically trained listeners in France and Great-Britain of the recordings of the "accent-revealing" sentences on a scale of 0 (native) to 5 (foreign). Language bias, seeking to reflect what would be considered the "base language" for a particular bilingual, was determined on the basis of a questionnaire where information was collected on main language spoken with family and friends, at school, etc. There were 10 English-bias bilinguals (4 "mid" and 6 "strong") and 11 French-bias bilinguals (4 "mid" and 7 "strong").

5.2 Production

VOT measurements were made of five repetitions of /pen/ and /ben/ for each speaker in each language. Means were then obtained according to subject group and language mode (Fig. 1). There is clear evidence of code switching in production, even though values obtained for bilinguals differ from monolingual values. The mean values obtained for "mid" bilinguals in their weaker language were furthest from the average monolingual values.

5.3 Perception

Mean labelling functions obtained for monolinguals and for bilinguals grouped according to language bias are presented in Figure 2. The mean phoneme boundary estimates were obtained using a maximum likelihood estimation technique, which fits a cumulative normal function to the data. For the Ben/VOT condition, mean boundary values of +1.4 ms were obtained for French monolinguals and +17.6 ms for English monolinguals. A sizeable phoneme boundary shift was obtained in the Pen/VOT condition for both groups of monolinguals with boundaries of -10.6 ms for English subjects and -22.4 ms for French subjects. Even though F1 onset is not contrastive in French, the presence of "abnormal" spectral characteristics therefore led to a greater proportion of voiceless responses by French listeners with only stimuli with greater than 20 ms of prevoicing consistently labelled as voiced. The labelling function obtained for English monolinguals was less sharp; even stimuli with prevoicing were not consistently labelled as voiced.

For the bilingual groups, each

graph contains four functions representing the labelling of the Pen/VOT and Ben/VOT conditions with French and English precursors. Phoneme boundaries (Table I) obtained for the Ben/VOT condition were similar for the French- and English-bias bilingual groups and intermediate to values obtained for monolingual listeners. There was little evidence of a significant shift in boundary between French- and English-precursor conditions as the 95% confidence interval ranges overlap considerably. For the Pen/VOT condition, English-bias listeners showed a much greater boundary shift relative to the Ben/VOT condition than French-bias listeners. As French monolinguals, French-bias bilinguals were able to consistently label stimuli with greater than 20 ms prevoicing as voiced despite conflicting spectral cues in the vowel. However, their labelling of the continuum was much less categorical than that obtained for monolinguals. Again, for both groups there was little evidence of boundary shift induced by a difference in the language of the precursor.

6. DISCUSSION

There is clear evidence of code-switching in production. The effect of strength of bilingualism was seen as weaker bilinguals showed a less complete shift in VOT between their productions of the French and English contrasts than stronger bilinguals. In perception, little evidence was found to support code-switching. On average, changing the language of the precursor did not generally lead to a significant shift in phoneme boundary, although evidence of code-switching may be found for individual bilinguals. A change in the spectral characteristics at vowel onset did however have a differential effect on labelling according to language bias. There is therefore some evidence for the theory that bilinguals have a "base language" which determines which speech pattern cues are used in perception. Indeed, bilinguals exposed to English early were shown to be more sensitive to changes in spectral characteristics of the vowel and showed this sensitivity both in French and English modes. Further support for a "base" language in bilinguals can be found at a different level of processing. Indeed, Cutler et al. [3] found that only French-dominant bilinguals made use of syllabic segmentation, which is appropriate for French but not English, even though all subjects in the study were strong bilinguals. There is therefore evidence from different sources that even in highly proficient bilinguals, one language dominates in terms of certain aspects of language processing.

7. REFERENCES

[1] WILLIAMS, L . (1977) "The perception of stop consonant voicing by Spanish-English bilinguals", *Perception and Psychophysics, 21*, 289-297.

[2] ELMAN, J.L., DIEHL, R.L. & BUCHWALD, S.E. (1977) "Perceptual switching in bilinguals", *J. Acoust. Soc. Am., 62*, 971-974.

[3] CUTLER, A., MEHLER, J., NORRIS, D., SEGUI, J. (1989) "Limits on bilingualism", *Nature*.

Table I: Phoneme boundary measures (ms VOT)

	Mean	95% conf.int.
FRENCH-BIAS BILINGUALS		
Ben/VOT E	11.5	6.9 → 15.9
Ben/VOT F	7.7	2.6 → 12.9
Pen/VOT E	-9.8	-15.8 → -3.8
Pen/VOT F	-14.7	-17.9 → -11.5
ENGLISH-BIAS BILINGUALS		
Ben/VOT E	8.0	* *
Ben/VOT F	9.7	-1.2 → 19.4
Pen/VOT E	-20.4	-32.2 → -12.1
Pen/VOT F	-21.4	-27.5 → -16.5

* not estimable

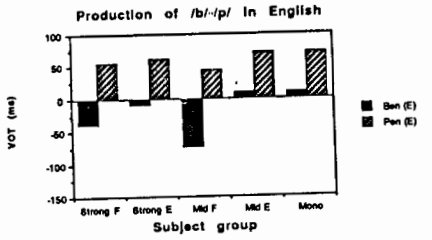


Figure 1: Mean VOT measurements for productions of /pen/ and /ben/ by monolingual speakers and bilingual speakers in each language.

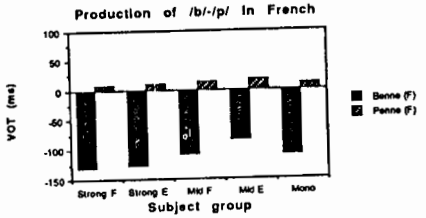
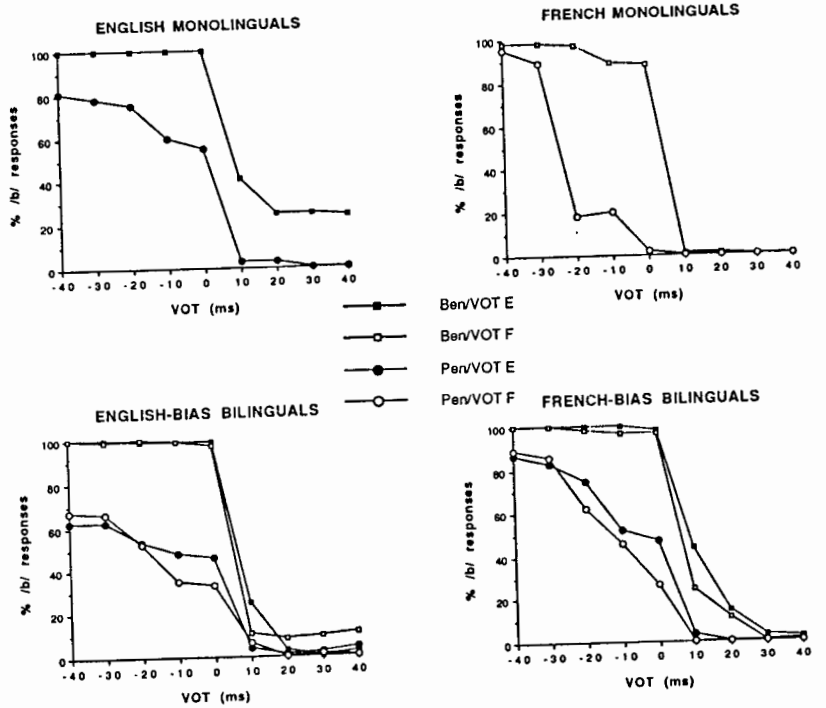


Figure 2: Mean labelling functions for monolinguals and bilinguals grouped according to language bias.



TEMPORAL CUES IN THE PERCEPTION OF THE VOICING CONTRAST IN RUSSIAN

S. M. E. Barry

Department of Psychology and Speech Pathology,
Manchester Polytechnic

ABSTRACT

The perceptual role of several temporal characteristics as cues to the voicing contrast in Russian is investigated. The following parameters were examined: duration of the closure, duration of the preceding vowel and duration of voicing during the closure. In identification experiments, all three factors contributed to cue the contrast. The duration of closure voicing was an important cue. Consonant and vowel duration contributed either when they were co-operating with the voicing cue or when the voicing cue was ambiguous. In addition, the results support a model where the absolute duration of voicing, rather than voicing duration relative to closure duration, is perceptually relevant.

1. INTRODUCTION

The relevance of temporal cues to the voicing contrast has been widely investigated, both with regard to English (closure voicing and consonant duration [5], vowel duration [8], the consonant/vowel ratio [2, 7], consonant and vowel duration as independent cues [6]), and, for example, French [3] and German [4]. This paper extends the investigation to Russian, testing temporal differences which were found in production [1]. It was found that, as in English, voiced consonants were shorter than their voiceless counterparts and the vowel preceding the consonant longer. The present perception experiments were designed to investigate whether these differences could be perceptual cues to the contrast, and if so how they interact with a third cue: the duration of voicing during the closure.

2. METHOD

2.1 Stimuli

Two pairs of tokens were selected from recordings previously analysed [1]. The tokens were /rota/ ['rɔtə] (military company) and /roda/ ['rɔdə] (sort, gen.sg.). A pair was taken from two female speakers' recordings (referred to below as Set 1 and Set 2 respectively). In Set 1, the /roda/ token contained voicing at the /d/ burst, as the stop was fully voiced; this cycle at the burst was not cut out in the editing (see below); in Set 2 voicing at the burst was absent, voicing during the closure having ceased before the burst.

The vowel and stop durations (in msec) of the tokens are as follows:

	Set 1		Set 2	
	stop	vowel	stop	vowel
/roda/	65	168	81	165
/rota/	113	145	114	144
difference	+48	–23	+33	–21

note: stop duration refers here to the duration of the hold phase only.

The four original tokens (sampled at 20kHz) were digitally edited as follows:
a) /t/ in /rota/ was shortened: to 65 msec in Set 1 and to 81 msec in Set 2.
b) /d/ in /roda/ was lengthened: to 113 msec in Set 1 and 116 msec in Set 2.
c) /o/ in /rota/ was lengthened by repeating a section of the steady state vowel: to 169 msec in Set 1 and 165 msec in Set 2. The stop duration was kept as in the original token.
d) /o/ in /roda/ was shortened by cutting out a section of the steady state vowel: to 144 msec in Set 1 and 143 msec in Set 2. The stop duration was kept as in the original token.

In each of these eight stimuli, therefore, one cue has been introduced which conflicts with the other information present in the token. A comparison of responses to the edited stimulus and to the unedited token will show the effect of this alteration.

To test the interaction of stop or vowel duration and voicing during the closure, in each of the above 8 stimuli, together with the 4 original tokens, the closure voicing was edited out in stages, in the case of /roda/ stimuli, or added in in stages, in the case of /rota/ stimuli. 12 series of stimuli were thereby created, each with a gradually increasing duration of voicing during the closure.

The series resulting from an original /rota/ token are referred to below as /rota/ stimuli, those from a /roda/ token as /roda/ stimuli.

2.2 Procedure

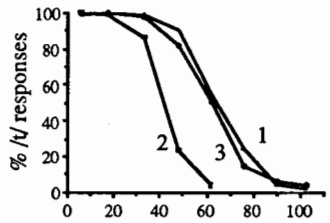
There were 10 presentations of each of the 78 stimuli. The experiment was in two parts: Set 1 and Set 2 stimuli being presented separately. Within each part the presentations were randomised, and preceded by 10 other words recorded by the same speaker. The experiment was carried out in Moscow, with 37 subjects who were native Russian speakers aged 17-40 resident in Moscow (18 female, 19 male). The subjects heard the stimuli played on a cassette recorder in quiet surroundings, 27 with headphones, 10 without. The format was a forced choice identification task, and subjects wrote their responses on prepared answer sheets.

3. RESULTS AND DISCUSSION

3.1 Voicing during the closure

Figures 1 and 2 present the mean responses to the /rota/ stimuli Set 1 and 2 and the /roda/ stimuli Set 2. It can be seen that voicing during the stop is an important cue. Full voicing leads to approximately 100% /d/ responses. Absence of voicing leads to 100% /t/ responses in the case of /rota/ stimuli, and in the case of /roda/ stimuli to 47.7%, 60% and 80.6% /t/ responses, depending on stop and vowel duration. Full voicing is, not unexpectedly, a sufficient cue to override all others, including in the case of /rota/ stimuli all other cues to a /t/ present in the original token. The absence of voicing does increase /t/ responses but is not, at least in the case of this particular token, a sufficient cue to a /t/.

a. /rota/ stimuli Set 1



b. /rota/ stimuli Set 2

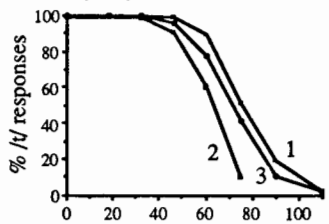


Figure 1: Responses to /rota/ stimuli with 1=original stop and vowel duration, 2=shortened stop, and 3=lengthened vowel.

22 of the 37 subjects gave a 100% /d/ response to the /roda/ Set 1 stimuli, even when the consonant or vowel duration had been altered and there was no voicing during the closure. Only two subjects gave a 50% or more /t/ response to any stimulus. The most likely characteristic leading to this almost total /d/ response is the presence of voicing at the burst.

For one subject only, there was an 80-100% /t/ response to all /rota/ stimuli, even with consonant or vowel duration changed. For this subject, at least in the case of these particular tokens, closure

voicing is not a sufficient cue to a /d/ and is being overridden by another cue or cues to a /t/ in the token. His results are excluded from the mean results for the /rota/ stimuli presented below.

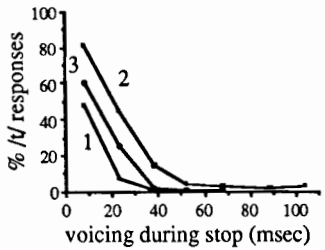


Figure 2: Responses to /roda/ stimuli Set 2 with 1=original stop and vowel duration, 2=lengthened stop and 3=shortened vowel.

3.2 Vowel duration

Figures 1a and b show a small decrease in /t/ responses to /rota/ stimuli with a lengthened vowel, where closure voicing is ambiguous. There is a small difference in the 50% crossovers between the two labelling functions: of 3.1 msec (Set 1) and 4.8 msec (Set 2). In both cases t-tests for paired samples show this difference to be significant (Set 1: $t = -3.40$, $p = 0.002$; Set 2: $t = -6.62$, $p < 0.001$).

Figure 2 shows that in /roda/ Set 2 stimuli the shortened vowel has an effect on responses where the closure voicing has been almost completely edited out. The combination of vowel duration and absence of voicing is a stronger cue than the absence of voicing alone, although not a sufficient cue to completely override others in the signal. In the case of the stimuli with 8 msec closure voicing, the difference in /t/ responses is 12.6%, which a paired samples t-test showed to be significant ($t = 4.89$, $p < 0.001$).

It is possible that the difference in responses is small because the amount by which the vowel duration had been altered was small. The alterations were this size, however, in order to be representative of the differences found in production.

3.3 Stop duration

In the case of /rota/ stimuli, the change to a /d/ percept occurs with less closure voicing when the stop is shortened than when it is not (figures 1a and b). The difference in the 50% crossover points was 23.2 msec (Set 1) and 14.1 msec (Set 2), which in paired samples t-tests were significant (Set 1 $t = -19.45$, $p < 0.001$; Set 2 $t = -12.62$, $p < 0.001$).

A possible interpretation of these results is that the shorter consonant duration is an additional cue to a /d/. However, as shortening the stop automatically increases the proportion of the stop which is voiced (for a given absolute duration of voicing), it is possible that the difference in response is due rather to the listeners using the voiced proportion of the stop as a cue, and not the absolute duration of voicing [3]. To test this, the responses were analysed as a function of the proportion of the closure which was voiced (figure 3).

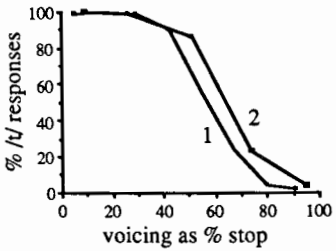


Figure 3: Responses to /rota/ Set 1 stimuli as a function of the percentage of the stop which is voiced (1=original stop duration 2=shortened stop).

The change to a /d/ response is now later (that is, when there is a greater voiced quotient) when the stop is shorter than when it is the original length. This was the case for Set 2 stimuli also. The difference between the 50% crossover points was 7.1% in Set 1 and 10.0% in Set 2, and was significant (Set 1 $t = 6.49$, $p < 0.001$; Set 2 $t = 9.54$, $p < 0.001$).

If the voiced quotient was the sole factor determining the response, the two labelling curves would be identical. This significant difference between the curves is unlikely to be due to the shorter consonant duration itself contributing, as the difference is in the opposite direction to that expected from production findings: here, a shorter stop appears to require a

greater voiced quotient to be perceived as a /d/. If, however, it is the absolute duration of the voicing, and not the voiced quotient, which is relevant in the perception of the contrast, the difference in figure 3 would be explained: a stimulus with the shorter stop will have a greater voiced quotient than a stimulus with equal absolute duration of voicing but a longer stop.

If it is the absolute duration of voicing which is perceptually relevant, figure 3 would not contradict the hypothesis that a decreased stop duration is contributing to a /d/ response: there would be no direct evidence as to whether stop duration is relevant or not.

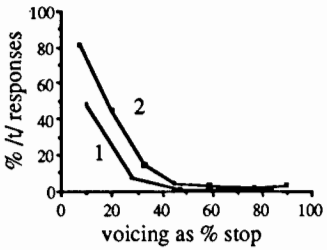


Figure 4: Responses to /roda/ Set 2 stimuli as a function of the percentage of the stop which is voiced (1=original stop duration, 2=lengthened stop).

If it is correct to conclude that the absolute duration of voicing is a cue, and not the voiced quotient, an explanation for the difference in the curves in figures 1a and b (voicing in msec) is that the difference shows the contribution of the decreased stop duration to the perception of a /d/.

This conclusion is supported by responses to the /roda/ stimuli (figures 2 and 4). Whether closure voicing is analysed in terms of proportion of closure duration (figure 4) or absolute duration (figure 2), a longer stop leads to more /t/ responses when there is little voicing during the closure. The difference between the two labelling functions in /t/ responses to stimuli with 8 msec voicing (32.6%) was significant in a paired samples t-test ($t = 6.99$, $p < 0.001$). The difference between the labelling functions when 10% of the closure is voiced is also significant (difference 24.4%, $t = 5.86$, $p < 0.001$). Stop duration is therefore contributing to cue a /t/, however voicing is analysed.

4. CONCLUSION

The duration of closure voicing is for most of the above subjects an important cue to the voicing contrast in Russian, although not in all cases does it override other cues present in the signal. The analysis supports a model in which the absolute duration of voicing is relevant and not the duration of voicing in proportion to that of the stop.

Stop and preceding vowel duration are additional cues to the voicing contrast in Russian. They have been analysed in this paper as independent factors; further analysis will investigate whether or not the results support a quotient model in which consonant to vowel ratio is the relevant perceptual cue.

5. REFERENCES

[1] BARRY, S.M.E. (1988), "Temporal aspects of the devoicing of word final obstruents in Russian", *Speech '88: Proc. 7th FASE Symposium*, 81-88.

[2] DENES, P. (1955), "Effect of Duration on the Perception of Voicing", *J. Acoust. Soc. Amer.*, *27*, 761-764.

[3] VAN DOMMELEN, W. (1983), "Parameter Interaction in the Perception of French Plosives", *Phonetica*, *40*, 32-62.

[4] KOHLER, K.J. (1979), "Dimensions in the Perception of Fortis and Lenis Plosives", *Phonetica*, *36*, 332-343.

[5] LISKER, L. (1978), "Rapid vs. Rabid: A Catalogue of Acoustic Features That May Cue the Distinction", *Haskins Lab. SR-54*, 127-132.

[6] MASSARO, D.W. & COHEN, M.M. (1983), "Consonant/vowel ratio: An improbable cue in speech", *Perception & Psychophysics*, *33*, 501-505.

[7] PORT, R.F. & DALBY, J. (1982), "Consonant/vowel ratio as a cue for voicing in English", *Perception & Psychophysics*, *32*, 141-152.

[8] RAPHAEL, L.J. (1972), "Preceding Vowel Duration as a Cue to the Perception of the Voicing Characteristic of Word-Final Consonants in American English", *J. Acoust. Soc. Amer.*, *51*, 1296-1303.

THE CONTEXT SENSITIVITY OF THE PERCEPTUAL INTERACTION BETWEEN F0 AND F1

Hartmut Traunmüller

Institutionen för lingvistik, Stockholms universitet,
S - 106 91 Stockholm, Sweden.

ABSTRACT

According to a known hypothesis, the perceived degree of openness in vowels is given by the CB-rate difference (tonotopic distance) between F1 and F0. Synthetic vowels and diphthongs with non-stationary F_0 and/or F_1 were used to find out whether it is the instantaneous F_0, its average, or the prosodic baseline, that is relevant here. Most subjects behaved in accordance with the basic hypothesis, but some attached a smaller weight to F_0. The results support the relevance of the prosodic baseline as well as that of the instantaneous value of F_0. Between speaker differences in behaviour were prominent.

1. INTRODUCTION

It is well known that the phonetic quality of phonated vowels, in particular their perceived degree of openness, or vowel "height", depends not only on the frequencies of their formants but also on their F_0. According to one hypothesis, the perceived openness is given by the tonotopic distance (CB-rate difference) between F1 and F0 [6]. Data on F_0 and the formants of vowels produced at different degrees of vocal effort and by speakers with differently sized vocal tracts are largely compatible with such an hypothesis [5, 7]. It is, however, still in question whether it is the instantaneous F_0, its average, or some other kind of context dependent reference value that is relevant here.

The tonotopic distance hypothesis was first proposed to explain the results of perceptual experiments with synthetic vowels [6]. Its quantitative validity has been questioned on the basis of results obtained in another perceptual experiment, in which the influence of F_0 turned out to be smaller [4]. The discrepancy can be explained if it is assumed that listeners relate F_1 to the prosodic baseline rather than to an instantaneous or average value of F_0 [8]. Such a baseline is obtained by interpolation between successive minima in the F_0-contour of the breath-group in question.

Data on F_0 in different styles of speech show that an invariant minimal value of F_0 is characteristic of each speaker [9]. That value of F_0 is normally reached close to the end of statements. It appears to be stable in various types of paralinguistic variations, such as the degree of involvement [1] and in different styles of speech [2, 3], at least as long as these do not involve an overall change in vocal effort. More precisely, the invariant value of F_0 is slightly above its minimum, and it might represent an average of the baseline.

If this is to be reflected in speech perception, listeners should, *in effect*, relate F_1 to an estimate of the speaker's prosodic baseline in judging vowel openness. According to slightly different hypotheses, the minimum F_0 in the whole breath-group or in a smaller unit of speech might be relevant instead. In order to test the various hypotheses, an experiment was performed with synthetic vowels and diphthongs in which either F_1 or F_0 varied or both varied in unison.

2. METHOD

2.1 Stimuli

The stimuli were synthesized digitally by means of a terminal analog of the vocal tract, using a three-parameter voice source and 8 formant filters in cascade. The excitation signal used imitated that observed, by inverse filtering, in a vowel produced by a woman. Thus, F_0 followed a natural intonation contour. The nominal F_0-values referred to in the following are amplitude weighted mean values. These were 161, 250, 347, 453, 569, and 697 Hz, representing steps of 1 Bark. The stationary positions of F_1 were 250, 347, 453, 569, 697, and 838 Hz. The formants above F1 were in all stimuli invariably at the following positions in Hz: 2 220, 3 406, 4 434, 5 050, 5 741, 6 785, 7 829.

The stimuli had a duration of 470 ms. Prospective diphthongs were obtained by frequency modulation of F1 and/or F0 with part of a sinusoid with a period of 360 ms, phased such that the nominal target values of F_1 and F_0 were reached 30 ms after the beginning and 80 ms before the end of the stimuli. The asymmetry was motivated by a final decrease in excitation amplitude.

The nominal F_0-targets for the diphthongs were 250 and 453 Hz (stimulus series 3a and 3b), 161 and 569 Hz (4a and 4b), and 250 and 347 Hz (5a and 5b). The targets of F_1 were in each sereis 1 Bark above those of F_0.

2.2 Subjects

The stimuli were listened to and transcribed phonetically by 20 subjects, recruited among the personnel and students of the institute. Their first languages were Swedish (12), German (2), Finnish, Estonian, Russian, Bulgarian, English, and Portuguese (1 each). The subjects reported no hearing disorders and they claimed good vocal proficiency in 4.7 languages, on average.

2.3 Procedure

The stimuli were presented binaurally through headphones in 8 series with 6 (first two series only) or with 9 stimuli each, as follows: (1) nominal F_0 = 161 Hz, F_1 rising in steps of 1 Bark. (2) Both F_0 and F_1 rising in steps of 1 Bark. The remaining series (3a to 5b) contained stimuli in which both F_0 and F_1 varied between the chosen target values. Each of these six series included also one sample of each combination of stationary target values: F_0 low, F_1 low; F_0 low, F_1 high; and F_0 high, F_1 high. Series a and b differed only in the order of presentation.

3. RESULTS

The stimuli were predominantly heard as front unrounded vowels with or without diphthongization. In some cases subjects heard front rounded vowels. The responses were computed according to the associated degree of openness as follows: [i y]: 1, [e ø]: 2, [ə]: 2.5, [ɛ œ]: 3, [æ]: 4. For diacritical marks "more (less) open" 0.5 was added (subtracted). In order to accommodate various diphthongs, the responses were quantified using four subsequent values according to the following model: [e]: 2222, [ej], [eˈ]: 2221, [ei]: 2211, [ᵉi]: 2111.

Fig. 1 shows the average perceived degree of openness in the vowel series with subsequently rising F_1 with and without rising F_0 (series 1 and 2). The last one of the four values assigned to each response was ignored. The vowels with the same F_0 and the same higher formants, but with subsequently rising F_1 were unanimously perceived as subsequently more open, from [i] to [æ] (upper line). The spread in perceived

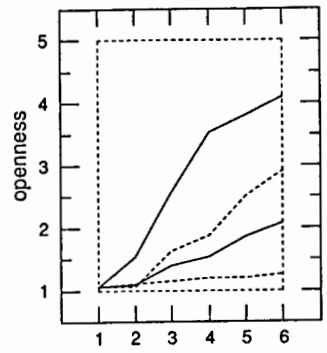


Fig. 1

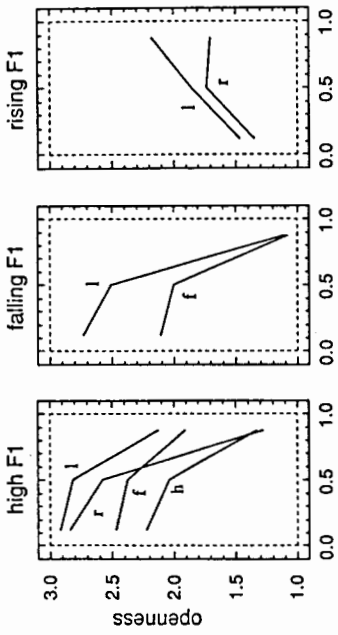


Fig. 3 (3a 3b 3c)

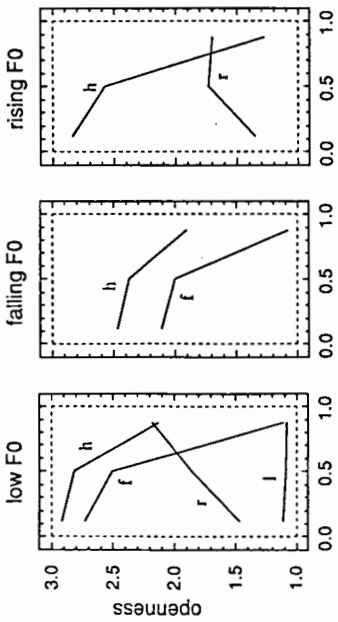


Fig. 2 (2a 2b 2c)

openness was small. As for the stimuli in which both F_1 and F_0 increased (lower line), ten subjects perceived essentially no change in openness, hearing all as [i] (lower dashed line), while the other ten were less uniform in behaviour. For them, only about 40 to 70 % of a shift in F_1 was compensated by an F_0-shift equal in Bark (upper dashed line). The subjects behaved in a unanimous fashion only up to F_0 = 250 Hz.

Fig. 2 shows the effect of variations in F_1 on the perceived degree of openness as a function of time from the beginning to the end of each stimulus in the series 3 to 5. The two non-terminal openness values have been averaged for these figures. The figure shows the results pooled over all subjects and over all three choices of extreme values for F_1 and F_0. There was no noticeable difference between the two orders of presentation. Fig. 2a includes the four cases in which F_0 was low, while F_1 was low (l), rising (r), high (h), and falling (f). In Fig. 2b, F_0 is falling, while F_1 is either high or falling. In Fig. 2c, F_0 is rising, while F_1 is either high or rising.

Fig. 3 is analogous to Fig. 2, but it shows the effect of variations in F_0 when F_1 is given. Fig. 3a includes the four cases in which F_1 was high, while F_0 was low (l), rising (r), high (h), or falling (f). In Figs. 3b and 3c, F_1 is rising and falling, respectively, while F_0 is either low or rising and falling with F_1.

The stimuli in which F_1 and F_0 were "stationary" were often heard as finally diphthongized. This tendency is exaggerated in the results, since even a slight degree of closing diphtongization in open vowels was often transcribed as [V^i] or [Vj].

4. DISCUSSION

The results of the first experiment show that the typical listener behaves quite precisely in accordance with the tonotopic distance hypothesis. The results of the large group of listeners who appear to attach a smaller weight to F0 are troublesome. Considering the quite high degree of naturalness of the stimuli, these results tell us that there will be large between speaker discre-

pancies in perceived phonetic quality even in natural speech produced at high vocal effort, in particular by children, and in soprano singing. As for the age-conditioned variation *per se*, which is also reflected in an approximately uniform shift in F_0 and F_1, there is a cue to vocal tract size in the formants above F2, which is likely to reduce between listener variation for that case.

Fig. 3 demonstrates clearly that the instantaneous F_0 (or a short time average) is of some importance. If the subjects were only sensitive to F_0 averaged over the whole stimulus, the contours in each panel would run in parallel, with a vertical displacement. If they were only sensitive to the F_0-minimum within each stimulus, the contours in each panel, except h in 3a, would coincide. If they were only sensitive to the baseline, all contours would coincide within each panel. On average, the data show a combination of baseline and instantaneous effects, the relative weight of the latter increasing from 0.36 to 0.68 during the course of the stimuli, but this does not hold for each subject.

The responses of individual subjects to the stimuli of Figs. 2 and 3 are not generally predictable from their responses to those of Fig. 1. This is shown in Fig. 4, in which the F_0-sensitivity (in % compensation) in the two types of context is shown for each subject. The comparison includes only the stimuli with "stationary" F_0. The correlation between the two sets of data is low (0.44). There were some subjects who relied entirely on the instantaneous F_0, while others relied on the baseline. The former appear along the diagonal ($y = x$), the latter where $y = 0$. Thus, between speaker differences turned out to be very prominent.

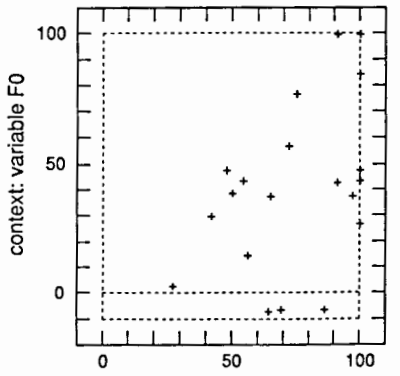


Fig. 4

In a previous experiment, in which the perception of F2' was in focus, it was also observed that some subjects behaved consistently in agreement with the tonotopic distance hypothesis, while others showed a reduced influence of F_0 and often a less consistent behavior [10]. The proportion of the latter was lower among speakers of Swedish than among speakers of Turkish. Apparently, it had been still lower in speakers of Austrian German [6]. This might then be correlated with functional load: The *minimum* number of openness distinctions which are necessary to describe the phonological distinctions in the vowel systems is two for Turkish, three for Swedish, and four for Austrian German. As for the balance between instantaneous F_0 and its baseline, the functional load of tone might be of importance, but there are no data to substantiate such a hypothesis.

5. REFERENCES

[1] Bruce, G. *Working Papers 23* (1982) 51–116, Dep. linguist., Lund univ.

[2] Graddol, D. and

[3] Johns-Lewis, C. in *Intonation in Discourse*, C. Johns-Lewis (ed.), Croom Helm, London & Sidney, 1986, pp. 221–237 and 199–219.

[4] Nearey, T. M. *JASA 85* (1989) 2088–2113.

[5] Syrdal, A. K.; Gopal, H. S. *JASA 79* (1986) 1086–1100.

[6] Traunmüller, H. *JASA 69* (1981) 1465–1475.

[7] " *Phonetica 45* (1988) 1–29.

[8] " *JASA 88* (1990) 2015–2019.

[9] Traunmüller, H.; Branderud, P.; Bigestans, A. *PERILUS X* (1989) 47–64. Inst. linguist. Stockholm univ.

[10] Traunmüller, H.; Lacerda, F. *Speech Comm. 5* (1987) 143–157.

PROBLEMS OF TRANSCRIPTION AND LABELLING IN THE SPECIFICATION OF SEGMENTAL AND PROSODIC STRUCTURE

Martine Grice and William Barry

Department of Phonetics and Linguistics, University College London

ABSTRACT

A consensus transcription by two independent phoneticians of four speakers' readings of a two-minute passage was compared with the hand-labelling of the same recordings. A broad phonetic level of transcription was employed as this level of representation is usual in speech technology applications; the same symbol inventory was used for labelling. A number of differences between transcription and labelling are discussed with reference to the theoretical problems of relating auditory symbolic representation to the acoustic signal; the mapping of transcribed elements onto temporally defined acoustic segments is less than straightforward.

1. INTRODUCTION

There is a long tradition behind a number of internationally established conventions for the auditory specification of the segmental structure of utterances. In the case of prosodic structure, the conventions are of a more language-specific and theory-bound nature. With the growing availability of digitised speech recordings, facilities for representing the speech signal graphically, and the urgent need for large annotated speech databases, these conventions are being challenged. Labellers are faced with the task of relating two phenomenologically different manifestations of speech: a symbolic representation using transcription symbols and a two dimensional transformation of the physical speech signal on the screen.

This paper addresses this theoretical dilemma. The cases of segmental structure examined include a)voiceless schwa b) the occurrence of glottal stop or constriction and c) linking and syllabic /r/. Though these are only a small number compared to the speech sounds transcribed and consistently segmented and labelled in these recordings, they are the product of normal articulatory processes, not freak events which could be safely ignored in the description of normal continuous speech. While they disappear as "noise" in the training distributions of stochastically based speech-recognition systems, they are in fact signal properties which, consistently labelled, could provide additional structural information for the generation of phonological rules. The prosodic investigation focusses on tone-group demarcation in the transcription: the consistency with which transcribers place tone-group boundaries and the relationship of these boundaries to labelled pauses.

2. SPEECH MATERIAL AND ITS REPRESENTATION

For the segmental analysis, four recordings of approximately two minutes each of the "Numbers Passage" from EUROM.0, the first CD-ROM database of the Esprit Project 2589 (SAM) were transcribed by two trained and experienced phoneticians. They were also hand-labelled by a number of SAM research assistants and cross-checked by the authors.

The level of detail given in the auditory transcription can be described as "broad phonetic". Only symbols available from the phonemic inventory of Southern Standard English were used, but changes to the phonemic structure of words resulting from continuous speech processes were captured, e.g. [k@m bi] for "can be" or [b{g k@Uld] for "bad cold". On the other hand, similar assimilatory pro-

cesses such as dentalisation of /n/ in sequences such as "on the" or "in that" are not distinguished from regular alveolar realisations. Syllabic sonorants were represented (e.g. [=m, =n, =l]).

The same inventory was employed for labelling which was based on visual scrutiny of the speech pressure waveform and simultaneous auditory examination of sections no shorter than one syllable in length.

3. TRANSCRIPTION VS LABELLING

3.1. Segmental issues

The discussion concentrates on those aspects of symbolic description which are sensitive to a) the difference in size of processing frames generally available in auditory transcription compared to those used at a labelling workstation, and b) the discrepancy between the transcription goal of merely perceiving a string of appropriate sounds and the need in labelling to annotate every part of the signal and to associate each element in the symbol string with a discrete signal segment.

a) Voiceless schwa?

The possible elision of schwa in extremely reduced syllables is well accepted for English (e.g."proprietary" ending in [t@ri] or [tri] [2]), but it is not usually given as a possible process in the weak form of "to". However, in the many occurrences of the phrase "to the", there were several cases of the /t/ release merging into the following interdental fricative /D/. Perceptually, the impression was of two, albeit extremely short, syllables, so the Broad Phonetic transcription was, logically, [t@D@]. The labelling decision was just as clear: in the case of stop-vowel sequences, the release burst and associated following frication are normally counted as part of the stop, periodicity in the signal is necessary for the labelling of a vocalic segment. Consequently, the absence of a part of the signal on which to attach a vowel label led to the labelled sequence [tD@].

Although a narrower phonetic transcription might have shown the schwa to be voiceless, it is doubtful whether, given the disyllabic percept, and the occurrence of a true vowel very soon after, that the voicelessness of the first syllable can be perceived in the normal flow of the phrase. Only by isolating the two syllables with a speech editor does this become apparent. The artificiality of the segmentation criteria is readily apparent; post-release aspiration, however short, has the function of identifying both a vowel and a consonant. In contrast, devoicing of liquids in stop-liquid sequences is a generally accepted phenomenon (e.g. in "place", "try"). Here transcribers and labellers can be in complete agreement: periodicity is not required, so the frication following the stop burst may be labelled as part of the liquid.

b) Glottal stop or constriction?

The glottal stop is not part of the phonemic inventory used for transcription and labelling. However, glottal stops are clearly audible in the recordings. It is well known that the glottal stop has a number of functions in English. It can reinforce voiceless plosives, replace /t/, and mark the vocalic onset of a stressed syllable. These three different functions correspond, at least in theory, to different manifestations in the speech signal: Reinforcing /p, t, k/, it occurs just before the period of silence resulting from the stop closure [2]. In both of the other two cases it can occur between vowels; glottal onset occurs at the beginning of *stressed* syllables and glottal /t/ replacement alone only occurs before *unstressed* vowels. The combination of glottal /t/ replacement + glottal vowel onset is differentiated from glottal vowel onset alone by the checked nature of the preceding vowel.

Thus, perceptually there are no problems in identifying them, and in a narrow transcription task their occurrence could be recorded. In a labelling task, however, they present manifold problems, since they cannot be treated consistently. A perceived glottal stop may consist of a period of total glottal closure similar to a stop closure, but more frequently it is characterised in the speech signal by either of the following:

i) a rapid increase in the duration of the laryngeal period, and, after a number of very long periods, a return to normal duration;

ii) a reduction in initial peak amplitude over one or two cycles, often accompanied by irregular period duration.

We use the cover term "glottal constriction" for the above two cases when there is no "glottal stop" as such.

Preceding a period of silence, lengthened or irregular glottal periods can arguably be considered part of the preceding vowel or sonorant (since the spectral properties are unchanged). They may, however, signal the preparation for a glottal onset of the following vowel (and are therefore a cue to the ensuing glottal closure), the reinforcement of a stop consonant (the closure phase of which constitutes the silence) or the replacement of a /t/ (where the silence is due to a glottal closure). All three can be auditorily distinguished and do not pose problems for the labeller at the broad phonetic level since the only symbol type permissible is a plosive which can be conveniently attached to the silence; the glottal reinforcement of vowels does not have to be marked.

In the case of /t/ replacement, there are problems with glottal constriction; the amplitude and period irregularity in the signal may constitute the only part of the signal which can be associated with the /t/. Furthermore, it does not need to be very accurately placed to result in an acceptable percept. It may be regarded as an "overlay" on a slowly changing vocalic or sonorant sequence, and it does not always occur in the transitional phase. In one extreme case, "point zero" was produced with the glottal /t/ replacement occurring at the *beginning* of the [n] segment (see figure 1). The most accurate labelling of the event was [poIt=n], although this aberrant sequence was only perceivable after isolation with the speech editor.

c) Linking and syllabic /r/

/@r/ and /r@/ sequences resulted in a number of transcription-labelling divergencies. They were consistently transcribed as diphonic, while in seven cases they were labelled as [=r]: in "numerals", "score and" and "number eight". The case of "numerals" is similar to that of words such as "preference" and "different", given in the literature [3] where compression can also occur. Two labelled versions were found: [njum=r@lz] and [njumr@lz], depending on whether two or three syllables were perceived. Syllabic /r/ is not generally said to occur as a manifestation of linking /r/. In "score and" and "number eight" it was found for both /r@/ and /@r/: [skO:=rn] and [nVmb=r eIt]. The nature of schwa and /n/ make their segmentation in the speech signal arbitrary, even when a schwa is clearly present. In the seven cases recorded as [=r], there was no schwa portion distinguishable from the /r/, and the /r/ was confirmed as syllabic by listening to it with only the preceding context (using the speech editor). Listening to the syllable sequence in context (transcription mode) reduced the clarity of the [=r] percept.

3.2. Tone-group demarcation

Common to most systems of intonational representation is the provision for intonational boundaries of some kind. Within the British approach, such units consist of the tone-group, containing an obligatory nucleus. Since the only prosodic labelling of Eurom-0 presently available is the specification of pauses, the consistency across transcribers in placing these boundaries and the relationship of these boundaries to labelled pauses were examined. To this end, eight experienced British phoneticians were asked to transcribe tone-group boundaries in the first two paragraphs of one speaker's rendition of the numbers passage.

Of the 43 boundaries which appeared in the transcripts, 32 were unanimously transcribed. 26 of these coincided with

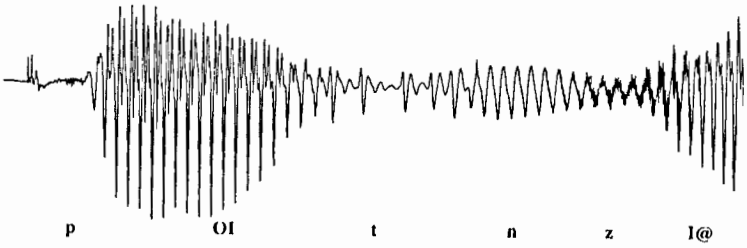


Figure 1 Glottal replacement of /t/ in the phrase "point zero"

pauses (>50ms) in the signal and six had complete agreement without a signal pause. Discrepancies in the transcriptions were found at 11 locations (none with signal pauses).

Four cases of disagreement occurred in the sequence "two, three, four, five, six, seven, eight, nine". Its tonal structure can be informally described as an alternating sequence of relatively high and relatively low shallow rises of the type:

tw^{o} thr_{ee} fou^{r} fiv^{e} si^{x} sev^{en} etc.

Two of the transcribers grouped together sequences of "high" rises followed by "low" rises and six marked a tone-group boundary between each item in the sequence.

The tonal and rhythmic patterns in the sequence suggest that there is a need for capturing a hierarchical structure beyond that of tone groups and intonational paragraphs. The list of digits might most satisfactorily be seen as a pattern of minimal, single-word units, co-ordinated, in turn, within larger ones. Whether the single-word units are below that of the tone group (cf Beckman and Pierrehumbert's intermediate phrase[1]) or above, is an issue for further theoretical discussion. What matters is that, for the transcription of continuous speech, a variety of prosodic boundary types might engender greater consistency amongst transcribers, a necessary prerequisite for the investigation of correspondences between transcribed boundaries and properties of the signal.

It is clear that the elements in the sequence are syntactically linear since they are components of a simple list. It would be interesting to ascertain whether transcribers would be more consistent in placing boundaries when a syntactic hierarchy is apparent such as in the phrase "red book, green pen, brown desk" where a similar tonal pattern might be produced:

red_{book} $gree^{n}$ $_{pe}n$ $brow^{n}$ $_{desk}$

A second area of inconsistency was in the case of a) "be composed", b) "no system" and c) "keep pace" A boundary was marked by five transcribers in a) and b) and by four in c). In a) and b), there were also discrepancies between the transcriptions of accented syllables. In c), all were transcribed keep pace. In all three cases, there was a high pitch on the first word and a low level pitch on the second as follows:

a) $be^{}$ composed b) $no^{}$ system c) $keep^{}$ pace

Transcribers who recorded a boundary after the second word also marked it as accented. Some transcribers may have been reluctant to place a boundary in these positions because the pitch on the second word was level, this being the most problematic of the tones in a theory where dynamic pitch is generally seen as an indicator of nuclearity.

A hierarchical intonational structure might facilitate the task of the transcriber here too. The intonation of the phrase in which b) and c) appear: "no system could keep pace", might thus have a larger unit consisting of both step-down contours, but each would also be a unit in its own right.

4 CONCLUSION

The divergencies discussed in the above sections highlight the theoretical problem underlying the association of auditory and signal-based analysis. At the segmental level, the facility for precise, de-contextualised replay may lead to more closely signal-linked percepts than is possible (or even desirable) in traditional auditory transcription. Also, the demand in labelling to allocate only one symbol to any given stretch of signal contradicts the known parallel nature of information transfer in speech. In fluent continuous speech, this will always lead to conflicts, however clearly segmentation criteria may be formulated. At the prosodic level, it appears that inconsistencies in marking tone group boundaries were mainly a result of the inability of the system as it stands to capture a hierarchical structure. As at the segmental level, overlapping units cannot be consistently forced into a linear string.

5 REFERENCES

[1] Beckman, Mary E and Janet B Pierrehumbert, 1986, Intonational Structure in Japanese and English, *Phonology Yearbook*, 3, 255-309.

[2] Roach, Peter J, 1979, *JIPA*, 9,1,2-6

[3] Wells, JC, 1990, *Longman Pronunciation Dictionary*, Longman.

This research was funded by Esprit Project 2589. Symbols are in SAMPA.

SONE-SCALED AND INTENSITY-J.N.D.-SCALED SPECTRAL QUANTISATION OF CHANNEL VOCODED SPEECH

R. Mannell

Speech, Hearing and Language Research Centre
Macquarie University, Sydney, Australia

ABSTRACT

Natural speech tokens were passed through a Bark-scaled channel vocoder simulation and the outputs of the 18 B.P. analysis filters were quantised at various multiples of the Sone scale and the intensity-j.n.d.-scale. The resulting synthetic speech was presented to a group of listening subjects and intelligibility scores were obtained for each type and level of quantisation. The results suggest that the Sone scale is preferable to the intensity j.n.d. scale at mid frequencies where many important speech cues are to be found.

1. INTRODUCTION

There is more than one way of measuring human perception of sound intensity. Apart from the measurement of intensity thresholds, there are three main procedures. One procedure involves the measurement of just noticeable differences (j.n.d.'s or difference limens) [5]. The second procedure involves the examination of which intensities are equivalence at different frequencies (the Phon scale) [4] The third procedure asks what changes in intensity are required to produce a doubling (for example) in the perceived loudness (Sones) [11]. A fundamental question that has still not been fully addressed is how these measures relate to each other and to the perception of speech. It might be expected that the Sone scale would be more relevant to speech perception than intensity j.n.d.'s as the former can be derived from both complex sounds and pure tones whilst the latter was originally derived from pure tones. Moore and Glasberg [8] argue that the loudness of even pure tones "depends upon the integration of loudness over a certain frequency region" (eg. 1 Bark or 1 ERB). The main disadvantage of the Sone scale is that it is very difficult to derive for individual subjects whilst it is relatively straightforward to determine the amplitude j.n.d.'s. This may explain the tendency for people working with cochlear implants to quote implant performance for individual subjects in terms of j.n.d.'s relative to the overall dynamic range [1].

2. PROCEDURE

A channel vocoder simulation developed for another project [7] was modified to incorporate a quantisation module after the analysis BP and LP filters (see figure 1). The vocoder had identical analysis and synthesis filter banks consisting of 18 Bark-scaled filters the outputs of which were demodulated by identical 50 Hz LP filters.

Two quantisation procedures were utilised, one based on the intensity- j.n.d. scale (henceforth the j.n.d. scale) and the other based on the Sone scale. The j.n.d. scale was taken from Gulick [5] (p115) and the values were logarithmically interpolated in the frequency dimension to obtain approximate j.n.d. curves for each of the 18 centre frequencies of the BP filters. For each centre frequency the 0 j.n.d. point was set as the threshold intensity and the 1 j.n.d. point was determined to be the threshold plus the j.n.d. value at the threshold intensity. The 2 j.n.d. point was determined to be the intensity at the 1 j.n.d. point plus the j.n.d. value at that intensity and so forth to give curves similar to that depicted in figure 2. The Sone scale was developed in the following way. Firstly the Phon values were determined (after Robinson & Dadson [10]) for each of the filter centre frequencies. For 40 Phons and above Sone values were derived from phon values using the formula of Kinsler et al [6]

$$L = 0.046 \times 10^{(Ln/30)}$$

(where L is loudness in sones, and Ln is loudness level in phons)

Below 40 phons this relationship no longer holds accurately and so values were derived from the data given in Fletcher [3]. This procedure directly produces the sone curves for each of the filter centre frequencies similar to the curve given in figure 2.

The tokens were quantised at the output of the analysis demodulation LP filters at 4 different j.n.d. levels (1, 2, 4 and 8 j.n.d.s') and at 6 different sone levels (0.2, 0.4, 0.8, 1.6, 3.2 and 6.4 sones) as well as a "normal" 16 bit quantisation utilising the same filters and forming the benchmark condition. This gave 11 sets of data in all. The quantisation curves (at 1000 Hz) for the 4 j.n.d. and the 6 sone conditions are shown in figure 3.

The test items were 11 vowels in an /h_d/ frame and 19 consonants in a CV frame (V=/a:/) spoken by a speaker of Australian English. These tokens were recorded to professional audio standards in an echo free room digitised and vocoded on a VAX computer. The tests were conducted in a sound treated room using calibrated TDH-49 headphones with standard cushions and circumaural seals. The test tokens presented unmasked at 70 dB s.p.l. (ref. 20 uPa). The 20 listeners were all native speakers of Australian English and none had a history of hearing or speech pathology and all were screened with a speech discrimination test which ensured that they were reliably able to identify monosyllabic words presented at 40 dB s.p.l. Relevant pairs of intelligibility conditions and classes were compared using the chi square test and tested for significant difference at the 0.01 level

3. RESULTS AND DISCUSSION

The intelligibility results for the 11 test conditions are shown in figures 4 and 5 for various phonetic classes. Figure 4 indicates that even the greatest levels of quantisation do not achieve a great deal of intelligibility loss for the vowels (as a class). The intelligibility of the vowels for the 3.2 and 6.4 sone conditions are nevertheless significantly lower the that of the 0.2, 0.4, and 0.8 sone conditions. It must also be noted that at about this level of quantisation the quality of the vowels deteriorates dramatically and they sound like they are spoken under water. It is interesting to note that some cochlear implant patients comment that the speech that they hear via their implant sounds like it is being spoken under water. For consonant intelligibility the 8 j.n.d. condition is significantly lower in intelligibility than the 16 bit condition whilst the 1.6, 3.2 and 6.4 sone conditions were significantly lower than the 16 bit and 0.2 sone conditions. An examination of both the curves and the above statistics suggests that the 8 j.n.d. condition may be equivalent in its effects on consonant intelligibility to either the 1.6 or the 3.2 sone conditions.

An examination of the results shown in figure 5 indicate fairly clear patterns for three of the classes (the stops being difficult to interpret). For the fricatives, 1.6, 3.2 and 6.4 sone results are significantly lower than the 16 bit and 0.2 sone results whilst the 8 j.n.d. results are significantly less than the 16 bit and 1 j.n.d. results. Examination of the curves suggests that for the fricatives the 8 j.n.d. condition seems to produce equivalent results to the 1.6 sone condition. For the nasals, similar results occur with significant drops in intelligibility at 1.6 sones and 8 j.n.d.'s and it would seem that the 0.8 sone and 4 j.n.d. conditions are equivalent in their effects upon intelligibility. In the case of the continuants the 6.4 sone condition is significantly lower than the 0.2 sone condition whilst all of the j.n.d. conditions are not significantly different. The equivalent points on these two curves appear to be the 3.2 sone and the 8 j.n.d. conditions.

In summary, for all phonetic classes there is no significant deterioration in intelligibility from 1 to 4 j.n.d.'s and from 0.2 to 0.8 sones. These conditions also show very little degradation (relative to the 16 bit case) in overall speech quality. Intelligibility deteriorates between 0.8 and 1.6 sones and between 4 and 8 j.n.d.'s and evidence from both the statistics and the intelligibility curves suggests that the 4 j.n.d. condition is approximately equivalent to either the 0.8 or the 1.6 sone condition. Neither the 0.8 sone nor the 4 j.n.d. condition ever display a significant drop in intelligibility relative to the 0.2 sone or the 1 j.n.d conditions respectively. These conditions can be considered the maximum levels of quantisation allowable before the intelligibility significantly deteriorates (at least for some classes) and they are also the coursest levels of quantisation that do not show a noticeable drop in speech quality. An examination of figure 3 indicates that the 4 j.n.d. curve intersects the 0.8 j.n.d. curve a little below 40 dB and that it intersects the 1.6 sone curve at about 50 dB (at 1000 Hz for a presentation level of 70 dB). This implies that the maximum allowable quantisation level is determined by the degree of quantisation down to about 40 dB and that about one quantisation step is all that is required below this level. It seems that the maximum degree of quantisation allowable before intelligibility and quality deterioration occurs is around 1 sone and that at the minimum intensity for which there appear to be significant cues (ie. down to about 40 dB) the j.n.d. curve which matches

the 1 sone curve the closest is the 4 j.n.d. curve.

A 70 dB presentation level was chosen for several reasons. Firstly, it is a comfortable listening level corresponding to the level of normal conversation. Secondly, the shape of iso-response auditory nerve tuning curves have consistent shape up to about 70 dB but increasingly distort above that level as saturation occurs [9]. Further, Dowell et al [2] found 60-70 dB but not 80 dB to be good presentation levels for cochlear implants. These similar figures imply that auditory nerve saturation is the limiting factor for both normally-hearing and cochlear implant subjects. It is reasonable to assume that we have adapted our normal speech levels to make use of that part of the intensity range (70 to 40 dB s.p.l.) where there is both sufficient intensity to pick up important cues up to 30 dB below the speech level and yet the intensity is not so high as to cause distortion of those cues through auditory nerve saturation.

It must be stressed that the curves at 1000 Hz are a fairly good representation of the sone and j.n.d. scales between 1000 and 4000 Hz, however as the frequency drops to 200 Hz or rises to about 10,000 Hz the 1 sone and the 1 j.n.d. curves become almost equivalent over the range of 40 to 70 dB. Many cues occur, however, in the frequency range where the curves in figure 2 and 3 apply and so the number of quantisation levels available would need to be determined from either the 1 sone or the 4 j.n.d. scale.

When cochlear implant performance is defined in terms of the number of j.n.d.'s in an overall dynamic range (eg. [1] dynamic range 2.6 to 16.4 dB and difference limens 0.2 to 0.8 dB) the number of available quantisation levels may actually be one quarter that implied by the quoted figures. For example, a dynamic range of 16 dB with difference limens of around 0.8 dB seems to imply the existence of 20 quantisation levels whilst it may be that there are only 5 quantisation levels available.

4. CONCLUSIONS

It seems that for much of the frequency range the maximum amount of quantisation that will not result in significant drops in intelligibility for at least some phonetic classes is 1 sone. In this frequency range and for the range of intensities that appear to contain most speech cues (70 to 40 dB for presentation levels of 70 dB) the 4 j.n.d. curve appears to produce similar results to the 1 sone quantisation level. These results support the notion that the reference point in sone calculations (40 phons or 40 dB at 1000 Hz equal to one sone) is not an arbitrary reference point but may be related to the effective data quantisation that occurs in the process of human speech perception. This is not a surprising finding when one realises that there is a power relationship between sones and phons above 40 phons (1 sone) but not below that point. It is reasonable that we would adapt our speech perception to the intensities with a more stable relationship to loudness.

5. REFERENCES

[1] BUSBY, P., TONG, Y.C., & CLARK, G. (1990), "Psychophysical studies on cochlear implant patients with early onset of profound hearing impairment", paper given at *Tactile Aids, Hearing Aids and Cochlear Implants: An International Conference*, Sydney.

[2] DOWELL, R., SELIGMAN, P., & WHITFORD, L. (1990), "Speech perception with the 22-channel cochlear prosthesis: A summary of ten years development", paper given at *Tactile Aids, Hearing Aids and Cochlear Implants: An International Conference*, Sydney, May 1- 3, 1990

[3] FLETCHER, H. (1953), *Sound and Hearing in Communication*, Van Nostrand.

[4] FLETCHER, H. & MUNSON, W.A. (1933), "Loudness, its definition, measurement and calculation", *J.A.S.A.* 5, 82-108.

[5] GULICK, W.L. (1971), *Hearing: Physiology and Psychoacoustics*, Oxford.

[6] KINSLER, L.E., FREY, A.R., COPPENS, A.B. & SANDERS, J.V. (1982), *Fundamentals of Acoustics* (3rd edn.), New York: John Wiley.

[7] MANNELL, R.H., & CLARK, J.E. (1991), "A comparison of the intelligibility scores of consonants and vowels using channel and formant vocoded speech", in *Proc. XII ICPhS*

[8] MOORE, B.C.J. & GLASBERG, B.R. (1986), "The role of frequency selectivity in the perception of loudness, pitch and time", in MOORE, B.C.J., *Frequency Selectivity and Hearing*, London: Academic Press

[9] PICKLES, J.O. (1986), "The neurophysiological basis of frequency selectivity", in MOORE, B.C.J., *Frequency Selectivity and Hearing*, London: Academic

[10] ROBINSON, D.W. & DADSON, R.S. (1956), "A re-determination of the equal-loudness relations for pure tones", *British J. Applied Physics* 7, 166-181.

[11] STEVENS, S.S. (1938), "A scale for the measurement of psychological magnitude: loudness", *Psychol. Rev.*, 43, 405-416.

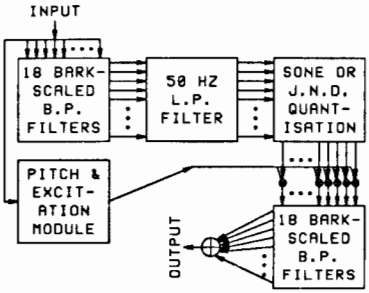


FIGURE 1. CHANNEL VOCODER

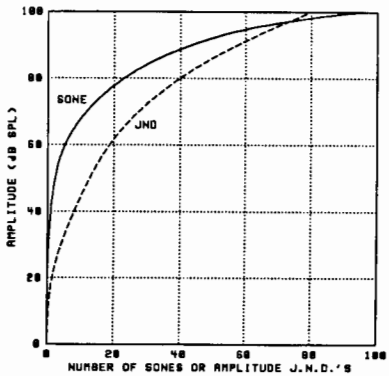


FIGURE 2. AMPLITUDE VERSUS NUMBER OF SONES OR AMPLITUDE J.N.D.'S BETWEEN EACH AMPLITUDE AND THRESHOLD AT 1000 HZ

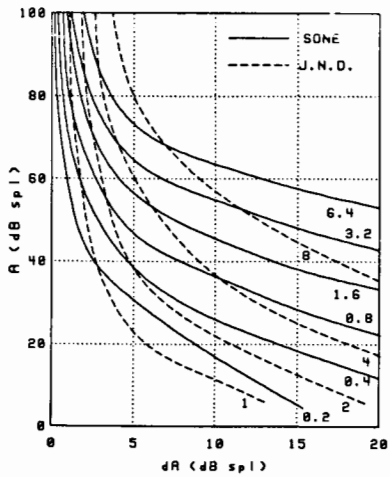


FIGURE 3. AMPLITUDE QUANTISATION CURVES (REFERENCE LEVEL [A] VS. QUANTISATION STEP [dA]) FOR VARIOUS LEVELS OF SONE AND AMPLITUDE J.N.D. QUANTISATION

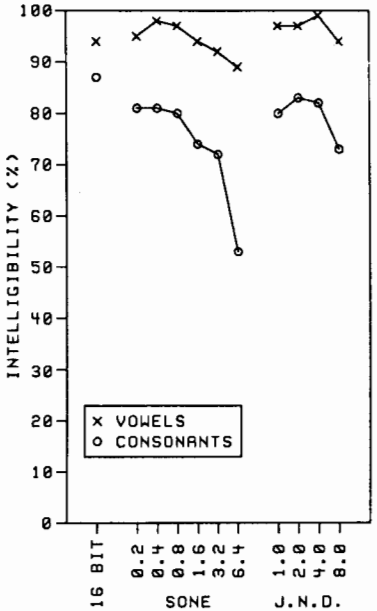


FIGURE 4. INTELLIGIBILITY SCORES FOR QUANTISED VOWELS & CONSONANTS

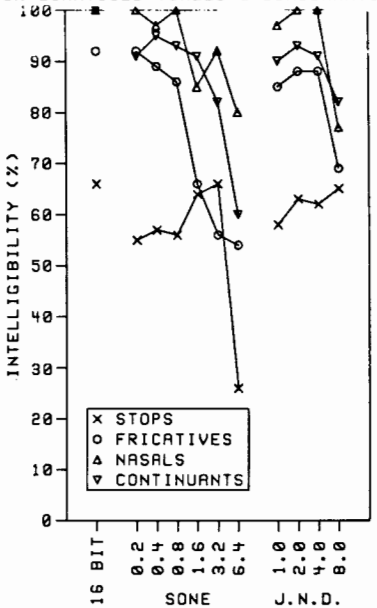


FIGURE 5. INTELLIGIBILITY SCORES FOR STOPS, FRICATIVES, NASALS, AND CONTINUANTS

PERCEPTUAL EVALUATION OF SPECTRALLY CONFUSING STOPS AND NASALS

Shigeyoshi Kitazawa
Shizuoka University, Hamamatsu, JAPAN.

ABSTRACT

Spectral similarity does not necessarily correlate to perceptual similarity. Bayesian classifier showed overlaps between consonants in spectral representation. The perceptual test of misclassified consonants was 90 % correct. Concerning 10 % of low intelligible consonants, half of the misperceptions corresponded to spectral deviations. The remaining misperceptions showed a systematic tendency which can be interpreted in terms of distinctive features.

1. INTRODUCTION

We intended to certificate the intelligibility of speech data used for speech recognition by machine. Since observed recognition errors may be due to low intelligible speech, we designed a perception test of stop consonants misclassified with a Bayesian classifier. This experiment unintentionally correlated to the hypothesis that speech sounds are perceptually decomposed into distinctive features.

This is to study the difference between machine and human being with respect to the recognition constituents. The reason why a machine does not distinguish speech sounds which a human can easily hear is discussed through the perception test of natural speech.

2. PROCEDURE

The context we have chosen was simple syllables consisting of a consonant and a vowel following. The language is French. The phonemes tested include 10 consonants /?,p,t,k,b,d,g,m,n,gn/ followed by 11 of 16 French vowels /a,o,eu,e,ai,ou,u,i,an,in,on/. We assumed a phoneme /?/ before the isolated vowel syllables. All of the syllables came from 40 male speakers living in Paris. The test sample consists of 4200 independent phonation of syllables. Most speakers are native French. In order to qualify articulation of the speech data base, 200 syllables selected randomly from 5 speakers' speech were presented to 11 listeners in a quiet listening room. French speaking listeners could identify consonants with more than 97 % of accuracy.

3. STATISTICS

As a certificate of the speech quality, syllables were classified according to the initial consonant[1]. The acoustic parameters were 28 LPC cepstrum coefficients using a 256 point Hamming window (effectively about 15 ms width under 16 kHz sampling frequency) shifted every 5 ms. Along these windows, the cepstrum coefficients are averaged over 3 consecutive frames resulting a set of 10 frames of smoothed cepstrum coefficients at every 10 ms.

The burst point of stop consonants and the release point (opening of the oral passage) of nasal consonants are determined as precisely as possible through visual inspection of waveforms. For initial vowels, initiation of vibration was inspected. The third analysis frame is at this critical point determined. With these procedure, consonant specific features are extracted.

The classification is based on the multi-dimensional statistical analysis of the above shown 290 scalars for each sample. The stepwise discriminant analysis selected around 50 to 70 elements of the vector. Then, Bayesian classifier determined the correct classification rates . In the separate analysis, 87 % for stops and 85 % for nasals were the scores.

Syllables tested were 4200 from 40 male speakers and 425 misclassifications were observed as shown in Table 1.

4. SPEECH PERCEPTION

Inspection of previous reports and our own experience show that syllables are highly intelligible if heard under low noise and wide frequency band condition. In our experimental paradigm, those syllables in which consonants are correctly classified are regarded to be intelligible, and those misclassified are subjects for perceptual experiments. Our preliminary experiment showed that most of the syllables were highly identifiable (97 % in average).

Syllables used for perceptual tests were 425 which were misclassified in the closed discriminant analysis[2]. This list contains all the possible syllables except /ou/, /teu/, /kon/. Each syllable were recorded on an audio cassette tape at a sampling rate of 16 kHz. The listeners heard the stimuli through headset, one syllable each 4 sec, and identified the syllables by writing.

All the 11 listening subjects are native French speakers. Records were kept of all responses.

5. RESULTS

Effective responses were 4620, among which 655 misperception were observed, therefore 86 % was correctly perceived. Half of the correct answers were unanimous among all listeners. The first finding means imperfection of recognizer, that is we missed some important features of consonants but puzzled with phantom features.

Among misperceptions, 462 are concerned with consonants(Table 2.), 237 of them coincide with machine errors, and the rest 225 were different perception from machine errors(Table 3.). Misperceptions concerning vowels was 231, which consists of 193 vowel errors and 38 consonant and vowel errors. Table 2. and Table 3. show in percent of each consonant presentation. Subtraction of Table 3 from Table 2 gives coinciding errors.

About half of consonant misperceptions coincided with misclassifications. This means acoustic features used for classification reflect perceptual similarities. Amongst all, 7 syllables are coincidentally misperceived by 10 of 11 hearers. Inspection of the waveforms showed that 3 errors from /b/ to /p/ and one /d/ to /t/ were not proceeded by prevoicing. In /bu/ to /u/ case, both prevoicing and burst were not observable. In /t/ to /p/ transition, very fast rising of amplitude at the onset without fricative noise was clearly observed.

The average correct response rate was 90 % which is 7 % lower than average. As usual, errors tended to accompany specific vowels or to concentrate to specific speakers and listeners. The score deviated from 87 % to 95 % between listeners. Five of the speakers also participated in the listening test. They can hear their own voice better than others.

Observing confusion matrices, we can find characteristic distributions. The perceptual confusions, Table 3, distributed along the diagonal and dencer in the upper triangular matrix. On the other hand, the matrix of machine recognition, Table 1, distributed differently. This is shown more clearly in Table 2 as the machine specific error distribution. Deviation to the lower triangular matrix is very significant comparing to the almost equal distribution in machine error (Table 1.). Another comparison with Bayesian classifier is in Table 4 as human specific perceptions. The distributions are a litte sparse to draw definite knowledge, however, rather frequent in the upper triangular matrix.

In these asymmetry of matrices, there is some specific characteristics of human perception. Confusions observed in Table 3 were sorted in terms of distinctive features as in Table 5. The table indicated meaningful tendency of perceptual transition. We will discuss in the following section.

6. DISCUSSION

Perception test of misclassified consonants is a unique experiment where several factors are combined; insufficiency of the features used by a classifier, difference of the perceptual space of speaker and hearer etc.. Since speakers hear their own voice, speaker recognize their speech to be correct. Definitely most of speeches convey sufficient acoustic information. Normally, the error rates of these speeches are very low, so it would take a long time to obtain accurate estimates of the error probabilities. However, misclassified consonants are low intelligible syllables or low intelligibility items which cause significantly higher error rates.

The importance of distinctive features in perception of consonants was demonstrated. For each feature, one feature specification (+ or -) tended to dominate over the other. As demonstrated in Tables 2, 4 and 5, there was

for each feature an asymmetry in the frequency of + and - feature specifications in error responses. With the exception of anterior, the dominant feature specifications are all "unmarked", according to traditional phonological theory. One plausible explanation for the dominance of unmarked feature specifications is that the low intelligibility of selected syllables leads to a simplification of the percept (i.e., a loss of information). In some of the perceptual shifts, acoustic features such as loss of prevoicing and weakened burst noise were observable.

The results from the present experiment are highly compatible with those from previous studies.

Previous paradigms include proximity estimates[3], identification of masked or distorted speech[4], dichotic presentation[5], recall test with the short term memory[6], and natives vs. non-natives[7]. However, much of this research dealt with listening conditions acoustically degraded or loaded stresses on listeners. Such researche has provided ample evidence that the number of *distinctive features* play an important role in perception of consonants and that the phonemes are not a perceptual unit. On the other hand, phonemes are a unit of classification.

The proximity estimates assume symmetry of the distance matrix. The analysises of MN test data also assumes symmetry of the confusion matrix. On the other hand, dichotic listening and short term recall tests are substantially asymmetric. Wickelgren did not mention about asymmetry of confusions or tendencies observed in distinctive feature system. Hayden explicitly indicated the feature specification dominance and suggested the perceptual system to favor the simpler (unmarked) feature specification in the presence of competing cues.

7. CONCLUSIONS

The purpose of this study was to reveal that the simple acoustic comparison is insufficient to explain perceptual differences of consonants. Human listeners can show essentially higher performance than machines but have different characteristics. The speaker independent acoustic analysis showed more than 90 % correct discrimination between consonant place of articulations. Those syllables misclassified by Bayesian recognizer are further examined. These outliers are an interesting set of examples providing an insight into human perception and production of speech. Most of them are phonetically perfect but uncovered by recognizers and a few of them are imperfect productions.

Perceptual experiments, using native listeners, exhibited a high intelligibility except for some acoustically confusing syllables. We found listeners made confusions under natural hearing condition. Half of the incorrect answers coincided with the misclassifications of the recognizer, perhaps through the similar evaluation of the features. Asymmetric distribution of the confusion matrix suggested that there are differences in strategy between human and machine.

The last point is important in relation to the hypothesis that speech sounds are perceptually decomposed into distinctive features. Analysis showed the tendency that perceptual system favors the simpler (unmarked) features in the presence of low intelligible cues. On the other hand, recognizers minimize the total errors by distributing errors among possible solutions.

The findings suggest that distinctive features play an important role for human perception of phonemes.

ACKNOWLEDGEMENTS

This work is an extension of cooperative work with Professor J.P.Tubach at ENST, Paris. People at ENST and ATR kindly participated in our perception tests. We appreciate conveniences offered from ENST and ATR.

REFERENCES

[1] KITAZAWA,S and J.P.TUBACH (1987), "Statistical discrimination of French initial stops," *Proc. European Conf. Speech Tech.*, 1, 91-94.

[2] KITAZAWA,S and J.P.TUBACH (1988), "Discriminant analysis and perceptual test of French stops and nasals," *Proc. 9tn Int. Conf. Pattern Recognition*, 1077-1079.

[3] BLACK,J.W.(1970), "Interconsonantal differences," *Essays in Honor of Claude M. Wise*, (Brownstein,A.J. et al. edt.) Artcraft Press,' Missouri, 74-96.

[4] MILLER,G.A. and NICELY, P. E. (1955), "An analysis of perceptual confusions among some English consonants," *J.A.S.A.*, 27, 338-352.

[5] HYDEN,M.E. and KIRSTEIN,E.

and SIGH,S.(1979),"Role of distinctive features in dichotic perception of 21 English consonants," *J.A.S.A.*, 65, 1039-1046.
[6] WICKELGREN,W.A. (1965), "Distinctive features and errors in short-term memory for English consonants," *J.A.S.A.*, 39, 388-398.
[7] SIGH,S. and BLACK, J. W. (1966), "Study of twenty-six intervocalic consonants as spoken and recognized by four language groups," *J.A.S.A.*, 39, 372-387.

Table 1. Machine Recognition Rate for Unknown Speakers

Classified	Actual Consonant									
	[ʔ]	[p]	[t]	[k]	[b]	[d]	[g]	[m]	[n]	[ɲ]
[ʔ]	.81	.08	.0	.0	.0	.0	.0			
[p]	.11	.84	.05	.02	.02	.0	.0			
[t]	.03	.04	.85	.08	.0	.2	.0			
[k]	.04	.04	.09	.90	.0	.0	.03	Not Examined		
[b]	.0	.0	.0	.0	.86	.03	.03			
[d]	.0	.0	.0	.0	.06	.92	.04			
[g]	.0	.0	.0	.0	.05	.03	.90			
[m]	Not Examined							.86	.10	.05
[n]								.09	.83	.08
[ɲ]								.05	.07	.86

Table 2. Intelligible Bayesian errors.

obs.	*real consonants*									
%	ʔ	p	t	k	b	d	g	m	n	ɲ
ʔ		7	2	3			6			
p	34			3	6					
t	12	13		40		4	3			
k	17	5	40		2		9			
b						15	3			
d					32		14			
g					30	27				
m									26	35
n								43		26
ɲ								14	28	

Table 3. Perceptual Confusions

obs.	*real consonants*									
%	ʔ	p	t	k	b	d	g	m	n	ɲ
ʔ	99	10	2	2	2		.3		.2	
p	.3	86	11	5	6	1		.2		
t		3	84	2		7				
k		1	1	88			1			
b		.8	.2		83	1	7		.2	
d			1		4	88	4			
g			.2	3	1	3	86			
m					4			98	5	
n								1	93	3
ɲ				.3			.8	.4	2	97
etc.	1		.3	.6	.4	.4				1

Table 4. Perceptions off the Bayesian errors.

obs.	*real consonants*									
%	ʔ	p	t	k	b	d	g	m	n	ɲ
ʔ		3	2	2	.6		.3		.2	
p	.2		4	4	4	1		.2		
t		1		.3		4				
k		.2	.5							
b		.8	.2			.4	2		.2	
d			1		.4		.8			
g			.2	3	1	1				
m					4				.3	
n								1		
ɲ				.3			.8		1	
etc.	1		.3	.6	.4	.4				.4

Table 5. Percentages of feature specification for perceptual errors.

features	*%+specification*	*%-specification*
Coronal	8.52	17.72
Anterior	16.16	5.06
Voiced	5.01	13.76
Consonantal	0.32	15.76

ANALYSE ACOUSTICO-PHONETIQUE DU MESSAGE VERBAL. SON RÔLE DANS LA RECONNAISSANCE LEXICALE.

Pierre-Yves Connan, François Wioland, Marie-Noëlle Metz-Lutz, Gilbert Brock

Institut de Phonétique - Université de Strasbourg II
22 rue Descartes - 67084 Strasbourg Cedex - France

ABSTRACT

This study deals with spoken language perception and comprehension during word-recognition processes. In this perspective, we tested the role of the acoustic-phonetic analysis for the activation of lexical representations. The experimental work is based upon a lexical decision task. We realized a real-time measurement of reaction-times for several listeners and in 3 different priming conditions to show facilitating or inhibiting effects of lexical access and to test the organization of the mental vocabulary. These effects are discussed in relation with theoretical modelisations of spoken word recognition, like the Marslen-Wilson "cohort model".

1. DOMAINE D'INVESTIGATION

Les travaux qui vont être décrits s'intéressent plus particulièrement au message verbal et à son traitement dans le cadre de la reconnaissance lexicale. On cherche à étudier grâce à différents protocoles expérimentaux les processus de perception et de compréhension, et plus précisément les interactions de ces divers processus lors de la reconnaissance lexicale. On a montré que les processus lexicaux, pour des sujets sains, facilitaient l'analyse acoustico-phonétique. Des études utilisant des mesures comportementales comme le temps de réaction ont mis en évidence l'influence des processus de compréhension sur la reconnaissance de la parole (Marslen-Wilson et al. 1981, 1984; Pisoni et Luce, 1987[5]; Wioland, Metz-Lutz, Brock, 1989[1]). Il a été démontré que les processus de compréhension et en particulier la reconnaissance auditive du mot sont mis en œuvre parallèlement à l'analyse acoustico-phonétique. Il faut insister sur le fait que ce processus d'analyse perceptive est à la base du traitement cognitif (Pisoni,1986; Marslen-Wilson, 1989) et qu'il est prioritaire; il est asservi aussitôt aux processus de compréhension. On va donc effectuer une approche en temps réel du traitement perceptif; cette approche se fait au cours du déroulement de la séquence de parole et elle a comme objectif de rendre compte de l'interaction entre les différents niveaux de traitement du signal.

2. ACCES AU LEXIQUE ET MODELISATION

Un certain nombre de modèles ont été proposés pour décrire l'accès au lexique interne. Ce dernier peut être défini comme l'ensemble des formes lexicales mises en mémoire par un locuteur donné.

2.1 Le modèle de recherche de Forster

Selon le modèle de Forster (1978) l'accès au lexique s'apparente au fonctionnement d'une bibliothèque. On va chercher à atteindre la cible par l'intermédiaire de fichiers et l'on dispose alors de deux principaux lieux de recherche : dans les fichiers (adresse de la cible) et dans la bibliothèque elle-même. Les fichiers périphériques et le lexique principal constituent ainsi les deux composantes de la structure du lexique. On suppose également l'organisation de sous-fichiers dans lesquels les éléments sont classés suivant un critère de fréquence. Lors de l'accès au lexique, le traitement va comporter une recherche jusqu'à ce qu'un appariement convenable soit obtenu; l'examen prend alors fin et une cote renvoie au lexique principal. L'ultime étape consiste enfin à comparer les deux représentations pour qu'une décision d'acceptation intervienne.

2.2 Modèle d'activation : le modèle de la cohorte

Soulignons ici que notre démarche expérimentale s'inspire du modèle développé par Marslen-Wilson (1978).
Le modèle était au départ de type interactif et il a évolué pour donner naissance à un modèle révisé, en 1987[3]. Le modèle actuel préconise trois fonctions fondamentales dans la reconnaissance de mots: l'*accès*, tout d'abord, où intervient le processus d'appariement d'une entrée sensorielle à plusieurs représentations lexicales qui forment alors une *cohorte*; la *sélection*, traitement qui permet de faire un choix approprié parmi les éléments de cette même cohorte, et l'*intégration*, qui correspond à l'insertion du mot dans le discours. Cette modélisation insiste sur la priorité de l'analyse d'ordre acoustico-phonétique. Il ne s'agit pourtant pas de prendre en compte toutes les caractéristiques acoustico-phonétiques d'un mot. Ainsi un premier groupe d'informations -basé sur un ensemble de traits acoustiques- est créé, et l'on peut alors délimiter un ensemble de candidats activés qui forment ce que l'on appelle la *cohorte initiale*. L'étape suivante consiste à réduire la cohorte initiale par élimination successive d'éléments de ce premier ensemble jusqu'à ce que seul le candidat adéquat soit conservé. Cette identification correcte s'effectue dans un minimum de temps et avant que la séquence vocale ait été totalement énoncée. On parle alors du *point de reconnaissance*, point, dans la séquence lexicale, où un mot est isolé par rapport aux autres candidats de la cohorte initiale. Une dernière caractéristique de cette nouvelle version du modèle met en évidence l'effet de fréquence, à savoir que pour des éléments fréquents il existe un niveau d'activation plus important; ceux-ci sont alors traités prioritairement par le processus de sélection.

3. PREMIERS TRAVAUX

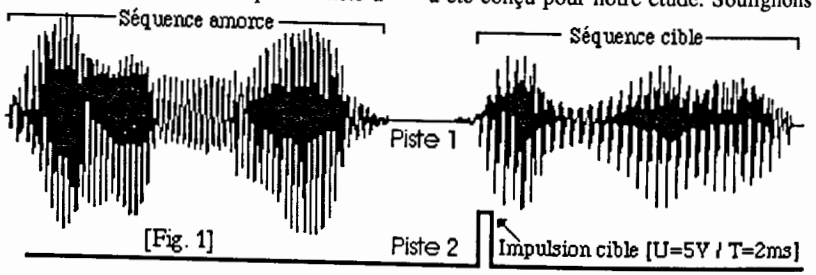
A la base, la démarche expérimentale s'inspire du modèle interactif de la cohorte et on s'est intéressé à l'interaction entre les processus de perception et de compréhension (M-N Metz-Lutz, F. Wioland, G. Brock, 1989). Cette première étude des processus d'analyse acoustico-phonétique a été effectuée à partir d'une tâche expérimentale de détection de syllabe : cette tâche consiste à identifier pendant l'écoute d'un message verbal un segment de parole sans signification (une syllabe CV), ce qui permet d'évaluer les processus perceptifs eux-mêmes. Le temps de réaction pour la détection correcte de la syllabe est enregistré dans différents contextes sollicitant tantôt une stratégie lexicale (une suite de mots trisyllabiques dans lesquels la syllabe cible se trouve en position initiale ou finale), tantôt une analyse phonético-acoustique seule (une suite de non-mots trisyllabiques portant la cible en position initiale ou finale). L'utilisation d'une stratégie lexicale se traduit, d'une part, par un allongement du temps de réaction pour la syllabe en début de mot, d'autre part, par la réduction du temps de détection pour la syllabe en finale de mot, indiquant l'activation d'une représentation lexicale phonologique du mot entendu.
De plus, si l'analyse acoustico-phonétique est la première étape qui initie les processus de reconnaissance lexicale, nous avons montré dans un contexte signifiant -mot ou énoncé- qu'elle est d'emblée soumise aux processus de reconnaissance lexicale (Wioland, Metz-Lutz, Brock, 1990[6]). En effet, le temps nécessaire à la détection d'une syllabe en début de mot est significativement plus long que pour la même syllabe isolée. Cet allongement du temps de détection n'est pas corrélé à la durée de la syllabe comme c'est la cas en syllabe isolée. Il semble donc que le processus perceptif lui-même soit déterminé par les processus de recherche lexicale.

4. APPROCHE EXPERIMENTALE

4.1 La tâche de décision lexicale

Dans les travaux sur l'accès au lexique, le paradigme expérimental classiquement

utilisé est la décision lexicale. Il est demandé au sujet de déterminer si des séquences de parole perçues auditivement, sont des mots ou des non-mots, c'est-à-dire si elles appartiennent ou non au répertoire des mots du français. Le mot ou le non-mot sur lequel doit porter la décision (la *cible*) suit de 400 ms un autre mot ou un non-mot, appelé *amorce*(en anglais,"prime"). Ce dernier peut être neutre, ou partager certaines caractéristiques avec le mot cible. Rappelons que le principe de l'accès au lexique consiste à établir une correspondance entre les caractéristiques perçues d'un stimulus et une représentation conservée en mémoire. La reconnaissance dépend également des contraintes liées à l'organisation du lexique interne. Selon cette organisation, on peut prédire des modifications du temps de décision lorsque certaines dimensions lexicales (sémantiques, syllabiques etc.) sont activées dans un contexte donné. On peut ainsi tester, en manipulant les condition d'amorçage d'un lexème donné, l'importance de ces dimensions lexicales pour l'accès au lexique interne. C'est ce que nous nous proposons d'étudier.



[Fig. 1]

4.2 Procédure expérimentale

L'installation se compose principalement d'un micro-ordinateur et de plusieurs interfaces permettant une normalisation poussée des divers signaux *cible* et *réponse*. Un enregistreur à bandes reproduit, à partir d'une première piste, la séquence de parole; sur une autre piste se trouve un signal inaudible, l'impulsion *cible*, qui indique le début du mot sur lequel va porter la décision (fig.1) et qui déclenche le chronomètre du micro-ordinateur. Le *temps de réaction* est calculé après que le sujet ait répondu et donc arrêté le chronomètre. Un logiciel de comptage des cibles et de chronométrage de ces temps de réaction, ainsi que leur mise en fichiers pour l'exploitation statistique, a été conçu pour notre étude. Soulignons que l'organisation physique du matériel verbal, enregistré au Laboratoire de Phonétique de Strasbourg, est vérifiée a posteriori par un enregistrement oscillographique qui permet d'indiquer le début du mot à identifier (fig.1).

4.3 Conditions d'amorçage

Les temps de réaction sont comparés suivant plusieurs conditions d'amorçage. De cette façon, on va tenter de mettre en évidence l'influence, facilitatrice ou inhibitrice, qu'exerce la première séquence verbale sur la seconde. Ces conditions sont au nombre de trois et l'amorce peut être alors soit de type syllabique (fig.2), soit morphémique ou encore sémantique: elles doivent permettre de tester l'organisation du lexique interne. En effet, l'existence de relations sémantiques entre les mots du lexique a été démontrée depuis les travaux princeps de Meyer et Schvaneveldt[4]. Selon la théorie de la cohorte,

Fig 2 - Extrait du corpus en amorçage syllabique

Cible n°	Type	Réponse	Amoçage	Binôme	amorce	cible
1	1	oui	+	M-M	galop	gamin
2	4	non		Nm-Nm	/tʀakaj/	/guʒɛ̃/
3	3	oui		M-M	sapin	propos
4	6	non	+	Nm-Nm	/mupɛ̃/	/mutwaʀ/
7	7	oui	+	Nm-M	/fimɛl/	figure

l'activation de certains membres du lexique repose sur l'identité du début des mots; on peut alors formuler l'hypothèse d'une organisation du lexique interne autour d'indicateurs de la forme phonologique, comme la syllabe initiale, ou de l'organisation morphologique du mot pour les lexèmes polymorphémiques. Pour ces derniers, on peut comparer l'importance des relations entre les items du lexique interne selon que la syllabe initiale a un statut phonologique ou morphémique.

5. DISCUSSION

Il faut insister sur le fait que la reconnaissance d'un mot résulte de la somme des informations sensorielles traitées en ligne. On admet que c'est sur la base de ces informations que le sujet va décider de l'appartenance du mot cible au lexique français. Sa réponse est fournie dès que les indices sont suffisants.

La finalité de la démarche expérimentale est d'étudier et de comparer les modifications du temps de décision. Dans notre protocole expérimental, l'amorce précède le début de la cible de moins de 500 ms; on cherche à tester la possibilité d'une activation automatique du lexique interne (Marcel, 1983[2]).

Une perspective différente nous amène à considérer l'importance du choix des différentes conditions d'amorçage; en effet, en précisant quel type d'amorçage favorise une décision d'acceptation plus rapide, il est possible de mettre en évidence le rôle propre des différentes informations concernant le mot phonologique ou morphologique et de préciser les modalités de la sélection finale. Au cours de l'analyse acoustico-phonétique, la reconnaissance d'un segment du mot analysé peut correspondre, par exemple, à un morphème (surtout un préfixe). On peut se demander s'il y a amorçage de tous les candidats morphémiques et comment se situe une telle activation en regard de l'amorçage syllabique, par exemple. La modification du temps de décision, d'autre part, reflétera les processus liés à l'activation de la cohorte et à la sélection du mot reconnu c'est-à-dire du bon candidat. S'il y a accélération cela montre que tous les mots de la cohorte restent activés. Dans le cas d'un allongement, on met en évidence le fait que la sélection du bon candidat repose sur l'inhibition des autres éléments de cette cohorte, alors différents du mot cible. Dans un même ordre d'idées, le fonctionnement même du processus de réduction de la cohorte initiale suscite un grand nombre d'interrogations. Le principe de la cohorte implique une adéquation entre l'entrée sensorielle et une représentation mentale. Le choix du bon candidat, par l'analyse acoustico-phonétique, amène à s'interroger sur le sort des autres candidats. Comment sont-ils réduits ou détruits ? Restent-ils activés ou sont-ils inhibés ?

Les résultats obtenus chez un groupe d'auditeurs francophones appartenant à différentes classes d'âge, cherchent à préciser le rôle de l'analyse acoustico-phonétique initiale pour la reconnaissance lexicale. Ils permettront de discuter l'importance des représentations lexicales phonologiques pour la reconnaissance des mots.

6. REFERENCES

[1] BROCK G., METZ-LUTZ M.N., WIOLAND F., 1989. Interactions entre perception et compréhension du langage oral : le temps de réponse à une syllabe sans signification considéré comme indicateur des processus de compréhension durant l'écoute d'un message verbal. *Mélanges de Phonétique générale et expérimentale offerts à Péla Simon*, vol.1, 155-181.

[2] MARCEL A.L., 1983. Conscious and unconscious perception : An approach to the relations between phenomenal Experience and Perceptual processes. *Psychology* 15, 238-300.

[3] MARSLEN-WILSON W.D., 1987.Functional parallelism in spoken word recognition. *Cognition* 25, 21-52.

[4] MEYER D.E., SCHVANEVELDT R.W., & RUDDY M.G., 1975. Loci of Contextual Effects on Visual Words Recognition, in P. Rabbit & S. Dormic (Eds.), *Attention & Performance*, V, N-Y, Academic Press.

[5] PISONI D.B. & LUCE P.A., 1987. Acoustic phonetic representations in word recognition. *Cognition* 25, 21-52.

[6] WIOLAND F., METZ-LUTZ M.N. & BROCK G., 1990. Speech perception during spoken language processing in French : How comprehension interacts with the perception of ongoing speech. *Clinical Linguistics and Phonetics*, 4, 303-318.

WORD SEGMENTATION IN MEANINGFUL AND NONSENSE SPEECH

Hugo Quené

Research Institute for Language and Speech, Rijksuniversiteit Utrecht
[quene@ruulif.let.ruu.nl]

ABSTRACT

This paper investigates the contribution of two phonetic word boundary markers on subjects' perceived word segmentation of ambiguous meaningful and nonsense word combinations. Both were realized in natural (both contrasting boundary positions intended) and synthetic speech (no boundary intended). Results show that markers are perceptually relevant, and that their contribution is the same for meaningful and nonsense stimuli. This suggests that phonetic markers are sufficient for word segmentation.

1. INTRODUCTION

In order to understand an utterance, its constituting words must be identified. To this end, a listener must (implicitly) locate the onset and offset points of words in an utterance. This *word segmentation* is argued to be a by-product of successful word recognition [1]. Listeners use stressed syllables as hypothetical word onsets. After recognition, a listener can anticipate on the subsequent word boundary and word onset (or attempt lexical access from the following stressed syllable, if the word is not yet recognized). However, this strategy only helps listeners in determining which syllables correspond to separate words. Word segmentation at the level of phonemes (or allophones), although necessary for word recognition, cannot be achieved through this strategy. Moreover, sensory word boundary information is indispensable in certain cases (e.g. words-within-words) [2].

In previous research [3,4,5], several acoustic-phonetic correlates of the (intended) word boundary have been identified. Roughly, two types of boundary phenomena can be discriminated: (1) 'explicit' boundary segments, e.g. laryngealisation or glottal stop (segmental, qualitative markers); (2) variations in segmental duration, e.g. word-initial consonant lengthening (durational, quantitative markers).

In this paper, it is examined whether these word boundary phenomena contribute to listeners' detection of word boundaries in connected speech, i.e. whether such phenomena are perceptually relevant. This can be investigated by means of manipulation of the boundary phenomena. If these markers are indeed perceptually relevant, then their manipulation should affect listeners' perceived word boundary position.

Given their (qualitative) nature, segmental boundary markers are less interesting for the present purpose. These "boundary segments" can only be perceived as the phonetic correlate of a (word or phrase) boundary. Hence, it is more interesting to assess the influence of *durational* cues on perceived word segmentation. This study concentrates on two such word boundary markers, viz. (a) the duration of the consonant adjacent to the word boundary, and (b) rise time of the vowel following the word boundary. In previous research [4,5], consonant duration was found to vary between 49 ms for intended /CVC#VC/, and 71 ms for intended /VC#CVC/; post-boundary vowel rise time varied between 19 ms and 13 ms, respectively [across 20 word combinations and 4 speakers].

In this study, the same type of stimulus material is used. Ambiguous two-word combinations may yield two distinctive sequences of two meaningful Dutch words (excluding segmentations involving geminates). For example, the combination /di(#)p(#)ɪn/ corresponds to the two Dutch two-word sequences **diep**

in "deep in" and **die pin** "that pin". In addition, ambiguous combinations yielding two *nonsense* words were included as stimuli. These enable a further test of the manipulated boundary markers: the perceptual relevance of the latter need not be limited to meaningful two-word combinations. The (unknown) intrinsic lexical effects on word segmentation are absent in nonsense word combinations.

In addition, even stronger evidence for the perceptual use of boundary markers can be obtained by using *synthetic* speech stimuli. Connected natural speech contains several acoustic-phonetic word boundary markers. The presence of all other (unmanipulated) boundary markers can be controlled in synthetic speech. If all other cues are absent, then any changes in the perceived boundary position between conditions can only be ascribed to manipulations of the phonetic word boundary markers.

In summary, the experiment reported here aims at providing evidence for the contribution of two durational boundary markers to listeners' perceived word segmentation, for combinations of either meaningful or nonsense words. These are realized (a) as /CVC#VC/ in natural speech [containing cues to the intended /C#/ boundary], (b) as /CV#CVC/ in natural speech [with cues to the intended /#C/ boundary], (c) in synthetic speech [containing no boundary cues]. In all realizations, boundary cues were manipulated by shortening and lengthening the durations of (a) the 'ambiguous' consonant adjacent to the word boundary, and (b) the rise time of the post-boundary vowel. Combining all conditions yields a 2x3x2x2 full factorial design.

2. EXPERIMENTAL METHOD

2.1. Stimulus material

Stimuli were constructed by combining a monosyllable (either a meaningful word or a non-word) having an ambiguous offset, with one having an ambiguous onset [e.g. /plat/ "plate", or /lat/ "late"]. Three types of boundary ambiguity were discriminated, depending on the number of intervocalic consonants and the possible positions of the word boundary:

Table I: Three types of word boundary ambiguity.

	nr.cons	ambiguity	
type 1	1	/V C#V/	/V #CV/
type 2	2	/VCC#V/	/VC#CV/
type 3	2	/VC#CV/	/V#CCV/

For all three types, the ambiguous boundary consonant could be either plosive, fricative or sonorant. However, type-3 word combinations with a sonorant boundary consonant are not allowed in Dutch. For each of the (3+3+2=) 8 remaining cells, 3 meaningful and 3 nonsense word combinations were constructed. For the meaningful (Dutch) word combinations, only monomorphematic words were used, and function words were avoided. Accent position was balanced between the two constituting words of a combination, and approximately balanced across boundary consonant categories and ambiguity types. For corresponding meaningful and nonsense combinations, the same member was accented.

In addition, 10 meaningful and 10 nonsense filler combinations were constructed, identical to the actual stimuli in all relevant aspects.

2.2. Natural speech material

Both contrasting versions of the 24 *meaningful* combinations were embedded in a meaningful sentence, which allowed only one segmentation. Corresponding sentences were as similar as possible with respect to number of syllables and words, stimulus position in sentence, etc. No important prosodic or syntactic break occurred within or immediately before or after the two-word sequence. In 9 out of 24 cases, it was necessary to extend the second of the two relevant words (in both versions) with a suffix, in order to fulfill these requirements. Both contrasting versions of the 24 *nonsense* combinations were embedded in a dummy carrier sentence, with a voiceless plosive immediately before and after the relevant sequence (for ease of excision).

All 2x2x24 stimulus sentences and 2x2x10 filler sentences were read twice by a professional speaker of Standard Dutch, and recorded on audio tape using high-quality equipment. Each two-word sequence was digitized (20 kHz sampling frequency, 9 kHz filtering, 12 bits) and excised from the carrier sentence [usually the second realization, unless affected by pausing, hesitation, mispronunciation, etc.]. Cuts were made at positive zero crossings (for nonsense sequences: after the preceding noise burst and within the following silent interval); no windows were applied. The resulting excerpts sounded natural, and did not suffer from

'clicks' at their onset or offset.

Natural sequences were processed identically to the (already processed) diphone source speech [6]. Digitized two-word sequences were fed into an LPC analysis (30 poles, window 25 ms, shift 10 ms). Subsequently, source type (voiced / unvoiced) and F_0 were established with a program using sub-harmonic summation and corrected if necessary. Filler combinations were digitized, excerpted and processed identically.

Subsequently, the analysis frames corresponding to (a) boundary consonant and (b) post-boundary vowel onset were established, by means of a segmentation program with auditory feedback and time-aligned displays of amplitude, voiced / unvoiced source, F_0, and original waveform. Vowel onset segments stretch from the first vowel frame to the frame with amplitude over 90% of the vowel peak amplitude (logarithmic).

2.3. Synthetic speech material

The 2x24 two-word sequences were generated by means of a diphone concatenation program [6]. In order to obtain the diphones used by this program, speech segments had been produced within (Dutch) nonsense words by the same speaker who realized the natural speech material in the present experiment (see above). From these utterances, the transition segments had been digitized (20 kHz, 12 bits), excerpted, and LPC-analysed.

The concatenation program was fed with phonetic transcriptions with accent symbols (no boundary symbols, "silence" phonemes or glottal stops). The output LPC analysis files (with marks for diphone and phoneme boundaries) were written to computer disk. The "accent" symbol yields a prominence-lending ('pointed hat') F_0 pattern on the appropriate vowel, superimposed on a declination line. After resynthesis, the diphone stimuli closely resemble natural stimuli. The crucial difference is that the synthetic speech does not contain any word boundary markers, since the diphones were originally realized word-internally. Again, filler combinations were input and concatenated identically.

Analysis frames corresponding to the two relevant intervals were established by the procedure described above, aided by the phoneme and diphone boundary marks in the LPC files. Vowel onset segments were not allowed to extend beyond the mid-vowel diphone boundary mark, nor beyond the F_0 turning point within the vowel (if accented).

2.4. Experimental conditions

The ambiguous boundary consonant and the post-boundary vowel onset were shortened (67%) or lengthened (134%) with regard to their original duration. Durations were manipulated by changing the number of samples for the appropriate frames [8]. Finally, the (2x2x (48+96)=) 576 manipulated two-word sequences, as well as the (2x2x10=) 40 unmanipulated fillers, were re-synthesised and stored on computer disk.

2.5. Stimulus tapes

The four stimulus conditions (meaningful--nonsense and natural--synthetic) were presented in separate blocks (pseudo-random order within blocks). Each block started and ended with 10 fillers. Four stimulus tapes were constructed, with counterbalancing between and within blocks. Tapes were recorded on DAT with 2.0 sec ISI (20 kHz, 9 kHz filter).

2.6. Subjects and procedure

Each tape was presented to 20 listeners (native Dutch, no reported hearing defects, language students) who received a small payment. They listened to the tapes over headphones (binaural) in a sound-treated booth. Their response sheet gave two possible responses (contrasting segmentations) for each stimulus; subjects were instructed to tick the appropriate one. Orthographic contrasts between responses were to be ignored. A short break was allowed between blocks on the stimulus tape.

Responses were fed (manually) into a computer, which calculated the rationalized arcsine [7] of the proportion of /#C/ responses (/V#CV/, /VC#CV/ or /V#CCV/, depending on combination type). The following section presents results of three tapes only, since remaining data are not yet available.

2.7. Results

The perceptual relevance of the manipulated boundary markers (a) consonant duration and (b) vowel rise time, should become apparent as a significant main effect of these factors. Influence of (c) the

speech source type and (d) the meaningful--nonsense difference on the perceptual relevance should become apparent as a significant interaction between these factors. In order to determine these effects, arcsine data were subject to an ANOVA with these four main factors. Two factors were added, viz. (e) the type of stimulus combination (8 types, fixed), and (f) the ambiguous combination (3 for each type, nested, random), see section 2.1. Main effects and relevant interactions are summarized in Table II; remaining interactions were all insignificant.

Table II: Summary of analysis of variance results.

	factor	F	df	p
(A)	cons.dur.	74.1	1,32	.001
(B)	vowel rise	20.5	1,32	.001
(C)	sp.source	149	2,64	.001
(D)	mean/nons	.4	1,32	n.s.
(E)	stim.type	.8	7,32	n.s.
(F)	stim.combi	28.3	32,1152	.001
AC		20.4	2,64	.001
AD		.1	1,32	n.s.
BC		6.1	2,64	.01
BD		1.4	1,32	n.s.
AE		3.0	7,32	.05
AF		2.0	32,1152	.01
CE		2.3	14,64	.05
CF		10.4	64,1152	.001
ACF		1.4	64,1152	.05

A Newman-Keuls post-hoc analysis on factor (C) showed a difference between natural stimuli, intended as /C#/, and both other speech source types ($p<.05$). Table III below illustrates the varying contribution of both boundary cues between the three source types (interactions AC and BC).

Table III: Mean percentage of /#C/ responses, for two manipulated boundary markers and three speech source types.

manipulation		NatC#	Nat#C	Synth
cons.dur.	Long	18	62	50
	Short	18	50	39
vowel ons.	Long	16	56	44
	Short	20	57	45

3. DISCUSSION

Both durational boundary markers under study are shown to contribute to word segmentation: manipulation affects subjects' perceived word boundary position. This perceptual relevance is identical in meaningful and nonsense conditions. Since the absence of anticipatory lexical information does not hamper word segmentation, phonetic cues seem to provide sufficient means to this end.

However, the natural speech conditions in Table III show that manipulations are only effective if they involve *post-boundary* markers (consonant in /#C/, vowel onset in /C#/). In addition, these data suggest that subjects pay primary attention to *un*manipulated markers in the natural stimuli, while the markers under study only play a secondary role. In general, subjects seem to perceive the intended boundary position on the basis of unmanipulated (probably segmental) cues. Durational cues contribute to this judgement, but only if the relevant speech segment follows the intended word boundary. The clustering of natural stimuli, intended as /#C/, with synthetic stimuli suggests that the latter also contain (uncontrolled) cues towards a /#C/ boundary position. Presumably, cues for *syllable*-initial position (in which the consonant diphones had been realized originally) were used for *word* segmentation in this experiment.

REFERENCES

[1] Cutler, A., & Norris, D. (1988) The role of strong syllables in segmentation for lexical access, *J. Experimental Psychology: Human Perception and Performance* 14 (1), 113-21.

[2] Frauenfelder, U.H. (1985) Cross-linguistic approaches to lexical segmentation, *Linguistics* 23, 669-87.

[3] Quené, H. (1987) Perceptual relevance of acoustical word boundary markers, *Proc. XIth Intl. Congress Phonetic Sc., Tallinn*, 6, 79-82.

[4] Quené, H. (1989) *The influence of acoustic-phonetic word boundary markers on perceived word segmentation in Dutch*. diss. Utrecht.

[5] Quené, H. (1991) Acoustic-phonetic cues for word segmentation. manuscript.

[6] Rijnsoever, P. van (1988) *From text to speech: User manual for Diphone Speech program DS*. IPO manual, 88.

[7] Studebaker, G.A. (1985) A "rationalized" arcsine transform, *J. Speech & Hearing Res.* 28, 455-62.

[8] Vogten, L.L.M. (1983) *Analyse, zuinige codering en resynthese van spraakgeluid*. diss. Eindhoven.

ETUDE DE LA PERCEPTION DES NOTES COURTES CHANTEES EN PRESENCE DE VIBRATO

Christophe d'Alessandro & Michèle Castellengo

LIMSI-CNRS BP133-91403 Orsay Cédex, France.
LAM, URA 868-CNRS, Tour 65-66, Université Paris VI, 4 place Jussieu,
75005 Paris, France.

ABSTRACT

This paper presents some results on perception of short vocal vibrato tones, using a method of adjustment. Means, standard deviations were computed and histograms plotted from 19 sets of responses (11 subjects, 28 types of short tones ranging from 1/2 vibrato cycle up to 2 vibrato cycles in steps of 1/4 cycle, 4 types of initial phase, mean frequency of 440 Hz, vibrato frequency of 6 Hz and vibrato amplitude of 100 cents). The main results are: A) for short tones, the pitch does not correspond to the mean frequency; B) the pitch depends on the shape of end of the tone; C) the pitch converges towards the mean frequency as the duration increases; D) the overal pattern of F0 has an influence on perception, and some simple patterns seem to behave better perceptually.

1 Introduction

L'étude de la perception de hauteur tonale dans la voix chantée en présence de vibrato de la fréquence fondamentale est un domaine qui a été relativement peu étudié. Une étude [5] (par la méthode d'ajustement) sur la perception de notes longues, qui comportent plusieurs cycles de vibrato par note, conclut à la perception d'une hauteur moyenne (arithmétique ou géométrique, ce qui est peu différent) pour des amplitudes et des fréquences de vibrato de l'ordre de grandeur de celle rencontrées dans le chant. Il est notable que cette perception moyenne est contestée par plusieurs auteurs (par exemple [4] p. 46) qui affirment que la moyenne mais aussi les deux hauteurs extrèmes peuvent être entendues, en fonction du contexte. Cependant, les différents auteurs s'accordent pour reconnaitre qu'en dehors de consignes ou de contraintes particulières, les sujets auditeurs percoivent la hauteur moyenne. Les résultats sur les notes longues ne peuvent s'appliquer pour expliquer comment sont appréciées les notes courtes, qui abondent dans les exécutions musicales. Lors d'une première étude sur les notes courtes [1] nous avions émis quelques hypothèses sur la base d'expériences préliminaires qu'il était nécessaire de reprendre de façon plus systématique. Cette communication présente une étude pour mesurer la hauteur tonale perçue lors de l'émission de notes courtes vibrées, hors de tout contexte musical, donc dans des conditions d'écoute de test psychoacoustique. Malgré la différence notable de cette situation avec celle d'une situation musicale réelle, nous pensons que les résultats peuvent se révéler utiles pour expliquer des phénomènes observés dans la production des chanteurs ou dans la perception des auditeurs.

2 Méthode

Une méthode d'ajustement a été utilisée pour mesurer la hauteur tonale des notes courtes. Les exemples synthétiques étaient produit par un synthétiseur à formants en parallèle. Les formants, maintenus fixes avec des valeurs ($F1 = 650Hz$, $F2 = 1100Hz$, $F3 = 2900Hz$, $F4 = 3300Hz$), correspondaient à un /a/. Les expériences ont été menées par 11 sujets, tous pourvu d'une éducation musicale. Certains sujets ont effectué plusieurs fois le test et 19 jeux de réponses ont été utilisé comme données expérimentales. Les stimuli étaient présentés de façon binaurale par un casque Beyer DT48 à 80 dB SPL. Pendant la présentation de la première série d'exemples, certains sujets ont attribués la différence entre les sons à une différence de timbre plutôt qu'à une différence de hauteur. La consigne donnée a donc été de se concentrer sur la différence de hauteur, mais l'influence de petites variations de hauteur sur le timbre mérite probablement une étude spécifique. Pour chaque test, les stimuli sont constitués d'une paire de sons: un son non vibré toujours identique et un son vibré que le sujet pouvait choisir parmis 12 sons de fréquences différentes ordonnées. Le vibrato suit une loi sinusoïdale autour d'une fréquence moyenne de 440 Hz, loi de fréquence (période 1/6s=167ms), ce qui correspont à une double croche avec la noire à 88, soit une note musicale courte mais courante dans la pratique musicale. L'amplitude totale du vibrato est de 100 cents (1/2 ton), la fréquence variant entre 428 et 450 Hz. Les fréquences des 12 sons d'appariement s'échelonnent entre ces deux limites, par pas de 2 Hz, les sujets devant répondre par un numéro entre 1 et 12. Le seuil différentiel de fréquence (pour des sons purs, à 440H et 80dB SPL) est légérement supérieur à 2 Hertz à cette fréquence. Les deux sons d'un même stimuli sont de durée égale, et séparés par un silence de 300 ms. D'un test à l'autre, la durée des sons varie en fonction du nombre fractionaire de cycles de vibrato présent.La plus petite durée utilisée dans ces tests correspond à une demi période de vibrato, soit une durée de 82 ms environ, et la plus grande à deux cycles complets, soit 343 ms. Les 7 durées utilisées s'étagent par palier d'un quart de cycle (41 ms). Pour chaque durée, quatres phases initiales ont été présentées, et 28 tests d'ajustement, correspondant à 28 formes ont donc été proposés, comme le résume la figure 1. Une expérience préliminaire, a permis de tester la perception de notes longues de dix cycles complets de vibrato, avec la même méthodologie, et donne des résultats identiques à ceux de raportés dans [5] pour les mêmes valeurs de vibrato et de fréquences fondamentales: la hauteur perçue est alors la moyenne soit environ 440 Hertz.

3 Résultats

Tous les sujets ont évoqué des difficultés pour juger la hauteur dans certains cas, a cause du mouvement du fondamental. Certains sujets ont même évoqué la possibilité d'entendre deux hauteurs, et la consigne a été d'ajuster à la meilleure hauteur possible (un tel phénomène est décrit dans [2] pour des glissandi de sons purs). Les résultats sont très cohérents malgré ces difficultés. La figure 1 résume les différentes formes (notées de 1 à 28) ainsi que les moyennes et écarts types obtenus. Les écarts types sont dans l'ensemble plutôt faibles, de l'ordre du seuil différentiel. Ils ne varient pas de façon significative d'une forme à l'autre.

3.1 Moyennes

La figure 2 reporte la moyenne des fréquences perçues en fonction de la forme (notée de 1 à 28, dans l'ordre). Les 7 courbes comportent 4 points chacune, pour les stimuli de même durée: l'axe des numéros de formes est aussi l'axe des durées croissantes, avec quatre phases initiales pour chaque durée. Pour une même durée les hauteurs perçues peuvent être trés différentes: les formes de plus courte durée peuvent, par exemple, suivant la phase initiale du vibrato, être perçues de 437 à 447 Hertz. En observant la courbe des 4 points d'une même durée, on remarque que ces formes se reproduisent à intervalle de 4: la courbe 1 ressemble à la 5, la 2 à la 6, la 3 à la 7. Les courbes de plus longues durées ont un ambitus plus petit. Cette "pseudo-périodicité" perceptive semble liée à la périodicité du vibrato: la forme 5 (resp. 6, 7) reproduit la forme 1 (resp. 2, 3) à une période de vibrato près, au début. La fin des formes semble donc prédominer perceptivement. Cette figure montre une convergence des hauteurs perçues, lorsque la durée augmente, vers la valeur moyenne perçue pour les notes longues. La figure 3 se déduit de la figure 2 en joignant les points, selon les durées croissantes, qui possèdent une même phase initiale (soit les point 1,5,9,13,17,21,25 pour la phase 1). Les 4 courbes obtenues oscillent, en convergeant vers la moyenne. Ces oscillations sont liées à la variation finale de la forme finale, et il semble que ce n'est pas la phase initiale qui domine perceptivement. La figure 4 se déduit de la figure 2 en joignant les points qui possèdent une demi-période finale commune. Les 4 courbes convergent de façon assymptotique, sans oscillations, vers la moyenne des notes longues. Il semble que la forme finale du son gouverne la perception de hauteur, avec une pondération due à la durée. Un tel phénomène a été observé [3] [2] pour différents glissandi de fréquence.

3.2 Histogrammes

Les histogrammes représentant le nombre de réponses obtenues pour une fréquence donnée et une forme fixée indiquent la dispersion des réponses, donc leur certidude. La figure 5 montre les 4 histogrammes obtenus pour une durée d'1/2 cycle. Deux situations apparaissent: les phases 1 et 3 donnent pics marqués, indiquant que les sujets ont perçu un peu de fréquences différentes; les phases 2 et 4 montrent un étalement des réponses, qui implique une certitude plus faible. L'examen des histogrammes, qui ne sont pas tous reproduits ici par manque de

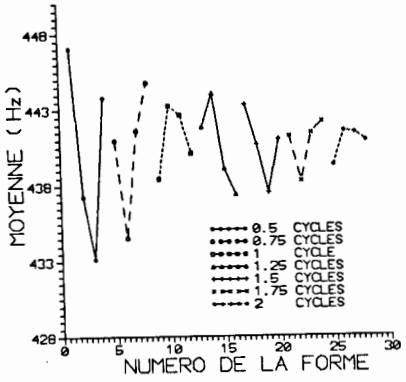


Figure 2

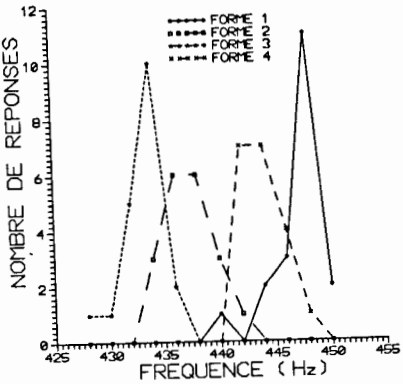


Figure 5

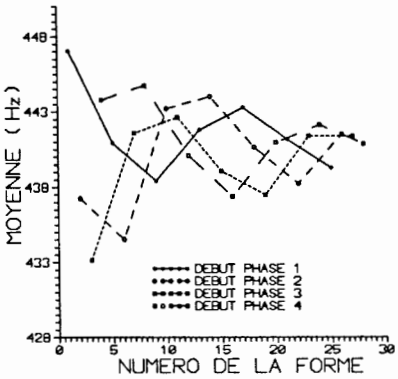


Figure 3

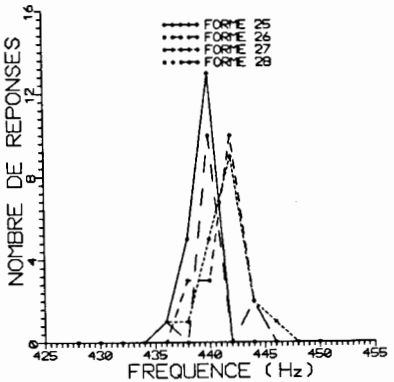


Figure 6

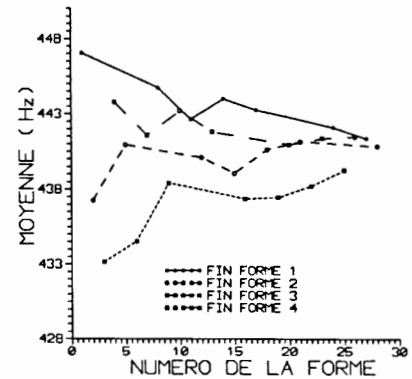


Figure 4

	Phase 1	Phase 2	Phase 3	Phase 4
0.5 cycles	forme 1 	forme 2 	forme 3 	forme 4 
Moy.	447.05	437.26	433.16	443.79
E.T.	2.34	2.23	1.92	1.87
0.75 cycles	forme 5 	forme 6 	forme 7 	forme 8 
Moy.	440.94	434.53	441.58	444.74
E.T.	1.92	2.19	3.16	1.66
1. cycles	forme 9 	forme 10 	forme 11 	forme 12 
Moy.	438.42	443.2	442.63	440.1
E.T.	2.63	2.01	2.5	2.1
1.25 cycles	forme 13 	forme 14 	forme 15 	forme 16 
Moy.	441.79	444	439.05	437.37
E.T.	1.99	2	2.78	1.5
1.5 cycles	forme 17 	forme 18 	forme 19 	forme 20 
Moy.	443.26	440.63	437.47	440.95
E.T.	1.91	1.77	2.09	1.81
1.75 cycles	forme 21 	forme 22 	forme 23 	forme 24 
Moy.	441.16	438.21	441.37	442.10
E.T.	1.54	2.39	1.77	2.35
2. cycles	forme 25 	forme 26 	forme 27 	forme 28 
Moy.	439.26	441.47	441.36	440.84
E.T.	1.19	2.09	2.22	1.80

Figure 1

place, donne une indication des formes de variation du fondamental qui semblent perceptivement préférables, surtout pour les sons les plus courts. Les sons plus long possèdent des histogrammes pointus qui indiquent une dispersion faible, comme le montre la figure 6.

4 Conclusions

Les tests d'ajustement pour des sons vocaux vibrés de courte durée indiquent que: A. les sons de courtes durées donnent lieu à des perception de hauteur très différentes de la hauteur moyenne; B. la demi-période finale permet de prévoir 4 sortes d'incidences sur la hauteur perçue: incidence haute pour une 1/2 arche positive (forme 1); incidence moyenne haute pour la 1/2 arche montante (forme 4); incidence moyenne basse pour la demi-arche descendante (forme 2); incidence basse pour la 1/2 arche négative (forme 3); C. avec l'accroissement de durée la hauteur perçue se rapproche de la moyenne; D. la forme globale du son importante, elle se combine avec les formes de fin et montre une pseudo-périodicité perceptive liée à la périodicité du vibrato; E. les formes courtes qui possèdent un seul maximum (1,3,10,12) sont perçues avec plus de certitude que les autres formes de même durée; F. lorsque la durée s'accroit la certitude des jugement devient égale et haute quelque soit la forme.

Références

[1] CASTELLENGO M., RICHARD G., d'ALESSANDRO C. (1989). "Study of vocal pitch vibrato perception using synthesis" Proceedings of the 13th Int. Cong. on Acoust. Belgrad, 113-116.

[2] NABELEK I. V., NABELEK A. K. and HIRSH I. J. (1970). "Pitch of tone bursts of changing frequency" J. Acoust. Soc. Am. 48(2), 536-553.

[3] ROSSI, M., (1971). "Le seuil de glissando ou seuil de perception des variations tonales pour les sons de la parole" Phonetica, 23, 1-33.

[4] SEASHORE C. E. (1938). *The psychology of music* Max-Graw Hill, New York.

[5] SHONLE J. I. and HORAN K. E. (1980). "The pitch of vibrato tones" J. Acoust. Soc. Am. 67(1), 246-252.

SPEECH INTELLIGIBILITY IN DEEP DIVING

Harry Hollien, Ph.D. and Patricia A. Hollien, Sc.D.

University of Florida
Gainesville, Florida, USA

ABSTRACT

Good communications are important in saturated diving. However, since heliox gas mixtures exhibit different sound transmission properties than normal air -- and high ambient pressures interact with them -- speech intelligibility tends to be degraded at depth. Attempts to improve communications here include the use of electronic processors and (in this case) modification of divers' speech. While it is known that divers can upgrade their speech at depth, little information is available about what they do to improve. Diver/talkers were trained to independently manipulate speech intensity, rate and FØ. Recordings were made at the surface and at 92.3m; intelligibility levels obtained via standardized listening sessions. The three vocal shifts that enhanced intelligibility were: low FØ, slow speech rate and high intensity.

1. INTRODUCTION

It must be conceded that divers are rather inefficient underseas workers [3]. Thermal effects, high pressures, weightlessness and especially poor communication, combine to limit a diver's ability to cope with the deep ocean environment and carry out reasonably complex tasks while doing so. In turn, divers' speech intelligibility is degraded primarily from exotic breathing gas mixtures, high ambient pressures, neural deficits, stress and hypothermia [6,8,11,12,15,17].

The four approaches which have been employed to restore the integrity of HeO2/P distorted speech can be found summarized in Figure 1. The use of trained decoders (D) is one remedial approach [10]; so is the use (C) of electronic devices [2,4,6,8, 14,16]. Third, attempts have been made to restructure language (B) as a compensation [1, 8]. However, while it must be noted that a totally new divers' lexicon probably would not be practical, appropriate data-bases are being collected and studied [13]. Finally, of the approaches portrayed in Figure 1, it is the articulatory characteristics (A) that may be both the easiest to

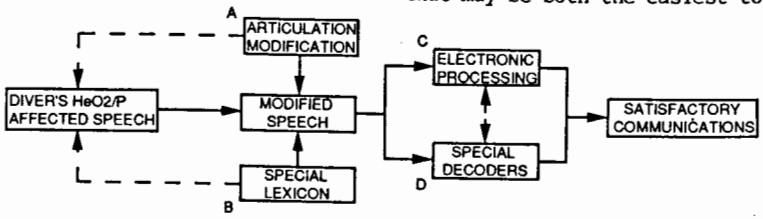


FIGURE 1

change and the most effective compensation for speech degradation in the HeO2/P milieu.

Figure 2 [7] will demonstrate that saturated divers experience severe reductions in communicative ability as a function of depth. The disparity here appears to be due to the effects of environment (reverberation, noise etc.), equipment, human variability and even HPNS (high pressure neural syndrome, 17). No simple solutions appear tenable; probably some combination of the four remedies will be necessary. While all require further study, the most critical need may be to determine how divers can modify their speech to become better communicators. This capability has not been addressed in the past.

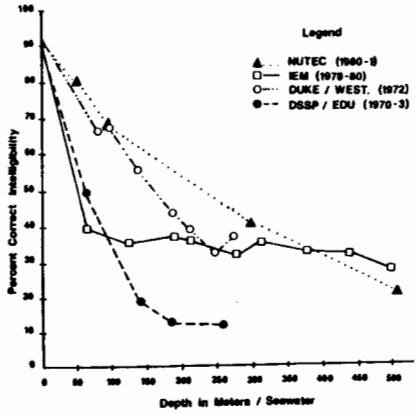


FIGURE 2

Finally, the speech changes which occur when divers attempt to communicate in HeO2/P environments have been documented to some degree. Briefly, these variations include nonlinear shifts in vowel formant structure [2,6] and (often) raised speaking fundamental frequency (FØ), at least on a behavioral basis [9, 12]. Other changes may or may not include shifts in VC ratios and the presence of nasal quality [7,8,12,15]. However, it should be stressed that these investigations have been focused on the (analysed) speech of diver/talkers -- when in environments which have varied extensively. Few investigators have attempted to study controlled/manipulated speech and these efforts have been confined primarily to shallow water [8].

2. PURPOSE

This study was conducted to investigate the effect of controlled articulatory modification on the intelligibility of divers' speech. To that end, diver/talkers were trained to separately alter, in turn, a single speech parameter while controlling all others.

3. METHOD

As stated, this investigation was carried out under highly controlled conditions with experienced divers who had completed a rigorous speech-control training program. It was conducted in the hyperbaric chambers located at the Westinghouse Ocean Research and Engineering facility, Annapolis, Maryland.

Subjects were twelve talker/divers (six males and six females) drawn from the University of Florida diver team. All were trained in phonetics and speech research, were certified/experienced divers and had served as subjects in previous experiments of this type. Moreover, to be included in the experiment, each had to demonstrate that he or she could produce the utterances with acceptable precision. Subjects were divided into equal groups; the first produced the required speech with normal sidetone, the second wore TDH-39 earphones into which was fed an 85 dB noise signal -- a procedure which essentially eliminated feedback. This approach permitted comparisons between talkers

who could hear their speech well enough to attempt enhancement to those who could not.

All talkers read eight Griffith's [5] minimal contrast word lists in the seven different speech modes with the sequence counterbalanced to avoid order effects. The speaking modes were: 1) normal articulation, 2) "most intelligible", 3) high fundamental frequency, 4) low FØ, 5) slow speaking rate, 6) fast speaking rate and 7) high vocal intensity. As stated, subjects received extensive training in using each mode (except the first two, of course) while keeping the others constant. Fundamental frequecy was monitored using the IASCP Fundamental Frequency Indicator (FFI), during training sessions and the dive; intensity by means of a calibrated sound level meter, and rate by means of a stop watch. During the dive, flashcards were employed at chamber portholes to caution talkers who were drifting from these rigid protocols. All procedures were carried out twice for both teams: first at the surface in air and secondly at 92.3m (300 fsw) in an environment consisting of 86% helium and 3% oxygen.

The recordings were made by means of calibrated, at depth, ElectroVoice 664 microphones coupled to Ampex 601 tape recorders (outside the chamber). It was concluded that, since less than 5-dB variations were obtained for frequencies up to nearly 10 kHz, the microphones were capable of functioning adequately at the experimental depth.

The experimental tape recordings were spliced, randomized and presented to groups of 12-15 listeners selected on the basis of (1) being native English speakers, (2) having normal hearing, and (3) being able to perform the listening task. Before their responses were evaluated, listeners were required to score at least 92% on a hearing screening test. Once the listening sessions were complete, the resulting data were tabled, analysed and statistical procedures applied.

4. RESULTS

One of the first contrasts was to compare the overall performance of the two groups. It was reasoned that, if training was effective, mean scores for the groups would show no differences. Indeed, they were found to be virtually identical; intelligibility levels at the surface slightly favored the group that spoke in quiet (95.3% vs. 92.8%). At depth, however, this difference was in favor of the group speaking in the high noise environment; here the difference was 2%. Further, no Lombard effect was observed in the speech of any of the talkers. These findings -- coupled with the expected male-female contrasts which were virtually identical at the surface and only slightly favored the men at depth -- served to demonstrate the robustness of the training.

Inspection of Figure 2 will reveal that a 30-60% reduction in speech intelligibility can be expected at depths around 100m. The present data are consistent with that prediction. Even though the subjects employed in this research were trained, an overall degradation of nearly 30% occurred at depth. Here the normal speaking condition (67.0%) again was found close to the overall mean and speech intelligibility was poorest for the two conditions of fast rate (61.0%) and high FØ (62.1%). The lowered FØ, however, resulted in a slightly better than average intelligibility level (69.0%) -- a finding that is not surprising since lower FØ also would tend to reduce (at least

slightly) the raised vowel formants. The three speaking conditions that demonstrated the best overall performance involved those where best intelligibility was attempted (73.0%), speaking intensity was increased (72.0%) and speaking rate was slowed (70.6%). Other than the maximum intelligibility condition, the only relationship found statistically significant (ANOVA) was loudness ($F=23.8$, df 11, 132); this factor also was significant when a post hoc Duncan's multiple range test was applied. Even though the trends for the low intelligibility conditions (fast rate; high F0) were consistant across the noise conditions and gender, they were not statistically different from the others.

5. CONCLUSIONS

It is suggested that saturated divers attempting to communicate in the HeO2/P environment can improve their performance if they consciously attempt to do so. Helpful modifications include: 1) lowered fundamental frequency level, 2) reduced speech rate, 3) increased speech intensity and 4) attempted articulatory precision. Of course, it is conceded that these speech modifications might not be equally effective for all ambient pressure levels, gas mixtures and noise levels. Nonetheless, they did result in speech improvement under the conditions of this experiment.

6. REFERENCES

[1] BAUME, A.D., et al. (1982) Procedures and Languages for Underwater Communication, UEG Tech. Note 26, London, 5-30.

[2] BELCHER, E.O. (1982) Model for Unscrambling "Helium Speech" Underwat Syst Des, 22-27.

[3] FLEMMING, N.C. (1973) The Efficiency of Scientific Diving Teams, CSL Lect.Ser., UF, Nov.

[4] GILL, J.S. (1972) The Admiralty Research Lab.Processor for Helium Speech, Proc. Helium Speech Conf., USN Sub.Med.Ctr., Groton, CT, 34-38.

[5] GRIFFITHS, J.D. (1967) Rhyming Minimal Contrasts Test, J.Acoust.Soc.Amer., 42:236-241.

[6] HOLLIEN, H. and HICKS, J.W., JR. (1981) Research on Hyperbaric Comm. IASCP/NUTEC-006/81, 1-26.

[7] HOLLIEN, H. et al. (1984) Motor Speech Characteristics in Diving, Proc. 10-ICPhS, Holland, Foris, 2:423-428.

[8] HOLLIEN, H. and ROTHMAN, H. (1976) Diver Communication, Underwater Research, London, Academic Press, 1-80.

[9] HOLLIEN, H. et al. (1977) Voice Fundamental Frequency Levels of Divers in Heliox Environments, Undersea Biomed. Res., 4:199-207.

[10] HOLLIEN, H. and THOMPSON, C.L. (1990) Effects of Listening Experience on Decoding Speech in HeO2 Environments, Diving for Science-90, 179-191.

[11] HOLLIEN, H. et al. (1973) Speech Intelligibility as a Function of Ambient Pressure and HeO2 Atmosphere, Aerospace Med., 44:249-253.

[12] MACLEAN, D.J. (1966) Analysis of Speech in a Helium Oxygen Mixture, J. Acoust. Soc. Amer., 40:625-627.

[13] MARCHAL, A., et al, (1990) DISPE: A Divers' Speech Data-Base, Proc. Third Austral. Conf. Speech Tech., 452-457.

[14] RICHARDS, M.A. (1982) Helium Speech Enhancement Using the Short-Time Fourier Transform, IEEE-ASSP, 30:841-853.

[15] ROTHMAN, H.B., et al, (1980) Speech Intelligibility at High Helium Oxygen Pressure, Undersea Biomed. Res., 7:265-275.

[16] STRAUME, O. (1980) Deep Ex 80: Diver Communication, NUI Report No. 39/80, Bergen, 1-15.

[17] VAERNES, R., et al. (1982) Central Nervous System Reactions During Heliox and Trimix Dives, Undersea Biomed. Res. 9:1-14.

Aspects théoriques et pratiques des études sur le système phonétique d'une langue.

L.Bondarko

Université d'Etat de Leningrad

ABSTRACT

A phonological theory cannot be developed if a phonologist ignores the speech activity of native speakers. Systematic analysis of Russian sound structure has revealed a discrepancy between units and operations used to describe the phonology of speech activity and those postulated by the 'classical' phonology.

L'étude théorique du système phonètique du russe contemporain langue littéraire a pour tâche de spécifier et (si besoin est) de corriger des idées concernant les relations entre le système des unités phonologiques et leurs réalisations phonétiques dans la parole.

Le moment est venu de comparer ces deux ordres de faits et d'en tirer des conclusions que aideraient à rénover les conceptions phonologiques.

Dans la série "trait distinctif - phonème - morphème - mot" chaque élément est généralement caractérisé comme un ensemble d'éléments d'un niveau plus bas: phonème - ensemble des traits distinctifs, morphème - séquense de phonèmes, mot - ensemble de morphèmes. Cependant, pour les locuteurs chaque unité linguistique est plus que la somme des unités d'un niveau plus bas. C'est pourquoi on doit se demander quel matériel linguistique suffirait pour qu'on puisse en tirer des conclusions sûres. C'est dans cet ordre d'ideès qu'on a procédé, il y a quelques années, à la création du fond phonétique russe (FPhR), qui doit constituer la base des recherches ultérieures. A présent ce fond comprend:

A.Des enregistrements sonores de différentes sortes: les syllabes CV où toutes les consonnes sont combinées avec toutes les voyelles, des mot isolés les plus fréquents et ceux qui présentent des variantes orthoépiques (3000 mots en tout); des textes suivis, contenant des mots fréquents. Tout ce matériel, enregistré dans la prononciation de 4 locuteurs, est conservé sous forme d'enregistrements sonores sur bande magnétique et en forme digitalisée dans la mémoire de l'ordinateur. Tout le matériel est présenté comme une série de syllabes ouvertes avec des marques de segmentation à l'intérieur. On procède à l'étude des caractéristiques de ces

syllabes; avec les auditeurs porteurs de différentes langues (y compris le russe) on fait des expériments, au cours desquels on modifie le matériel primitif de diverses manières pour étudier le rôle de certaines caractéristiques acoustiques pour la perception de la parole [1].

B.Un système de transcription automatique qui donne la possibilité de faire une transcription phonématique ou proprement phonétique des textes orthographiés et d'obtenir différentes caractéristiques statistiques.

C.Des dictionnaires des morphèmes russes (à la base des 3 dictionnaires morphologiques) conservés dans la mémoire de l'ordinateur, ce qui rend possibles des études statistiques.

Pour le problème qui nous occupe l'analyse de la forme sonore des morphèmes présente un intérêt particulier.

D. Des dictionnaires grammaticaux et morphologiques, ce qui donne la possibilité de comparer les structures phonétiques des constituants de mot (morphèmes) avec leur distribution dans les dérivés et formes grammaticales de mots. Ces dictionnaires sont la base pour les études sur des réalisations sonores des formes grammaticales du mot russe.

En combinant les résultats de toutes ces recherches on peut déterminer, avec assez de sûreté, certains procédés phonologiques employés par les sujets parlant et qui concernent les structures phonémiques et les réalisations sonores des unités significatives.

Selon la tradition, parmi les trois caractéristiques phonétiques du système vocalique russe (antériorité-postériorité, degré d'ouverture et labialisation) seules les deux dernières sont considérées comme relevantes, puisque l'antériorité de la voyelle dépend, dans une grande mesure, de la consonne précédente, dure ou mouillée. Dans certaines études expérimentales, cependant, on a démontré que cette affirmation est incorrecte, car les modifications articulatoires des voyelles après les consonnes mouillées représentent un "geste articulatoire" d'une nature particulière: c'est le passage de la langue d'une position avancée et élevée, propre aux consonnes mouillées, à la position reculée, propre aux voyelles postérieures. Le degré d'ouverture, comme le montrent les données du FPhR, est aussi variable, et dépend:

1) des habitudes personnelles des locuteurs et;

2) de certaines circonstances de la parole. Ainsi, en prononçant les syllabes du type CV ceux des locuteurs qui ont un débit rapide ont les [o] et [e] assez fermés, ce qui a pour résultat la perception des voyelles comme [u] et [i]/[ы]. Des modifications d'ouverture plus grandes encore sont observées dans les textes suivis, où les voyelles sont généralement réalisées comme plus fermées qu'il ne faut d'après leur caractéristique phonologique.

Il s'en suit que la variation d'une caractéristique phonétique ne donne pas le droit de la considérer comme irrélevante. La différence entre les voyelles antérieures et postérieures dans le système

phonologique du russe est dûe au fait que l'opposition entre les consonnes dures et mouillées en syllable CV n'est guère possible que devant les voyelles postérieures: /'sadu/ (au jardin, Dat.) - /s'adu/ (je m'assoirai) (Cepedant la nouvelle possibilités de réaliser cette opposition consonantique devant /e/: /pas't'el'/ 'lit' - /pas'tel'/ 'pastel' peut entraîner des changements dans le système vocalique russe).

L'analyse des formes sonores des morphèmes permet aussi d'en tirer quelques conclusions concernant le niveau phonologique du russe. L'un des fondements de la phonologie c'est l'idée sur la priorité de la fonction distinctive du phonème. L'étude du matériel phonétique - composition phonématique des morphèmes - a montré, contrairement à ce qu'on attendait, que cette fonction ne se réalise que dans un nombre assez restreint de cas. L'analyse des racines les plus fréquentes du type CVC (il y en a près de 700) montre qu'il y a très peu de paires minimales différenciées par les voyelles: 279 racines n'ont pas de paires minimales; 113 racines ne peuvent former qu'une seule opposition vocalique, trois oppositions vocaliques sont possibles dans 23 contextes consonantiques, et il n'y a que 6 contextes consonantiques où 4 oppositions sont possibles.

Les 35 consonnes qui sont possibles à l'initiale de morphèmes auraient pu former une série de 35 quasi-homonymes; cependant, la série la plus longue n'en compte que 11 (des cas semblables sont très rares), les séries les plus fréquentes ne comptent que de 2 à 5 racines.

Ainsi, la fonction distinctive du phonème n'est réalisée que dans une petite partie des cas théoriquement possibles. On peut voir cette faiblesse relative de la fonction distinctive du phonème en étudiant la différenciation des formes grammaticales de mot. En russe, la principale information grammaticale (pour les substantifs c'est le genre, le nombre, le cas) est portée par les désinences, c'est-à-dire par la partie post-tonique du mot que est sujette à une forte réduction phonétique. Dans cette situation les facteurs phonétiques l'emportent sur les facteurs phonologiques, c'est-à-dire sur la nécessité de différencier les formes grammaticales. L'homonymie des désinences est surtout fréquente dans la parole continue, où l'on observe aussi une homonymie grandissante des autres morphèmes. Donc, un locuteur russe ne rencontre que rarement des cas où la différenciation des unités significatives dans le texte se fasse grâce aux oppositions phonématiques.

Une des questions les plus importantes pour toute théorie phonologique c'est l'interprétation phonématique des segments sonores. Le plus souvent on procède à l'identification phonématique d'un son recourant aux oppositions en position forte, c'est-à-dire en position de différenciation maximale. Pour les voyelles russes c'est la syllabe accentuée. Parmi les morphèmes du russe c'est la racine qui est le plus souvent accentuée, les suffixes le sont bien plus rarement, les préfixes - plus rarement

encore.

Le traitement des donnéls du RDD [2] (110000 mots) a permis de diviser tous les cas d'apparition de préfixes comportant une voyelle en 2 groupes selon la présence ou l'absence d'accent. Il y a très peu de cas où le préfixe porte l'accent de mot (à l' exception de - qui est souvent accentuè):

voyelle du prfixe (orthographe)	accentuée	inaccentuée
e (бе́з-дар - без-да́рный) nullité inepte	13	3291
u (при́-иск - при-ня́ть) mine accepter	24	1495
o (о́т-пуск - от-пусти́ть) congé lâcher	80	6174
a (ра́з-ум - раз-у́мный) raison raisonnable	1399	6226
(вы́-пуск - вы-пускать) émission émettre	1353	1041

La comparaison de ces chiffres fait penser qu'il est peu probable qu'un locuteur russe, qui, comme on sait, tient compte, dans la parole, des caractéristiques probabilitaires, détermine la qualité phonématique des voyelles en recourant à la position forte. Il semble raisonnable de supposer que pour les locuteurs la "position typique" est aussi importante que la "position forte" pour le linguiste.

Ainsi, la réalité du phonème pour les locuteurs est liée en premier lieu à son rôle dans la structure phonétique du mot. Les cas où le modèle phonémique du mot est réalisé intégralement pendant la production de la parole ne sont pas plus fréquents que ceux où ce modèle est seulement esquissé: p. ex. dans (Acc. du rouge' adj. Nom. fém.) la suite posttonique /uju/ est réalisée phonétiquement comme groupe de voyelles [ ], où [ ] est la réalisation de la séquence /j/+/u/.

Dans les actes de perception de la parole la réalité du système de phonèmes se manifeste comme les facultés des locuteurs de rétablir la forme phonémique du mot en utilisant une information phonétique défectueuse.

Le système phonologique dont se servent les porteurs de la langue n'est pas en tous points semblable à celui établi par le linguiste, ce qui nous oblige à une étude minutieuse du matériel linguistique et de l'activité langagière des locuteurs.

References

[1] Bulleten fonetitcheskogo fonda rousskogo yazyka N1 (1988), N2 (1989), N3 (1990) Bochum-Leningrad.

[2] Worth D., Kozak A., Johnson D. 1970, Russian Derivational Dictionary, N.-Y.

LA SÉMIOLOGIE DE LA PHONOLOGIE

Jaap J. Spa

Universiteit van Amsterdam/Université de Provence

Although human language is the most elaborate semiotic that exists, man invented other sign systems to remedy the flaws inherent to the use of language. For the same reason scientists had recourse to non-linguistic signs: Phonologists proposed iconic diagrams to circumvent the linearity of discourse. As phonology developped further other sign species were created — the meaning of which could not be transmissed by language — to describe newly discovered facts.

1. INTRODUCTION

Le langage est le plus complexe de tous les systèmes sémiologiques utilisés par l'homme. Le corollaire de cette complexité est le caractère extrêmement sophistiqué des messages linguistiques. Pourquoi alors l'homme a-t-il éprouvé le besoin de se doter d'autres systèmes de signes ? C'est que la communication langagière est sujette à des restrictions, dues à la nature de l'instrument ou aux circonstances dans lesquelles on peut l'employer: Un défaut dans l'appareil récepteur, l'ambiguïté fréquente du message linguistique, l'absence de code partagé etc. (voir [9] pp 21-27) nécessitent le recours à un substitut. Ainsi toutes les anthroposémiotiques non linguistiques ont-elles été inventées pour pallier une ou plusieurs limitations du langage: L'anthropo-sémiologie est à la linguistique ce que la médecine est à la biologie humaine.

2. L'ICÔNE

Voici donc pourquoi une sémiologie de la linguistique, et partant une sémiologie de la phonologie est possible. Hagège [3] , p. 1 a affirmé: "La linguistique (...) étudiant la langue le fait en langue." A mesure que les théories linguistiques se développent, le langage peut s'avérer être un outil défectueux pour les formuler. Le linguiste a dès lors recours à des signes non linguistiques possédant des propriétés qui font défaut au langage: Les éléments d'un message linguistique n'ont entre eux que des rapports linéaires, "horizontaux". Si le linguiste veut exprimer des rapports "verticaux" et "horizontaux" entre deux concepts théoriques, il ne peut le faire au moyen du langage. Ainsi Hellwag, [4] p. 25, a été amené à poser son triangle vocalique, un signe iconique bidimensionnel, pour représenter les relations entre les voyelles de l'allemand:

u ü i
o ö e
a ä
a

L'iconicité de ce signe consiste à représenter verticalement l'aperture et horizontalement le lieu d'articulation, ce qui correspond grosso modo à ce qui se passe dans la réalité. Si on avait voulu rendre les rapports "vertical" et "horizontal" au moyen d'énoncés linguistiques on aurait pu y parvenir mais au prix de formulations laborieuses. Un simple coup d'oeil sur l'icône, cependant, suffit pour rendre ces rapports évidents. L'iconicité de ce signe n'est pas pour autant parfaite: le rapport "horizontal" entre [ü] et [i] ou entre [ö] et [e] ne correspond pas à ce qui se passe dans la réalité.

Les consonnes avaient déjà fait l'objet d'une représentation iconique 150 ans plus tôt par Montanus, [5] p. 19. Ni Hellwag, ni Montanus n'ont été les premiers à avoir employé des icônes en phonologie. Cet honneur revient à Pānini (cf. [6], p. 69). Mais sa description était tombée dans l'oubli, de sorte qu'elle devait être réinventée par les Européens Montanus et Hellwag.

Outre les schémas iconiques élaborés par Hellwag, Montanus, Pānini, d'autres icônes furent proposés dont l'objectif fut de donner une image plus fidèle de la réalité, notamment quant à la façon dont les sons étaient produits. Ainsi les images ci-dessous de John Wilkins, citées par [1] , p. 115, représentent-elles ce qui se produit dans les cavités pharyngale, nasale, buccale au moment de l'articulation des sons [t, d, s, l] :

Dans ces icônes on aperçoit la représentations visuelle d'un certain nombre d'événements dont les principaux ne sont pas visibles à l'oeil nu ou même pas visibles du tout mais seulement audibles ou éventuellement tangibles. C'est pourquoi je proposerais pour ce type d'icône l'appelation d'icône synesthésique , c.-à-d. un icône dont le signifiant s'adresse à un seul des 5 sens mais dont le signifié relève de sens différents.

3. LE SIGNAL ET LE SYMBOLE

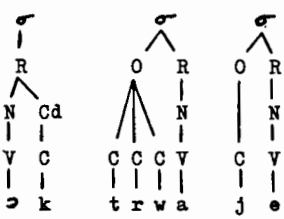
Sebeok, [8] pp 231-248, reconnaît 6 types de signes. Outre l'icône dont il a déjà été question, il distingue le signal, le symbole, le symptôme, le nom, l'indice. La flèche (→) de la phonologie générative peut être considérée comme un signal, parce que selon Sebeok ce dernier demande une action de la part de celui qui le perçoit. En l'occurence l'action consiste à réécrire une séquence. / (= dans le contexte) peut dès lors être considéré comme un symbole, un signe pour lequel signifié et signifiant sont liés par une convention arbitraire. Les trois autres types ne sont pas employés, à ma connaissance, dans la littérature linguistique.

Pourquoi des signes comme → et / se sont-ils substitués à des expressions langagières ? C'est parce que celles-ci sont trop longues et trop souvent utilisées. L'homo significans aime se servir de signes mais pas toujours des mêmes. Il renâcle devant l'obligation de réemployer chaque fois le même énoncé quand

un moyen plus simple s'offre à lui. En outre la formalisation de la linguistique permet la déduction rapide de prédictions faites à partir d'assertions théoriques généralisantes, de sorte que celles-ci, le cas échéant, peuvent être rapidement falsifiées.

4. L'ICÔNE MÉTAPHORIQUE

Les icônes dont il a été question avaient la propriété dite iconicité topologique (cf. [10], p.88) c.-à-d. la configuration de leurs parties constitutives est semblables à celle d'originaux qui existent dans la réalité. Récemment la phonologie a vu naître des icônes métaphoriques (cf. [10], p. 89): La configuration des parties constitutives de ce dernier type reproduit un original qui n'a qu'une existence métaphorique: Une personne haut placée n'est pas quelqu'un qui se situe quelques mètres au-dessus du sol mais qui est à la tête d'une hiérarchie. Une telle métaphore est à la base de la description suivante de la structure syllabique:



Les symboles σ sont hiérarchiquement supérieurs à O et R, qui à leur tour dominent N et Cd. Ensuite viennent V et C et en bas de l'échelle se trouvent les sons concrets. Une telle hiérarchie peut être exprimée également par un schéma plat, unidimensionnel:

[[[[ɔ]]] [[[k]]]] etc.
σ R N V Cd C

Cette possibilité n'existe pas pour l'icône topologique qu'est le triangle vocalique car l'original n'est pas quelque chose de linéaire.

5. LE SIGNE TRIDIMENSIONNEL

La toute dernière méthode pour décrire scientifiquement les sons est le schéma tri-dimensionnel, illustrant le fait que la séquence sonore est le résultat de différentes strates, reliées à un axe central, le squelette, sur lequel se trouvent des unités chronologiques. Un tel schéma combine les propriétés des icônes topologique et métaphorique. Le tout prend à peu près la forme d'un écouvillon. C'est dire qu'on pourra bientôt s'attendre à ce que les manuels de phonologie soient conçus comme certains livres pour enfants où il suffit d'ouvrir une page quelconque pour voir se former une image tri-dimensionnelle.

6. L'ECRITURE

Le plus vieux domaine de la sémiologie de la phonologie est l'écriture. Chez les anciens Egyptiens les signes de l'écriture étaient encore de nature iconique parce que, quand un mot avait comme segment initial le v de vache, ce v ainsi que tous les autres v étaient rendus par une tête de vache. On peut dès lors parler d'un icône dérivé. Celui-ci se transforme en symbole (arbitraire) lorsqu'il est emprunté par une autre langue pour exprimer le segment v, alors que dans cette langue le mot pour vache ne commence pas par v. L'écriture a été conçue pour remédier à la volatilité de la langue parlée.

7. CONCLUSIONS

1. Quatre limitations inhérentes à la communication langagière ont provoqué l'introduction de signes substitutifs en phonologie: la volatilité du discours oral, la linéarité du langage, la monotonie de l'emploi fréquent de messages linguistiques

identiques, l'inadéquation du langage de se prêter à des déductions fiables.

2. Le langage est la plus élaborées des anthropo-sémiotiques. Cela se voit aussi au nombre d'entrées dans [2] comparé à celui dans [7]. Aussi est-ce la linguistique qui peut fournir la clé donnant accès à l'anthroposémiologie.

RÉFÉRENCES

1 ABERCROMBIE, D. (1967), "Elements of General Phonetics", Edinburgh: Edinburgh University Press.

2 DUBOIS, J. et alii (1973), "Dictionnaire de linguistique", Paris: Librairie Larousse.

3 HAGÈGE, C. (1988), "Leçon Inaugurale", Paris: Collège de France.

4 HELLWAG, C.F. (1967), "Dissertation Inauguralis Physiologico-Medica de Formatione Loquelae 1781", H. Mol Ed.,Amsterdam: Instituut voor Fonetische Wetenschappen, Publicatie no. 10.

5 MONTANUS, P. (1964), "De Spreeckonst 1635", W.J.H. Caron Ed., Groningue: Wolters.

6 MOUNIN, G. (1974), "Histoire de la linguistique des origines au XXe siècle", coll. Le linguiste, Paris: Presses Universitaires de France.

7 REY-DEBOVE, J. (1979) "Lexique sémiotique", Paris: Presses Universitaires de France.

8 SEBEOK, Th. A. (1974), "Semiotics: A survey of the state of the art", Current trends in linguistics, vol. 12, La Haye: Mouton.

9 SPA, J.J. (1985) "Sémiologie et Linguistique, Réflexions préparadigmatiques", Amsterdam: Rodopi.

10 VAN ZOEST, A. (1978) "Semiotiek", Basisboeken, Baarn: Ambo

VOWEL PALATALIZATION IN MONGOLIAN

Jan-Olof Svantesson

Lund University, Sweden

ABSTRACT

Most Mongolian languages have gone through a process of palatalization which has affected the vowel and consonant systems in different ways in different languages. In this paper, phonetic data are given from the Khalkha dialect where consonant palatalization is contrastive, and where vowels preceding palatalized consonants have been umlauted. The umlauted vowels are realized as diphthongs, and at least for some speakers they contrast with original diphthongs with *i* as the second element. The contrast is realized as differences in spectral timing.

1. BACKGROUND

1.1 The vowel system

Classical Mongolian had seven vowels (shown below) and a vowel harmony system based on palatality with three front vowels *e, ø, y,* three back vowels *a, ɔ, u,* and one neutral vowel *i.* It is believed that the oldest stages of the language had a back unrounded vowel *ɨ* as well. There has been a vowel shift in East Mongolian languages (Mongolian proper and Buriat), by which the vowel *u* became a pharyngeal ([–ATR]) vowel *ɷ*, and the front vowels *y* and *ø* became *u* and *ɵ*, respectively [3][4] (in Southern Mongolian dialects, e.g. Baarin, *e* became *ɔ*, as well). At the same time, the phonetic basis of vowel harmony shifted from palatality to pharyngeality ([ATR]), the vowels *a, ɔ, ɷ* being pharyngeal ([–ATR]), *e, ɵ, u* non-pharyngeal ([+ATR]) and *i* neutral [4].

Classical	Khalkha	Baarin
i y u	i u	i u
e ø a ɔ	ɷ	ɷ
	e ɵ	ɔ ɵ
	a ɔ	a ɔ

Vowel length is contrastive in modern Mongolian, but only in the first syllable of a word.

1.2 Palatalized consonants

The vowel *i* caused palatalization of both consonants and vowels. Consonants preceding *i* were palatalized, and in many cases the conditioning vowel disappeared (especially when word final) or became assimilated to a following vowel, in particular when that vowel was *a: ama>am* 'mouth', *ami>amʲ* 'life'; *bara>bar* 'to finish'; *bira>bʲar* 'strength'. Palatalization did not always take place when an *i* followed, however: *miqa > max* 'meat'.

In this way a whole class of palatalized consonant phonemes appeared in Khalkha (*bʲ, pʲ, mʲ, wʲ, dʲ, tʲ, nʲ, lʲ, rʲ, gʲ, xʲ*), contrasting with the corresponding plain consonants. (In Khalkha, *l* is realized as a lateral fricative [ɮ].)

The palatalized consonant phonemes in Khalkha have a limited distribution, occurring only in words with pharyngeal vowels. In non-pharyngeal words there is no contrast between palatalized and plain consonants, a fact that indicates that palatalization of consonants took place before the vowel shift that converted the front vowels *y* and *ø* to *u* and *ɵ*.

1.3 Palatalized vowels

The palatalized consonants in pharyngeal words have in their turn palatalized (umlauted) preceding vowels. Thus, *ɷ, ɔ, a* have umlauted allophones, here written as *ɷ̈, ɔ̈, ä*, before palatalized consonants

In some Southern Mongolian dialects, e.g. Baarin, the umlauted vowels are realized as monophthongs *ʏ, œ, æ*, but in Khalkha they are diphthongic. Both short and long vowels were umlauted in a similar way.

Another source of palatalized vowels is original diphthongs with *i* as the second element, *ɷi, ɔi, ai*. In Khalkha they are retained as diphthongs, but in Baarin they became monophthongs, merging with the umlauted vowels. There is also a non-pharyngeal diphthong *ui* in Khalkha (*y* in Baarin). Instead of expected **ei* or **ɵi*, *e* is found both in Khalkha and Baarin.

2. PHONETIC INVESTIGATION

2.1 Method

The data presented here are based on recordings of three male speakers of Khalkha Mongolian, XB, DD and BB. They were born, grew up and are still living in Ulaanbaatar. Their age was 36, 26 and 21 years, respectively. A word-list illustrating various phonetic phenomena, including palatalization, was recorded. Each word was read in isolation 3-5 times by each informant. The recording was made in Ulaanbaatar using a cassette recorder of fairly high quality. The recordings were analyzed using the MacSpeech-Lab II digitizer and analysis programs.

2.2 Results and discussion

2.2.1 Umlauted vowels vs. i-diphthongs

The umlauted vowels *ä, ɔ, ɷ* and the *i*-diphthongs *ai, ɔi, ɷi* were compared by measuring F_1 and F_2 at the beginning and end of the vowel, and at three intermediate equidistant points. The words *ä:l', ail, ɔ:l', ɔir, ɷ:l', ɷil* were used. The results are shown in Figures 1-3 and in Table 1. The simultaneous equality of F_1 and F_2 was tested with Mahalanobis' D^2 test [2, p. 480] after converting the formant frequencies to the mel scale.

The umlauted vowels and the corres-

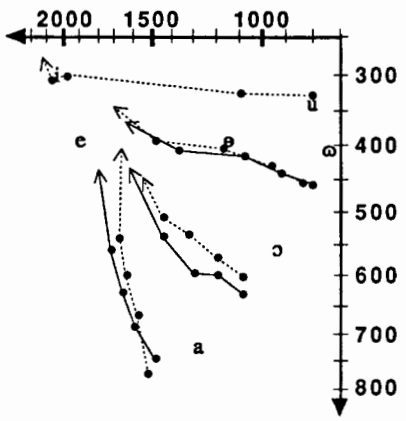


Figure 1. F_1-F_2 plot for all vowels of speaker BB. The *i*-diphthongs *ui, ɷi, ɔi, ai* are shown as dotted lines and umlauted vowels *ɷ, ɔ, ä* as solid lines. The average formant values (of 5 tokens) of monophthongic vowels are also shown.

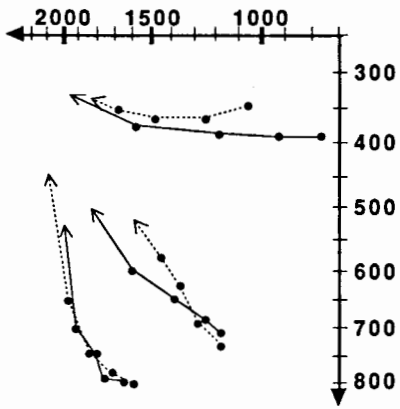


Figure 2. F_1-F_2 plots for speaker DD. Umlauted vowels *ɷ, ɔ, ä* are shown as solid lines and *i*-diphthongs *ɷi, ɔi, ai* as dotted lines.

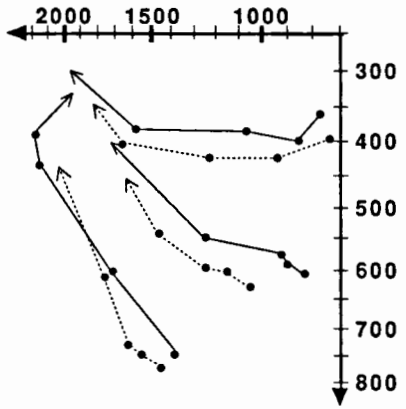


Figure 2. F_1-F_2 plots for speaker XB. Umlauted vowels *ɷ, ɔ, ä* are shown as solid lines and *i*-diphthongs *ɷi, ɔi, ai* as dotted lines.

ponding *i*-diphthongs have similar paths in the F_1-F_2 plane, starting at a point in the neighbourhood of the corresponding non-umlauted vowel and ending in the *e~i* area. According to the test results (see Table 1), *ä* and *ai* are significantly different for speaker XB and also for DD, *ɔ* and *ɔi* are different for BB and DD, and *ω* and *ωi* are different for DD and XB. Some of the differences are perceptually very salient. Although it is difficult to find invariant features which differentiate umlauted vowels and *i*-diphthongs for all speakers, the three pairs differ in similar ways for each speaker. The difference lies in the timing structure of the diphthongs rather than in the starting point, end point or direction of the diphthong path.

The spectral timing of diphthongs often differs between different languages [1], but it is an unusual feature for a language to have diphthongs with the same general start and end points but which nevertheless contrast because of their spectral timing.

2.2.2 *Palatalization of consonants following* i*-diphthongs*

Palatalized and plain consonants do not contrast after *i*-diphthongs. The quality of a consonant in this position was checked by measuring F_1 and F_2 at the beginning of the second vowel *a* in the word *ailar* and comparing to plain *l* in *ba:lar* and palatalized *l'* in *ä:l'ar*.

The results were (means of 5 tokens):

	BB		*DD*	
	F_1	F_2	F_1	F_2
ba:lar	501	1202	669	1387
ailar	505	1452	438	2002
ä:l'ar	438	1594	365	2138
tests:				
ba:lar ~ ailar	p<.01		p<.001	
ba:lar ~ ä:l'ar	p<.001		p<.001	
ailar ~ ä:l'ar	p<.05		p<.05	

As these results show, *l* in *ailar* is palatalized, but slightly less than the contrastively palatalized *l'* in *ä:l'ar*.

This seems to be the only case of progressive palatalization in Khalkha.

2.2.3 *Influence of* i *on preceding consonants*

Both plain and palatalized consonants can occur before *i* in pharyngeal words, as in *ba:lig* and *ä:l'ig*. The possible influence of *i* on a preceding plain consonant was checked by measuring F_1 and F_2 at the end of the first vowel *a:* in the words *ba:lar* and *ba:lig*, with the following results (5 tokens of each vowel for BB and DD, 4 for XB):

	BB		*DD*		*XB*	
	F_1	F_2	F_1	F_2	F_1	F_2
ba:lar	593	1246	647	1213	672	1275
ba:lig	652	1321	645	1311	642	1418
test:	ns		ns		ns	

F_2 is slightly higher at the end of *a:* in *ba:lig* than in *ba:lar*, no doubt because of coarticulation effects, but this difference is not significant in my material according to Mahalanobis' test. Since the vowel *i* is the historical source of palatalization in Mongolian, it is somewhat paradoxical that *i* does not palatalize preceding consonants in Khalkha.

2.2.4 *The quality of* i

The vowel *i* is neutral in vowel harmony, but only in a restricted sense. Words with only this vowel are always non-pharyngeal, and in pharyngeal words, *i* occurs only in suffixes. The quality of *i* in pharyngeal words is influenced by the preceding consonant. In order to check this, F_1 and F_2 were measured at the beginning and middle of the *i* vowel in the words *ba:lig*, *ä:l'ig* and *ailig*, i.e. following a plain, palatalized and non-contrastively palatalized consonant. The results are shown below where the first row for each word shows the beginning of the *i* vowel and the second row the centre:

	BB		*DD*		*XB*	
	F_1	F_2	F_1	F_2	F_1	F_2
ba:lig	473	1572	370	1949	445	1887
	465	1768	381	2083	449	2128
ailig	354	1820	246	2135	354	2044
	370	1896	367	2211	363	2194
ä:l'ig	353	1871	348	2083	–	
	351	1958	359	2181	–	
tests:	*BB*		*DD*		*XB*	
ba:lig ~ ailig	p<.001		p<.01		p<.05	
	p<.05		p<.05		p<.05	
ba:lig ~ ä:l'ig	p<.001		p<.01		–	
	p<.05		p<.01		–	
ailig ~ ä:l'ig	p<.05		ns		–	
	ns		ns		–	

(5 tokens of all words for BB and DD; 4 tokens of *ba:lig* and 3 of *ailig* for XB.)

There is a large difference between *i* following plain and palatalized consonants (as is the case for the other vowels as well). The difference is still present at the middle of the vowel but is smaller there. Thus, coarticulation between a plain consonant and a following *i* does not lead to palatalization of the consonant, as is the case in many languages, including Old Mongolian, but rather to "de-palatalization" of *i*, resulting in lower F_2 and higher F_1, a relation which is characteristic within each pair of non-pharyngeal vs. pharyngeal vowels (*e~a*, *ɵ~ɔ*, *u~ω*; cf. Figure 1).

In Baarin and other South Mongolian dialects, the contrast between plain and palatalized consonants seems to have disappeared, and *i* has split into two phonemes, non-pharyngeal ([+ATR]) *i* and pharyngeal ([–ATR]) *ɪ*, thereby repairing the asymmetry of the harmony system which resulted from the loss of *i* in Old Mongolian.

3. CONCLUSION

Mongolian has gone though a palatalization cycle. First *i* palatalized preceding consonants and was then lost in many cases. The contrastive function of the lost vowels was transferred to a palatalized/plain contrast in the consonant system, supplemented by the appearance of umlauted vowels, realized as diphthongs, before palatalized consonants. This is the stage found in Khalkha. In Baarin and other Southern Mongolian dialects, umlauted vowels have become monophthongs and carry the contrast, contrastive palatalization having disappeared, at least partly, from the consonant system.

In Khalkha, there is a contrast between umlauted vowels and original *i*-diphthongs, both being realized as diphthongs, but differing in their spectral timing. In Baarin these two sets of vowels have merged.

REFERENCES

[1] LINDAU, Mona, Kjell NORLIN & Jan-Olof SVANTESSON (1990), "Some cross-linguistic differences in diphthongs", *Journal of the International Phonetic Association, 20:1*, 10-14.

[2] RAO, Radhakrishna (1965), "*Linear statistical inference and its applications*", New York: Wiley.

[3] RIALLAND, Annie & Redouane DJAMOURI (1984), "Harmonie vocalique, consonantique et structures de dépendance dans le mot en mongol khalkha", *Bulletin de la Société de Linguistique de Paris 79*, 333-83.

[4] SVANTESSON, Jan-Olof (1985), "Vowel harmony shift in Mongolian", *Lingua 67*, 283-327.

Table 1. Mean values of F_1 and F_2 at five equidistant points in the umlauted vowels and *i*-diphthongs. The number of tokens of each vowel is given as well as test results for Mahalanobis' test, for each point testing whether the two vowels have the same F_1 and F_2 values.

	BB						DD						XB					
ä:l'	743	683	623	563	435	5	802	795	759	698	528	5	748	603	435	390	340	3
	1482	1591	1654	1708	1716		1697	1784	1828	1923	2016		1401	1722	2144	2188	1958	
ail	772	669	598	536	405	5	802	778	753	647	457	5	780	761	735	617	440	3
	1531	1580	1645	1691	1672		1621	1776	1842	2013	2092		1478	1573	1709	1877	2013	
test:	ns	ns	ns	ns	ns		ns	ns	ns	ns	<.05		<.05	ns	<.01	<.01	ns	
ɔ:l'	628	598	596	533	421	5	707	680	650	592	500	4	612	585	571	558	408	1
	1091	1172	1308	1450	1605		1187	1248	1394	1615	1853		857	911	938	1265	1713	
ɔir	601	569	539	506	449	5	735	688	620	582	519	5	630	607	594	544	460	3
	1104	1200	1330	1461	1594		1192	1289	1363	1471	1596		1052	1160	1260	1478	1623	
test:	ns	<.05	<.01	ns	ns		ns	ns	ns	<.001	<.01		ns	ns	ns	ns	ns	
ω:l'	466	441	416	408	362	5	395	398	387	377	333	4	367	405	381	385	295	3
	816	925	1081	1368	1678		765	932	1194	1608	1965		789	879	1077	1605	1985	
ωil	451	424	402	389	348	5	354	372	376	358	340	3	395	426	431	413	354	3
	865	984	1178	1496	1713		1088	1265	1491	1695	1854		757	943	1265	1700	1831	
test:	ns	ns	ns	ns	ns		<.05	<.05	ns	ns	ns		ns	ns	ns	ns	<.05	

UNIVERSALS OF NASAL ATTRITION

Bruce Connell and John Hajek

Phonetics Laboratory, University of Oxford

ABSTRACT

The claim that there is a hierarchy governing the attrition of nasals according to place of articulation is put to the test in this paper by examination of cross-linguistic data from two language groups which are unrelated genetically and geographically: the Romance dialects of Northern Italy and the Lower Cross group of South-Eastern Nigeria. Results of this new survey provide interesting food for thought: developments in the Northern Italian dialects support, to a large extent, predictions that follow from phonetic considerations. However, the Lower Cross languages at first appear to contradict expectations. This suggests that other factors may need to be taken account of, before a true universal tendency, if one exists, can be established.

1. UNIVERSAL TENDENCIES OF VN SEQUENCES

There have been numerous studies of the diachronic development of both vowels and nasals in VN sequences, e.g. [1, 4, 7, 8, 9, 11]. As a result, many generalizations have been made with purported universal or quasi-universal effect. With reference to the distinctive nasalization of vowels, we note such claims as: (1) nasalization affects low vowels first, before spreading to higher vowels; (2) front vowels are nasalized before back vowels of similar height; (3) stressed vowels are nasalized before unstressed vowels. As for nasal consonants, it has been variously claimed that : (1) nasal consonants are preferentially deleted when in tautosyllabic rather than in heterosyllabic position; and (2) weakening and deletion of N, i.e. N-attrition, occurs first in word-final position before spreading to N+C clusters. The particular claim we wish to examine here is the suggestion that the development of N-attrition is universally governed by, among other things, a parameter of place of articulation, i.e., that N-attrition will predictably affect one place of articulation before spreading in a determinable fashion to other places of articulation. Of special interest is the fact that opinions conflict as to the precise nature of this place of articulation parameter. Chen [1] states explicitly that N-attrition and nasalization of the preceding vowel affects anterior nasals [m, n] before spreading to posterior [ŋ]. Some fine-tuning of Chen's claim can be made, since it is also obvious in the diagrammatic formalizations he presents of relative backness and of nasalization processes in Chinese dialects, that he expects N-attrition and subsequent distinctive vowel nasalization to affect /Vm/ sequences before spreading to /Vn/ and then finally to /Vŋ/. Chen claims that historical developments in a sample of Chinese dialects support his observations.

Hombert [7, 8] in an analysis of the historical development of nasal consonants and N+C clusters in the Teke languages (Bantu B70) of Central Africa agrees that vowel nasalization and N-attrition affect /Vm/ (< Proto-Bantu */Vm/, */Vmb/) before /Vn/ (< */Vn/, */Vnd/). However, Hombert [7] also notes that the velar nasal in /Vŋ/ (< */ŋg/) is frequently deleted, but without evidence of expected vowel nasalization. To account for the apparently anomalous behaviour of the velar nasal, he can make only the unsatisfying suggestion that loss of /ŋ/ is an independent and unrelated phenomenon.

In contrast to Chen and Hombert, Foley [4] suggests that the order of any place of articulation parameter of N-attrition is reversed: N-attrition affects the velar nasal preferentially, before spreading along the parameter to affect /n/ and only then to /m/, i.e., ŋ > n > m; cf. also Lightner [9]. Historical developments in Portuguese, and German are cited as exemplifying the suggested directionality of the parameter.

We wish to constrain possible scenarios of sound change by providing phonetic explanations for them. Given this approach, it appears at the outset that Foley and Lightner's parameter of N-attrition is perhaps the most plausible, since it is most consistent with the phonetic observations of Ohala [10] who notes (p.297) that, "the alternation [ŋ] ~ Ṽ should be more common than the alternation of other nasals with Ṽ", as suggested by perceptual experiments. This is explained (following House 1957), "by noting that the velar nasal has primarily just a single resonating cavity with a small, perhaps negligible side-cavity, unlike other nasals, and thus negligible anti-resonances with large bandwidths and is more like that of a nasalized vowel than are those of any other nasal." Following from this, "is the prediction that of all nasal consonants one would expect [ŋ] to be most prone to change or deletion. Insofar as the zero of [ŋ] is situated in the more attenuated higher frequencies, it is less perceptible than the zeroes of other nasals, and thus make [sic] [ŋ] just that much less of a nasal."

However, while there are good phonetic reasons to claim that [ŋ] is inherently weak and most prone to loss, Ohala's prediction as to the inherent phonetic propensity of [ŋ] towards attrition is apparently not entirely borne out by cross-linguistic data. The extremely limited sets of data used by Foley [4] and Lightner [9], do concur; on the other hand, Chen's work cited earlier claims that [m, n] are weaker than [ŋ], while Hombert's suggests that devising a place hierarchy may not be such a straightforward matter as was at first thought. The apparent contradiction in data may be the result of intervening elements such as: (1) poor methodology and interpretation of data; and (2) language-specific factors that may have an over-riding effect on the expected operation of universal factors normally governing N-attrition. In order to distinguish between language-specific factors, still to be ascertained, and universal factors, an examination of developments in languages other than the small set of languages referred to above may cast some light on the purportedly universal nature of N-attrition.

2. NORTHERN ITALIAN

We have examined in detail the diachronic development of nasal consonants from Latin to the present day in a sample of Northern Italian dialects. While Latin final consonants were lost after the Classical period, the loss of final atonic vowels, combined with the rise of /ɲ/ (< L. Lat. /nj/), resulted in a new set of permissible final consonants, and a three-way contrast in final nasals: /m, n, ɲ/, cf. the limited set of examples in Fig. 1.

Latin	balneu	cane	fame
Proto-N.I	*/baɲ/	*/kan/	*/fam/
Imolese	[be:ɲ]	[kɛ̃:]	[fɛ̃:m]
Lughese	[ba:ɲ]	[kɛ̃:]	[fɛ̃:m]
Bolognese	[ba:ɲ]	[kæŋ]	[fa:m]
Riminese	[ba:ɲ]	[kɛ:n]	[fɛ:mə]
Bergamese	[baŋ]	[ka:]	[fam]
Milanese	[baŋ]	[kã:]	[fam]
Cairese	[baŋ]	[kaŋ]	[fam]
Tavetsch	[boŋ]	[cawn]	[fɔm]
gloss	bath	dog	hunger

Fig.1: Developments of PNI /*ɲ, *n, *m/.

Despite appearances, loss of word-final /n/, and related distinctive vowel nasalization are historical phenomena in Bolognese and Cairese. The presence of velar [ŋ] in place of historical word-final /n/ in Bolognese and Cairese is a recent development and, for reasons not given here, is not evidence of place shift from alveolar to velar, but is the result of a process of consonantization

of nasalized glides that developed subsequent to the deletion of word-final /n/, cf. Hajek [4, 5].

It is evident that in Northern Italian dialects, [n] is significantly more prone to N-attrition than either [m] or [ŋ]. Attrition of final [ŋ] is completely unknown in our sample, while loss of final [m] is recorded very sporadically in Imolese and Lughese. For historical reasons we can make no conjectures about the place of [ŋ] along any place of articulation parameter. However, developments in Northern Italian indicate the following order for the other nasals: n > m > ɲ. The suggested phonetic explanation for such an ordering in Northern Italian is based on duration measurements of nasal consonants in various Romance languages which correlate with diachronic developments concerning N-attrition in Northern Italian: temporally shorter nasals (i.e. [n]) are more prone to reduction and loss than longer nasals (i.e. [m], and in particular [ɲ]), cf. Hajek [6]. With regard to the relative ordering of [n] and [m], developments in Northern Italian support Foley and Lightner's suggested parameter.

3. LOWER CROSS

Nineteen languages have been identified in the Lower Cross group (Connell [2, 3]), a sub-branch of Benue-Congo situated in S-E Nigeria. Across the group, a common inventory of possible consonants in final (pre-pausal) position exists: /b d k m n ŋ/. Among certain languages of the group there is a strong tendency towards final consonant attrition, and while this has affected the different languages to varying degrees, broadly speaking, the group may be divided into two camps: those that have retained and those that are losing final Cs. While we focus here only on nasals, there seems to be little difference between oral and nasal consonants with regard to a possible place of articulation hierarchy regarding attrition, though it seems that oral consonants disappear at a faster rate than nasals.

To date only a preliminary analysis has been conducted; this, though, is sufficient to determine certain broad trends. For this, a set of 200 words was used for comparison across the group. Table One gives figures for a representative selection of languages in the group which reflect the extent of final N-attrition across the group according to place of articulation. Figures given are a count of the number of instances of final nasals found in the 200-wordlist for each language.

	m	n	ŋ
Anaang	24	20	33
Ibibio	20	21	32
Ibino	18	16	28
Ebughu	8	7	30
Ekit	9	5	27
Oro	8	3	24

Table 1: Final-N retention in LC.

These figures show clearly that [ŋ], contrary to prediction, is the 'survivor', while both [m] and [n] appear to go at approximately the same rate (only Oro does [n] appear tobe going faster than [m]). It is not surprising, given that the side-cavity for [m] and [n] are closer in size to each other than either is to that of [ŋ], that these two should behave in a similar manner; however since the phonetic structure of these consonants does seem to be playing some role in their attrition, it is not clear why other expectations do not obtain. One hypothesis is that [ŋ] is produced with greater nasal airflow, making it perceptually stronger. Preliminary aerometric investigation of Ibibio supports the claim that [ŋ] is produced with greater nasal airflow, and this may be the case across the group.

Further insight into the attrition of nasals can be gleaned in examining this phenomenon in LC in other than final position. The same three nasals occur in what may be referred to as ambisyllabic position, and while there is no strong evidence for complete loss of consonants differentially across the group, there is a strong tendency for consonants to weaken in this position. This weakening is to a large extent governed by speech rate and style, with greater reduction apparently occurring in faster and more casual styles.

In this ambisyllabic position, however, the opposite tendency to that of final position is found; i.e., /ŋ/ reduces more than either of /m/ or /n/. Typical realizations of the latter two in this position are as tapped stops, [m̆, n̆], whereas /ŋ/ is often a nasalized approximant [ŋ̈] or may even be deleted. This we consider to be an articulatory phenomenon, rather than acoustic; i.e., the articulation of /ŋ/, being achieved with the tongue dorsum, will be more affected by consonant/vowel coarticulation than either of /m/ or /n/.

4. CONCLUSIONS

Work reported here has attempted to clarify some of the phonetic factors that may be involved in the attrition of nasals. It is obvious both from previous studies and from our own work that attempting to account for this phenomenon simply in terms of a place of articulation hierarchy will not work, and that other factors need also to be considered. Among those identified in the present work are the relative durations of the consonants in question, their position within the word, and possibly the degree of nasality. It is also apparent that where /ŋ/ disappears more readily than other nasals, this may be due to articulatory, rather than acoustic, considerations. Regarding a place of articulation hierarchy, once other considerations, such as the role of duration in Northern Italian, have been taken into account, it appears that differentiating between [m, n] on one hand, and [ŋ] on the other, is the most that can be done at this point.

5. REFERENCES

[1] CHEN, M. (1974) "Metarules and Universal Constraints in Phonological Theory" in L. Heilmann (ed.) *Proceedings of the Eleventh International Congress of Linguists*, 909-924, Bologna: Il Mulino.

[2] CONNELL, B. (1990) *"Sound Correspondences, Lexicostatistics, and Lexical Innovation in the Lower Cross Languages"* Paper presented to the 20th Colloquium on African Languages and Linguistics, Leiden, The Netherlands. 3 - 5 Sept. 1990.

[3] CONNELL, B. (1991) *"Phonetic Aspects of Consonantal Sound Change in the Lower Cross Languages"*, PhD. dissertation, Edinburgh (in prep.).

[4] FOLEY, J. (1977) *"Foundations of Theoretical Phonology"*, Cambridge: Cambridge University Press.

[5] HAJEK, J. (1990) "The Hardening of Nasalized Glides in Bolognese" in P.M. Bertinetto and M. Loporcaro (eds.) *Certamen Phonologicum 2*, Turin: Rosenberg & Sellier (in press).

[6] HAJEK, J. (1991) *The Interrelationship between Vowels and Nasals: a Case Study in Northern Italian*, D. Phil dissertation, Oxford (in prep.)

[7] HOMBERT, J.-M. (1986) "The Development of Nasalized Vowels in the Teke Language Group (Bantu)" in K. Bogers and H. van der Hulst (eds.) *The Phonological Representation of Suprasegmentals*, 359-379, Dordrecht: Foris.

[8] HOMBERT, J.-M. (1987) "Phonetic Conditioning for the Development of Nasalization in Teke" in *Proceedings of the Eleventh International Congress of Phonetic Sciences*, 2: 273-276, Tallinn: Academy of Sciences of the Estonian S.S.R.

[9] LIGHTNER, T.M. (1973) "Remarks on Universals in Phonology" in M. Gross, M. Halle, and M.-P. Schützenberger (eds.) *The Formal Analysis of Natural Languages*, 13-50, The Hague: Mouton.

[10] OHALA, J.J. (1975) "Phonetic Explanations for Nasal Sound Patterns" in C.A. Ferguson, L.M. Hyman and J.J. Ohala (eds.) *Nasàlfest: Papers from a Symposium on Nasals and Nasalization*, Stanford: Language Universals Project, Stanford University, 289-316.

[11] SCHOURUP, L. (1973) "A Cross-Language Study of Vowel Nasalization", *Working Papers in Linguistics* 15: 190-221.

PHONOLOGICAL STRUCTURE AND ABSTRACT SPECIFICATION

Ewan Klein & Steven Bird

Centre for Cognitive Science, University of Edinburgh, Scotland.

ABSTRACT

Much recent work in theoretical phonology has revolved around issues of *representation*. As the structures grow in size and complexity, it becomes increasingly difficult to represent them and reason about them. In this article we shall explore techniques from computer science for abstract specification in order to provide a solution to these representation and reasoning problems. Our starting point is the assumption that phonological representations are simply rather special kinds of *data types*. Once this connection has been noted, techniques for specifying and accessing data structures can be carried over to phonology, with some interesting results. Example applications to metrical structure and feature geometry will be provided[1].

1 ABSTRACT DATA TYPES

In this section we illustrate how a particular theory of phonological representation may be recast as a definition of a certain class of data structures, i.e. as a specification of an *abstract data type*[2].

Specifications are written in a conventional format consisting of a declaration of *sorts*, operation symbols (*opns*), and equations (*eqns*). Preceding the equations we list all the variables (*vars*) which figure in them. As an illustration, we give below a specification of the data type BINARY TREE, where the leaves are labelled σ.

BINARY TREE =

sorts:	leaf, netree < tree
opns:	σ : → leaf
	$\langle \_,\_ \rangle$: tree tree → netree
	left _ : netree → tree
	right _ : netree → tree
vars:	T_1, T_2 : tree
eqns:	*left* $\langle T_1, T_2 \rangle = T_1$
	right $\langle T_1, T_2 \rangle = T_2$

The *sorts* line lists the three sorts **leaf**, **netree** and **tree**. The < sign indicates that **leaf** and **netree** are subsorts of **tree**. Anything of sort **leaf** or sort **netree** is also of sort **tree**, and anything of sort **tree** also has the sort **leaf** or **netree**, but not both. The operation symbol $\langle \_,\_ \rangle$ (where '_' marks the position of the operator's arguments) is a called a *constructor*: it builds trees out of trees (and indeed out of leaves and non-empty trees, since operators are defined for all subsorts of their domain sort)[3]. *left* _ and *right* _ are known as *selectors*: they pull trees into their component parts. The equations specify the behaviour of the two selectors.

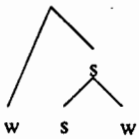
Suppose we now wish to modify the definition of binary trees to obtain metrical trees. These are binary trees where branches are ordered according to whether they are labelled 's' (strong) or 'w' (weak). We encode this

[1] We are grateful to Michael Newton and Jonathan Calder for their comments on this paper. Our work is supported by ESPRIT Basic Research Action 3175 (DYANA).

[2] For a more thorough definition of these terms, see [7]. There are various systems for the computational implementation of abstract specifications, e.g. [8,9].

[3] In general, let σ, σ', and τ be sorts such that $\sigma' < \sigma$, let f be an operator of rank $\sigma \rightarrow \tau$, and let t be a term of sort σ'. Then $f(t)$ is defined, and is a term of sort τ. From a semantic point of view, we are saying that if a function assigns values to members of particular set X, then it will also assign values to members of any subset X' of X. See [5,11] for discussion of this approach to inheritance.

information by augmenting the left or right angle bracket of our $\langle \_, \_ \rangle$ constructor with 's' according to whether the left or right branch is considered strong.



All trees have a distinguished leaf node called the 'highest terminal element', which is connected to the root of the tree by a path of 's' nodes. The specification is as follows:

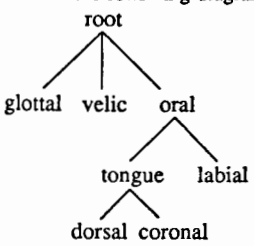
METRICAL TREE

sorts:	**leaf, netree < tree**
opns:	σ : → **leaf**
	$\langle \_, \_ \rangle$: **tree tree** → **netree**
	$\langle \_, \_ \rangle_s$: **tree tree** → **netree**
vars:	L : **leaf**, T_1, T_2 : **tree**
eqns:	*hte* $L = L$
	hte $\langle s T_1, T_2 \rangle$ = *hte* T_1
	hte $\langle T_1, T_2 \rangle_s$ = *hte* T_2

The equations state that the highest terminal element (*hte*) of a tree is the highest terminal element of its strong subtree. Another way of stating this is that the information about the highest terminal element of a subtree T is *percolated* up to its parent node, just in case T is the 's' branch of that node.

2 FEATURE GEOMETRY

The particular feature geometry we shall specify here is based on the articulatory structure defined in [4]. The five active articulators are grouped into a hierarchical structure involving a tongue node and an oral node, as shown in the following diagram.



This structure is specified below. The nine sorts and the first three operations describe the desired tree structure, using an approach which should be familiar by now. However, in contrast with our previous specifications, this specification permits ternary branching: the third constructor takes something of sort **glottal** and something of sort velic and combines them with something of sort **oral** to build an object of sort **root**.

FEATURE GEOMETRY =

sorts:	**glottal, velic, dorsal, coronal**
	labial, tongue, oral, root < gesture
opns:	$\langle \_, \_ \rangle$: **coronal dorsal** → **tongue**
	$\langle \_, \_ \rangle$: **tongue labial** → **oral**
	$\langle \_, \_, \_ \rangle$: **glottal velic oral** → **root**
	$\_$ *coronal* : **tongue** → **coronal**
	$\_$ *dorsal* : **tongue** → **dorsal**
	$\_$ *tongue* : **oral** → **tongue**
	$\_$ *labial* : **oral** → **labial**
	$\_$ *glottal* : **root** → **glottal**
	$\_$ *velic* : **root** → **velic**
	$\_$ *oral* : **root** → **oral**
vars:	C : **coronal**, D : **dorsal**, T : **tongue**,
	L : **labial**, G : **glottal**, V : **velic**, O : **oral**
eqns:	$\langle C, D \rangle$ *coronal* $= C$
	$\langle C, D \rangle$ *dorsal* $= D$
	$\langle T, L \rangle$ *tongue* $= T$
	$\langle T, L \rangle$ *labial* $= L$
	$\langle G, V, O \rangle$ *velic* $= V$
	$\langle G, V, O \rangle$ *oral* $= O$
	$\langle G, V, O \rangle$ *glottal* $= G$

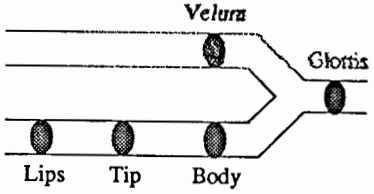
The selectors (e.g. *coronal*) occupy most of the above specification. Notice how each selector mentioned in the *opns* section appears again in the *eqns* section. Consider the *coronal* selector. Its *opns* specification states that it is a function defined on objects of sort tongue which returns something of sort coronal. The corresponding equation states that $\langle C, D \rangle$ *coronal* $= C$. Now C has the sort coronal and D has the sort dorsal. By the definition of the first constructor, $\langle C, D \rangle$ has the sort tongue. Furthermore, by the definition of the *coronal* selector, $\langle C, D \rangle$ *coronal* has the sort coronal. So the equation $\langle C, D \rangle$ *coronal* $= C$ respects the sort definitions.

Selectors can be used to implement structure-sharing (or re-entrancy). Suppose that two segments S_1 and S_2 share a voicing specification. We can write this as follows: S_1 *glottal* $=$ S_2 *glottal* . This structure sharing is consis-

tent with one of the main motivating factors behind autosegmental phonology, namely, the undesirability of rules such as [α voice] → [α nasal]. The equation *S glottal = S velic* is illsorted.

Now we can illustrate the function of selectors in phonological rules. Consider the case of English regular plural formation (**-s**), where the voicing of the suffix segment agrees with that of the immediately preceding segment, unless it is a coronal fricative (in which case there must be an intervening vowel). Suppose we introduce the variables S_1, S_2 : **root**, where S_1 is the stem-final segment and S_2 is the suffix. The rule must be able to access the coronal node of S_1. Making use of the selectors, this is simply S_1*oral tongue coronal* (a notation reminiscent of paths in feature logic, [10]). The rule must test whether this coronal node contains a fricative specification. This necessitates an extension to our specification, which will now be described.

Browman & Goldstein [4:234ff] define 'constriction degree percolation', based on what they call 'tube geometry'. The vocal tract can be viewed as an interconnected set of tubes, and the articulators correspond to valves which have a number of settings ranging from fully open to fully closed. These settings will be called *constriction degrees* (CDs), where fully closed is the maximal constriction and fully open is the minimal constriction.



The net constriction degree of the oral cavity may be expressed as the maximum of the constriction degrees of the lips, tongue tip and tongue body. The net constriction degree of the oral and nasal cavities together is simply the minimum of the two component constriction degrees. To recast this in the present framework we employ our notion of percolation again. In order to simplify exposition, the definition of *max* and *min* are omitted. Moreover, we will also assume, without giving details, that a 'constriction degree' (CD) value is specified for every gesture and is selected by the operator *cd* .

CD = FEATURE GEOMETRY +

sorts: **obs, open < cd**
clo, crit < obs
narrow, mid, wide < open

opns: $\_\ cd$: **gesture → cd**
max, min : **cd cd → cd**

vars: C : **coronal**, D : **dorsal**, T : **tongue**, L : **labial**, G : **glottal**, V : **velic**, O : **oral**

eqns: $\langle G,V,O\rangle\ cd\ =$
$max(G\ cd\ , min(V\ cd\ , O\ cd\ ))$
$\langle T,L\rangle\ cd\ = max(T\ cd\ , L\ cd\ )$
$\langle C,D\rangle\ cd\ = max(C\ cd\ , D\ cd\ )$

There are five basic constriction degrees (**clo**, **crit**, **narrow**, **mid**, and **wide**), and these are grouped into two sorts **obs** and **open**.

Using the above extension, the condition on the English voicing assimilation rule could be expressed as follows[4], where *Crit* : **crit**:

S_1 *oral tongue coronal cd* ≠ **crit**

If this condition is met, the effect of the rule would be:

S_1 *glottal cd* = S_2 *glottal cd*

This is how we say that S_1 and S_2 have the same voicing.

Now the manner features can be expressed as follows (omitting **strident** and **lateral**).

MANNER FEATURES = CD

opns: $+\_$: → **bool**
$-\_$: → **bool**
son $\_$: **root** → **bool**
cont $\_$: **root** → **bool**
cons $\_$: **root** → **bool**
nas $\_$: **root** → **bool**

vars: R : **root**, G : **glottal**, V : **velic**, O : **oral**
Open : **open**, *Obs* : **obs**, *Clo* : **clo**

eqns: *son* $R = +$ *iff* $R\ cd$ = *Open*
cont $\langle G,V,O\rangle = -$ *iff* $O\ cd$ = *Clo*
cons $\langle G,V,O\rangle = -$ *iff* $O\ cd$ = *Obs*
nas $\langle G,V,O\rangle = -$
iff $V\ cd$ = *Open and* $O\ cd$ = *Obs*

[4] A proviso is necessary here. Just because there is a critical CD at the tongue tip does not mean that a fricative is being produced. For example, the lips might be closed. We can get around this problem with the use of CD percolation (as already defined) and the equation S_1*oral* = **crit**. Further discussion of this option may be found in [2].

It follows directly from the above definitions that the collection of noncontinuants is a subset of the set of consonants (since **clo** < **obs**). Similarly, the collection of nasals is a subset of the set of consonants. Note also that these definitions permit manner specification independently of place specification, which is often important in phonological description.

3 CONCLUSIONS

We began this article by pointing out the difficulty of defining and using complex phonological structures. In addressing this problem we have used a strategy from computer science known as abstract specification. We believe this brings us a step further towards our goal of developing a computational phonology.

This approach contrasts with the finite state approach to computational phonology [1,6]. Finite state grammars have employed a rigid format for expressing phonological information, and have not hitherto been able to represent the complex hierarchical structures that phonologists are interested in. Our approach has been to view phonological structures as abstract data types, and to obtain a rich variety of methods for structuring those objects and for expressing constraints on their behaviour.

We have briefly examined the idea that data can be structured *in terms of* sorts and operations on items of specific sorts. We also explored the organization of data into a hierarchy of classes and subclasses, where data at one level in the hierarchy inherits all the attributes of data higher up in the hierarchy. Inheritance hierarchies provide a succinct and attractive method for expressing a wide variety of linguistic generalizations. A useful extension would be to incorporate *default inheritance* into this system.

Further exploration of these proposals, we believe, will ultimately enable the mechanical testing of predictions made by phonological systems and the incorporation of phonological components into existing computational grammars.

4 REFERENCES

[1] ANTWORTH, E. L. (1990). *PC-KIMMO: A Two-level Processor for Morphological Analysis.* Dallas: Summer Institute of Linguistics.

[2] BIRD, S. (1990). *Constraint-Based Phonology.* Ph.D. Thesis. University of Edinburgh.

[3] BIRD, S. & E. KLEIN (1990). Phonological events. *Journal of Linguistics, 26,* 33–56.

[4] BROWMAN, C. & L. GOLDSTEIN (1989). Articulatory gestures as phonological units. *Phonology, 6,* 201–251.

[5] CARDELLI, L. (1988). A semantics of multiple inheritance. *Information and Computation, 76,* 138–164.

[6] DALRYMPLE, M., R. KAPLAN, L. KARTTUNEN, K. KOSKENNIEMI, S. SHAIO & M. WESCOAT (1987). *Tools for Morphological Analysis.* CSLI–87–108. CSLI, Stanford.

[7] EHRIG, H. & B. MAHR (1985). *Fundamentals of Algebraic Specification 1: Equations and Initial Semantics.* Berlin: Springer Verlag.

[8] GOGUEN, J.A., J.W. THATCHER & E.G. WAGNER (1976). An initial algebra approach to the specification, correctness and implementation of abstract data types. In R. Yeh (ed.) *Current Trends in Programming Methodology IV: Data Structuring.* 80–144. Englewood Cliffs, NJ : Prentice Hall.

[9] GOGUEN, J.A., & T. WINKLER (1988). *Introducing OBJ3.* Technical Report SRI-CSL-88-9, SRI International, Computer Science Laboratory, Menlo Park, CA.

[10] KASPER, R. & W. ROUNDS (1990). The logic of unification in grammar. *Linguistics and Philosophy, 13,* 35–58.

[11] SMOLKA, G. & H. AIT-KACI (1989) Inheritance hierarchies: semantics and unification. *Journal of Symbolic Computation, 7,* 343–370.

ADJUNCTION IN SYLLABLE STRUCTURE

David Michaels

University of Connecticut

ABSTRACT

When strings of phonological segments are organized into syllable structures, diverse phonological processes fall together as adjunctions of syllable-free segments to syllabified segments.

In French a sequence of vowel and nasal consonant contract to a nasalized vowel before a consonant (1a), but not before a vowel (1b).

(1) a. bon garçon
b. bon ami

Also, obstruents delete before consonants (2a), but not before vowels (2b).

(2) a. petit garçon
b. petit ami

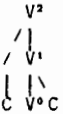
Both nasalization and deletion occur as well before a strong boundary (utterance finally, for example). But neither nasalization nor consonant deletion occur before a feminine ending as in (3a,b), which has the effect of an overt vowel in this context, though it is not pronounced.

(3) a. bonne fille
b. petite fille

Nasalization in (1a) is consistent with an analysis of French as a CV language where only consonant positions preceding vowels are licensed. Sonorant consonants like n, however, when syllable-free (in a CVN sequence, for example, where only C and V occupy licensed syllable structure positions), adjoin to the preceding vowel position, giving a nasalized vowel. In (1b), on the other hand, there is a vowel following n which licenses it as onset. Since onsets are obligatory, adjunction to the preceding vowel is ruled out. In (2a), t, an obstruent, followed by a consonant g deletes since to be phonetically interpreted it must be licensed by a following vowel, as it is in (2b). Obstruents, unlike sonorants, cannot adjoin to a preceding vowel. In (3a,b) the feminine ending serves as the licenser of the preceding consonant. Thus, consonant licensing is effected by an overt vowel or by a grammatical entity like the feminine ending which may be phonetically empty.

In the above account, consonant licensing is determined by syllable structure. I assume every syllable is headed by a vowel and every consonant must be incorporated into a syllable, that is, licensed by a vowel. Typical syllable structures are illustrated in (4), where V^0 is the nucleus, V^1 the rhyme which optionally dominates a C-category (coda), and V^2 the syllable category, which obligatorily dominates a C-category (onset).

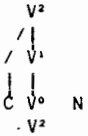
(4) a.



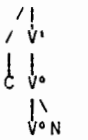
b.

In the unmarked case, a vowel in French licenses consonants in onset position only as in (4b). This property of French syllables is consistent with nasalization and deletion phenomena such as those illustrated in (1-3). In particular, nasalization is analyzable as the adjunction of the nasal consonant to the preceding vowel position in the case where there is no following vowel to license it as onset. Deletion of obstruents in the same context follows from their not being licensed as onsets to a following vowel and their inability to adjoin to a preceding vowel. Thus, nasalization is represented as the adjunction of a syllable free nasal segment as in (5a) to a syllable nucleus position as in (5b).

(5) a.



b.



Adjunction creates a new category, identical to the one which dominates the vowel, which now dominates both the vowel and the nasal, that is, a new syllable nucleus which is a nasalized vowel.

A salient fact about the analysis outlined above is that it requires no language particular rules. The theory of syllable structure, where vowels project syllables and consonants must be in positions licensed by vowels, is assumed to be universal. The fact that languages divide up into CV and CVC types is a general parametric division established for each language on readily available data. The CV type language appears to be the unmarked case. Thus, for example, the child would need evidence that vowels licensed following consonants to set the CVC parameter for his language. The kind of evidence that is salient for this setting would include geminate clusters of consonants, which by definition require that one be coda, the other onset of separate syllables.

The only rule that is required by the analysis is a general (presumably, universal) adjunction rule. That is, a rule which says adjoin anything to anything else within the limits set by the inherent categories of segments and the universal principles of syllable structure. By definition adjunction does not create new syllable structure positions, it merges a syllable-free segment into an already established position. Thus, adjunction is structure preserving. The limits set by the inherent properties of segments seem to be general compatibility limits. Thus, a nasal consonant is not incompatible with a syllable nucleus position through adjunction, but an obstruent is generally incompatible with a vowel nucleus position under the same circumstances. In traditional terms, a nonvocalic segment cannot be adjoined to a vocalic segment, and vice versa.

Japanese is also a CV language. However, in Japanese, sequences of two consonants across a syllable boundary are allowed under quite specific conditions.

These conditions are illustrated in the following verb forms.

(6)	Present	Past
a.	taberu	tabeta
b.	wakaru	wakatta
c.	yomu	yonda
d.	yobu	yonda
e.	toku	toita
f.	togu	toida
g.	hanasu	hanasita

In the past forms (6b,c,d) there are sequences of two consonants (tt, nd, nd). In each case the features for place, voice and continuance form a kind of long component across the sequence. In addition, if the first consonant is voiced, it is also nasal. In (6a), the stem is tabe and the endings are ru in the present and ta in the past. In the past tense, the remaining examples show ta or its voiced alternate da. In the present the r of the ending seems to merge into the final consonant of the stem to which it is attached. Thus, in the present tense cases (6b) through (6g), the final consonant of the stem surfaces before the residual u ending. In (6e,f), in the past, stem final k and g are lost and the vowel i emerges in their place. In (6g), in the past, the vowel i emerges after the stem final consonant and palatalizes it.

I assume that in (6a) the basic forms of the present and past tense suffixes surface following a vowel final stem. Thus we get the analysis in (7).

(7) tabe+ru, tabe+ta

The tabe- case illustrates the basic CV pattern of Japanese where every vowel licenses a single consonant to its left as onset. It is where this pattern is perturbed in the remaining examples of (6) that rather complex adjustments in segmental structure come into play. The perturbation is caused when consonant initial suffixes are added to consonant final stems creating a sequence C_1, C_2 with no intervening vowel. In the case of the present tense suffix, the r appears to merge completely with the final consonant of the stem which precedes it. Assume that r in this case is a maximally unmarked liquid, a kind of archisegment R such that it is nondistinct from any other consonant and that this accounts for its chameleon like nature. Then, if the final consonant of the stem is syllable free and adjoins to R in the position licensed by the following vowel, we get the analyses in (8).

(8)			
	b.	wakar+Ru	wakaru
	c.	yom+Ru	yomu
	d.	yob+Ru	yobu
	e.	tok+Ru	toku
	f.	tog+Ru	togu
	g.	hanas+Ru	hanasu

In the past tense, the examples in (6e,f,g) suggest that the suffix is -ita. In (6g), the stem final s is syllable free and adjoins to the initial i of the suffix which palatalizes it. In (6e,f), the stem final velar consonant is syllable free and adjoins also to the initial i of the suffix. In these cases, however, we get velar elision, a not uncommon result of the adjunction of a velar consonant to a high, front vowel. For example, in Kasem, also a CV type language, nouns suffixed by a are singular, by i are plural as in bakada, bakadi. In stems that end in a velar consonant such as dig-, the velar is preserved in the singular before a (diga), but lost in the plural before i (di, from intermediate di+i).

In the remaining cases of the past tense in Japanese illustrated in (6a,b,c,d), the i of the suffix deletes. Where the stem is vowel final as in (6a), suffix i must adjoin to it since it has no onset. Since, i is nondistinct from the stem final vowel, it must be a relatively unmarked vowel. Where the stem ends in a consonant as in (6b,c,d), the suffix i deletes and the final stem consonant assimilates for place to the t of the suffix. In addition if the stem final consonant is voiced it voices the suffix t and is itself realized as a nasal.

The interesting fact about these last cases is that Japanese allows the CVC structure just where the coda C has an identity relation with the following onset. The identity relation appears to involve the agreement of voice, continuance and place, the long component mentioned above. The intervening i is retained in the (6e,f,g) cases because the velars in (6e,f) do not agree with the following dental for place and the s in (6g) does not agree with the following stop for continuance. If this agreement pattern represents a kind of adjunction of the stem final, syllable-free segment to the licensed suffix initial consonant, then we have the suggestion of an explanation for this interesting array of assimilations and deletions. The suffix initial i appears to be a defective vowel which will delete wherever it can (or alternatively, an epenthetic vowel which is inserted only where it must be). The conditions surrounding the deletion of i are linked to the possible adjunction of the consonants which surround it. Where the adjunction is possible because the consonants are compatible with respect to voice, place and continuance, then i can delete. Where they are not compatible, i cannot delete. Alternatively, in an epenthesis analysis, i is inserted where the stem final consonant cannot adjoin to the suffix initial consonant becuase of an incompatibility.

SOME PHONETIC BASES FOR THE RELATIVE MALLEABILITY OF SYLLABLE-FINAL VERSUS SYLLABLE-INITIAL CONSONANTS

S. Y. Manuel

Communication Disorders and Sciences
Wayne State University, Detroit, Michigan USA

ABSTRACT

Syllable-final consonants seem to be more susceptible to assimilation and deletion than are syllable-initial consonants. We are using instrumental data to augment earlier claims about the phonetic bases of such asymmetries.

1. INTRODUCTION

It is often noted that syllable-final consonants are more malleable (subject to deletion, lenition and assimilation) than syllable-initial consonants [1,3]. If speech sounds weren't made by a real mouth, then V-C sounds might be mirror images of C-V sounds, and there would be no phonetic bases for such asymmetric phonological behaviors. However, speech sounds *are* produced by real vocal tracts, and there *are* certain asymmetries in the acoustics of movements into and out of consonant closures, as outlined by Ohala and Kawasaki [8]. Some of these asymmetries are simply due to the way aerodynamic forces change over time, given symmetric movement into and out of consonant closure. Other acoustic asymmetries may be due to the relative timing of oral vs. velar and/or glottal gestures, differences which result in different overall vocal tract shapes at closure implosion versus closure release. Importantly, these asymmetries are such that various features of consonants should be more salient at releases than at implosions.

Here we compare some of the acoustic characteristics of particular consonants, in order to determine possible sources for certain phonological asymmetries. Thus, we are following John Ohala and others [2,4,5,7,8,10,11] who have sought to find, in the architecture and acoustics of the vocal tract, explication of (at least some) phonological behaviors.

2. SYLLABLE-INITIAL VERSUS SYLLABLE--FINAL STOP CONSONANTS

2.1 Syllable-initial stop consonants

During the closure interval of an oral stop consonant, air pressure builds up in the oral cavity as air flows through the glottis into the mouth. At the release of the stop, a brief burst occurs at the oral constriction. This noise source, which has an abrupt onset, has cue value for manner feature - generally it is absent in sonorant consonants and it does not have an abrupt onset for fricatives. The burst is filtered by the cavity in front of the constriction, and consequently its spectrum varies as a function of the place of constriction. As the oral cavity continues to open, the primary acoustic source switches to the larynx. This laryngeal source is also filtered by the changing oral tract shape, producing consonant-to-vowel formant transitions. Since the formant transitions reflect the changing oral tract configuration from the constriction to the following vowel, they too are strong cues as to the place at which the oral constriction was made. The rate of the transitions, determined by the rate of the articulatory movements, is a further cue to manner, as stop-vowel movements are much more rapid than glide-vowel movements [6].

If the vocal folds are sufficiently adducted at the release of the oral constriction, the vocal folds will vibrate immediately following the release and the laryngeal source will be periodic. If the glottis is appreciably open at the moment of oral release, the laryngeal source will be aspiration noise until the vocal folds have adducted sufficiently to permit glottal vibration. The acoustic correlates of aspiration and the way in which this source couples with the oral tract has the consequence that the first formant is very weak in amplitude, and the upper formants are noise-excited [9]. These differences in source characteristics show up in the formant transitions, and the transitions at the release of a stop consonant are cues to the voicing feature for the consonant.

Thus syllable-initial consonants have rich information in that (1) **bursts** provide a good source of information for manner and place and (2) **formant transitions** provide a good source of information for manner, place, and voicing.

2.2 Syllable-final stop consonants

On the other hand, the acoustic consequences of movement from a vowel *into* a stop consonant constriction are less rich. As the talker moves from a vowel into a consonant, in most cases there is little or no noise generation, but formant transitions are observable as long as the glottal source continues during the movement into the constriction. However, unlike releases of consonants into vowels, movement from vowels into consonants does not entail a burst. As noted above, the burst at a C-V release is due to pressure buildup in the mouth, but this pressure buildup occurs precisely because the oral tract is closed during the consonant, and therefore is not relevant to the movement *into* the constriction.

In syllable-final position, manner distinction between the glides and stops may not be as well maintained as it is in syllable-initial position, since there seems to be a tendency for gestures to diminish at end of syllables [5].

2.2.1 Implementing voicing distinctions for syllable-final stops may put place distinctions at risk.

If fully voiced, the transitions from a vowel into a stop closure provide little information about voicing of the consonant, since they should be identical for both voiced and voiceless consonants. If these consonants are unreleased (as they often are when utterance-final or when followed by another consonant), then how are voicing distinctions maintained? Languages use several strategies, including cutting off voicing very near oral closure for voiceless stops by:

- opening the glottis, or
- constricting the glottis (making a glottal stop)

If timed to occur by the time oral closure is achieved, either of these devoicing gestures would aid in preventing vocal fold during the beginning of the oral closure period.

In English, it is quite common to achieve devoicing for syllable-final voiceless stops by making a glottal stop very close to oral implosion [3]. Note that if this glottal stop is timed too late with respect to oral closure, then it won't serve to distinguish voiced and voiceless stops, since at the moment of oral implosion the vocal folds would still be in a configuration suitable for vibration. To ensure devoicing, a better strategy would be to error on the side of making the glottal stop relatively early.

However, if the glottal stop is timed *too* early with respect to oral closure, the formants won't be excited during the V-C closing movement. This in turn would result in loss of place information. Eventually speakers may not bother to make the oral gesture at all, and this presumably explains the development of dialects that have glottal stops instead of oral stops in final position.

Figures 1a-c show acoustic data from a study [7] which used simultaneous electropalatographic (EPG) recordings, fiberoptic views of the larynx, and acoustic recordings, to illustrate this point. In each of the figures is shown the F1 movements toward the end of an English word that phonologically ends in /t/. What we are interested in is whether or not F1 falls toward the end of the vowel which precedes the /t/, as F1 fall is an acoustic consequence of oral closure (F1 fall can clearly be seen in the reference word that ends in /d/).

The data in Figure 1a are from a speaker who heavily glottalized his /t/, but who actually made an alveolar

closure at about the acoustic end of the vowel, as evidenced by our EPG data. Acoustic evidence of the oral closure can be seen in the F1 fall.

In contrast, Figure 1b shows the same speaker's production of a /t/ in *absolute utterance-final* position. EPG data indicated that while the speaker *did* make an alveolar closure, he made it very late with respect to the end of the vowel. Fiber-optic data indicate that in this case devoicing was achieved with a glottal stop. As can be seen by the slight fall in F1, there is some acoustic evidence that at the end of the vowel the speaker is beginning to make an oral constriction.

Figure 1c shows data from a different speaker. The EPG signal indicated that this speaker did not make an oral constriction for the /t/ in the target utterance. Rather, the vowel was terminated solely by making a strong glottal stop. And, we see very little fall in F1, which is expected since the speaker *in fact did not make an oral closure* (there may or may not have been some residual of a *weakened* oral closing gesture, one that did not result in apico-alveolar closure).

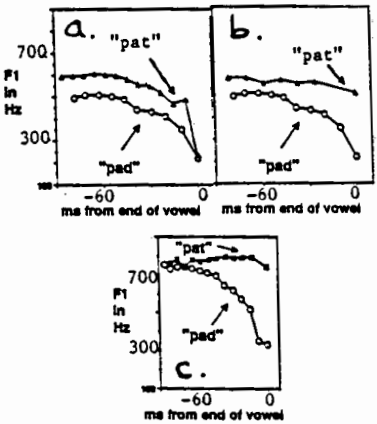


Figure 2a-c. Acoustic evidence of presence or absence of oral closure prior to cessation of vocal fold vibration.

3. SYLLABLE-INITIAL VERSUS SYLLABLE-FINAL NASAL CONSONANTS

During the oral closure period of a nasal consonant, the mouth is closed and the velum is lowered, with the acoustic consequence that much of the sound is filtered and propagated through the nasal cavities. When the oral occlusion is released, there is a sudden increase in the amount of energy in the sound as the principal sound output switches to the mouth opening. This increase is especially marked in the F2 region for labial and alveolar consonants, since the nasal murmur tends to have weak energy in this region. Conversely, with the velum down, oral implosion will result in a sudden *decrease* in the amplitude in the F2 region. This sudden change in F2 is a cue to the implosion or release of the consonant, and the F2 frequency indicates the place of articulation for the consonant.

The amount by which the amplitude of the F2 peak changes as a function of opening or closing the oral cavity is expected to be dependent on the degree to which the velum is lowered [10]. If the velum is very low, then a high percentage of the energy will go through the nose, regardless of whether or not there is a change in the oral constriction, and the resulting change in the amplitude of F2 will be relatively small as the oral constriction is made or released. However, if the velar-pharyngeal opening is somewhat smaller, then the sudden change in the oral constriction will have a very large effect. Thus the consonant/nonconsonant (oral constriction vs. no oral constriction) distinction is most clearly maintained when the velum is not too low.

Several studies [e.g. 11] suggest that the velum is generally lower at the time the oral constriction is being made for a syllable-final nasal consonant than it is at the time oral constriction is released for a syllable-initial consonant. Thus we might expect that going into a syllable-final nasal consonant there is less of a V-C demarcation than the C-V demarcation we get when moving from a nasal consonant into a vowel. If this is the case, then syllable-final nasals might be expected to delete more often than syllable-initial nasals, since they would be less salient to listeners.

An example is shown in Figures 2a and 2b. Figure 2a shows smoothed spectra for the 30 ms periods preceding

(solid line) and following (dashed line) the nasal release of [m] in [#mIb#]. There is a 21 dB increase in the F2 region from the spectrum of the nasal murmur to the spectrum of the vowel.

Figure 2b shows spectra for the periods preceding (dashed line) and following (solid line) the oral implosion for the [m] in [#bIm#]. Here there is only a 14 dB difference between the vowel and the consonant, in the F2 region. We have observed similar patterns for a number of utterances and speakers. We are proceeding to quantify these differences, and crucially, we will be looking at the acoustic signal of utterances for which there is accompanying velotrace [5] date to indicate velum height.

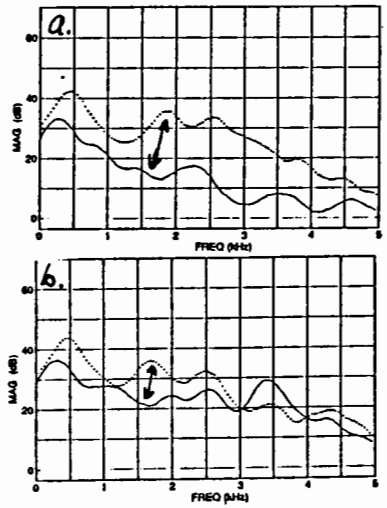


Figure 2a-b. Spectra before and after /m/ closure and release. Arrows point to F2 region.

4.SUMMARY

These are just a few of the differences between the acoustic consequences of making and releasing consonant constrictions. As we have seen, some differences, such as the presence or absent of a burst, are simply due to physics. Other asymmetries are due to the ways in which velar or glottal gestures are used to implement the features voicing and manner. Implementation of these features may involve different or differently timed gestures, depending on syllable position. These articulatory asymmetries can put at risk the saliency of certain other features, particularly in syllable-final position. Listeners may not hear, for example, that a speaker has actually made a consonant closure, or a closure at a particular place. These listeners might reasonably assume that a consonant was not made (deletion), or that it was made at some place other than what the articulatory facts would have revealed (assimilation). When those listeners take their turn at speaking, they may articulate in the way they *assume* other speakers do - omitting, weakening, or assimilating syllable-final oral gestures.

5. REFERENCES

[1] Bell, A. & Hooper, J. B. "Issues and Evidence in Syllabic Phonology", in A. Bell & J. Hooper (Eds.) *Syllables and Segments* (pp 3-24). Amsterdam: Holland.

[2] Browman, C. P., & Goldstein, L. (1990), "Articulatory Gestures and Phonological Units", *Phonology, 6*, 201-151.

[3] Brown, G. (1977). *Listening to spoken English.* London: Longman.

[4] Kingston, J. (1985) *The Phonetics and Phonology of the Timing of Oral and Glottal Gestures,* Univ. Calif. Berkeley Dissertation

[5] Krakow, R. A. (1989), *The articulatory organization of syllables: A kinematic analysis of labial and velar gestures,* Yale University Dissertation.

[6] Liberman, A. M., Delattre, P. C., Gerstman, L. J., & Cooper, F.S. (1956) "Tempo of frequency change as a cue for distinguishing classes of speech sounds", *Journal of Exp. Psych., 52,* 127-137.

[7] Manuel, S.Y. & Vatikiotis-Bateson, E. (1988) "Oral and glottal gestures and acoustics of underlying /t/ in English", *J. Acoust. Soc. Am., 84, Suppl 1,* S84(A).

[8] Ohala, J. J. & Kawasaki, H. (1984), "Prosodic phonology and phonetics", *Phonology Yearbook 1,* 113-128.

[9] Stevens, K.N. (1990) "Noise at the glottis during speech production", *J. Acoust. Soc. Am, 87,* Suppl 1.

[10] Stevens, K. N. (forthcoming) *Acoustic Phonetics.*

[11] Ushijima, T., & Sawashima, M. (1972), "Fiberscopic observation of velar movements during speech", *Annual Bulletin No. 6, RILP,* Univ. of Tokyo.

L'INTERPRETATION PHONOLOGIQUE DES SEGMENTS DIPHTONGOIDES

T.Tchalakova

Institut pédagogique supérieur
Choumen, Bulgarie

The study shows that in the basis of biphonemic identification of the glide segments Ibi, Ibĭ for Bulgarian and Russian speakers is the use of different acoustic and linguistic factors arranged in graduation respectively.
Zones of similarity and identification for Bulgarian and Russian speakers are determined.

I.LE BUT principal de cette communication est la découverte d'opérations applicables lors d'une interprétation phonologique des segments diphtongoïdes Ibi, Ibĭ des locuteurs russes et bulgares, ainsi que la détermination des paramètres acoustiques principaux pour leur identification biphonémique.

2. MATERIEL et méthodologie de la recherche.

Des syllabes du type IbiI, Ibĭi, isolées des mots, ont été utilisées comme matériel pour l'analyse acoustique et perceptive. Un annonceur (speaker) russe qui possède les normes prononciatives actuelles. Les syllabes ont été évaluées par des Bulgares et des Russes - 20 auditeurs de chaque langue par groupe. Les données reçues ont été traitées par la méthode de l'analyse dispersionnelle.

3.I. Un travail préalable montre que les Bulgares marquent le son vocal IbiI dans la syllabe Cbi comme IəI, Ibi, Ibă, Ibŭ. La fragmentation de IbiI en deux segments est probablement déterminée par une identification du segment transitoire de IbiI au son vocal IəI dans la langue bulgare, tandis que son segment en IiI est identifié respectivement avec Iăi ou Iŭi. Dans 75 - 85% des cas les syllabes du type Ibiă sont déterminées par les Bulgares comme trisegmentaires.

3.2. L'analyse dispersionnelle des ensembles à un facteur - pour les indices qualitatifs /2/, présente la possibilité de fixer le poids de chacun des facteurs linguistiques /3/ qui influencent les variations de la valeur des segments de la parole /tableau I/.

Le poids des facteurs montre que les paramètres acoustiques principaux fixant l'identification biphonémique des segments diphtongoïdes chez les Bulgares sont l'intervalle FII stationnaire - FII initial /ΔFII/, ainsi que la longueur des segments diphtongoïdes Ibi ou Ibĭă /s/, le rapport du secteur transitoire et du secteur en IiI est d'une valeur très proche /tableau I/

Tableau I. LE POIDS DES FACTEURS η^2_x ET DES RANGS R

LES FACTEURS LINGUISTIQUES	Bulgares η^2_x	R	Russes η^2_x	R
Qualité du consonnant suivant /palatalisé- non-palatalisé/	10,4%	6	4%	5
Présence - absence de /ĭ/	9%	5	27%	6
Qualité du consonnant précédant /labial - non-labial/	4%	I	0,9%	4
Début - fin du mot	6,6%	4	0,3%	2,5
Syllabe accentuée-non-accent.	4,6%	3	0,3%	2,5
Quantité des syllabes dans le mot	4,5%	2	2,5%	3
$\sum \eta^2_x$	39,1%		35%	
Quantité des facteurs principaux		6		6
LES FACTEURS ACOUSTIQUES	η^2_x	R	η^2_x	R
La dimension (grandeur) FII	20%	2	5%	I
La longueur du segment diphtong.	28%	3	10%	2
a/longueur du secteur transitoire/	26%		7%	
b/longueur du segment en IiI	23%		20%	
La dimension FII - final	3%	I	17%	3
$\sum \eta^2_x$	100%		59%	
Quantité des facteurs principaux		5		

Si nous prenons l'espace bidimensionnel entre l'axe des abscisses - x /s ms/ et l'axe des ordonnées y/ΔFII Hz/, nous notons que plus de 60% de la valeur biphonémique se localise dans ΔFII≥ 500 Hz et s≥ 80ms /fig.I/. Dans des paramètres inférieurs à ceux cités ci-dessus, mais pas en-dessous de ΔFII=450 Hz et s=40 ms, la valeur varie de I5 à 55% et dépend sans doute aussi de la capacité de l'ouïe d'identifier les deux segments - le segment transitoire et le segment en IiI - /I/.

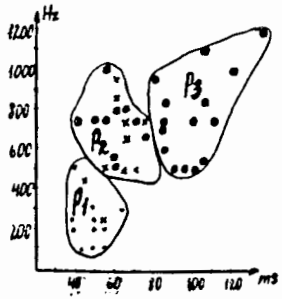


Fig.I.Valeur biphonémique du segment diphtongoïde PI 0-I5%,P2 20-55%,P3 60-85%

Là il faut noter que dans l'étendue P3 se trouvent des réalisations de syllabes /Cbl/ ainsi que des réalisations de syllabes Cbшl.

3.3.La fixation du son I ш I par les auditeurs russes dans les syllabes isolées Cbl-de 5 à 40%,est observée uniquement dans des cas isolés,en position devant des consonnes palatalisées et en fin du mot. Dans les syllabes Cbшl la valeur varie de 25 à 95%.Comme le bruit produit par /ш/ dans ces syllabes est très faible et enrégistré sur un spectrogramme seulement dans l'une des réalisations des syllabes Cbшl,il est évident que les Russes emploient aussi d'autres paramètres acoustiques très importants.Prenant en vue le poids des facteurs linguistiques,nous pouvons supposer qu'une information pour la présence ou l'absence de I ш I est comprise dans la nature de la liaison intersyllabique,plus précisément le niveau de FII dans le deuxième secteur transitoire est informationnel. L'analyse dispersionnelle des facteurs acoustiques a montré que FII-fin. exerce la plus grande influence sur l'hiérarchie des autres facteurs.Sur la figure 2 est marquée la valeur biphonémique des segments diphtongoïdes par les Russes dans

les coordonnées x-/longueur du segment dipht.en ms/, y-/niveau en Hz de FII-fin/

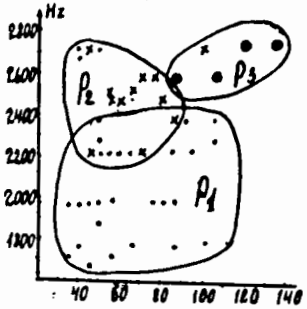


Fig.2.Valeur biphonémique du segment diphtongoïde PI=0-I5%,P2=20-55%,P3=60-95%

L'évaluation des syllabes à 2 segments commence dès FII-fin.≥2250Hz et s≥45ms; dans cette étendue se situent les réalisations de /кы/et de/кый/ également. Comme bisegmentaires d'une plus grande certitude/étendue P3/ sont marquées ces réalisations des syllabes /кый/ qui se caractérisent avec FII fin.≥2600 Hz et s≥ 80 ms.

Il s'avère que pour la perception de I/ы/I par les Russes,lors de l'absence ou de la faible réalisation du bruit de /ы/,il est nécessaire que la valeur finale de FII soit égale ou plus grande que FII-stat., tandis que la longueur du secteur en IiI soit suffisante /tableau I/.

4. Sur la base de ces données peuvent être faits les raisonnements suivants: Pour les locuteurs des deux langues,certains facteurs linguistiques et acoustiques sont très importants,en déterminant la segmentation du courant sonore et en formant des rangs d'une gradation différente.Il est possible d'établir ainsi les différences dans les opérations que les Bulgares et les Russes exécutent au niveau syllabique.Ces facteurs marquent la zone de ressemblance et la zone d'identification des segments diphtongoïdes /кы/,/кый/ pour les locuteurs des deux langues.

Bibliographie:

Ii.I FANT,G.(I964),"Acoustic theory of speech production",The Hague:Mouton& Co.

I2.I ПЛОХИНСКИЙ Н. "Алгоритмы биометрии", Москва.

I3.I STERN,A.(I987),"Linguistic factors in speech perception",Vol.I,Tallinn.

LENITION PROCESSES AND 'GLOBAL PROGRAMMING PRINCIPLE'

T. Szende

Linguistics Institute, Budapest, Hungary.

ABSTRACT

A general abstract principle referred to as the 'global programming' is considered which universally governs (i) lenition processes, (ii) slips of the tongue, and (iii) child language restrictions turning their underlying PRs into defective realizations. GPP claims: defective realizations come about due to the distortions of the (next-to-phonetic) phonemic representations (i) by replacing a fine-controlled scaling in the speech production by an overgeneralized plan or (ii) by eliminating constituents the phonological system of a language is made up of. These processes, (i) rule restriction and (ii) elimination of constituents result in narrowing the information space in which an underlying PR is posited.

It is the purpose of this paper, (i) to give a general account of (a polynomial word-level) 'phonological representation' considerably different from that as conceived of by post-*SPE* phonologies. Second, (ii) to argue that the multidimensional interrelationship existing between a low-level PR and a (next-to-phonemic) phonetic representation may be best described in terms of what I call the Global Programming Principle. Finally, (iii) to show, and to exemplify on Hungarian, what kind of consequences the GP hypothesis might have on our conception of rules and rule domains.

(i) Word-level phonological representations (PRs) are taken to be stratified abstract objects. One and the same representation may assume various forms at the different levels of abstraction, depending on the actual perspective taken. As I exemplified on a Hungarian word form (*lássa* 'see' Imp1PSg, cf. [14], 132) submitting it to analyses (= segmentations and structural parsings) of varying depth both in a syntactico–morphophonological and a phonemic perspective, we get different, but equally valid, results in terms of the set of primitives as well as their arrangement. In this instance the stratified PR includes following layers:

a. /la:ʃʃa/

b. //#la:t#/-v-/#j#/-*ms*-/#j(a e)#/p

c. ///#latV*back*#//=//#sV#/-
-/# C(a e)#///

(Needless to say, (a), (b), and (c) are to be considered as partially independent abstract forms; they do by no means appear to be the variants of one and the same basic word form.) The different results we get will be interrelated and derivable from one another.

In the example, above, (b) is related to (a) via morpheme structure rules and phonological rules; the same obtains for (c) and (b) with the proviso that this relation may involve non-productive morpheme structure rules and phonological rules (along with productive ones). Whereas (c) obviously has no role in speech production, i.e. it is 'extra-conscious' with respect to both speaker and listener, (b) is an active component of the speaker's mental processes at a 'pre-conscious' (*vorbewusst*) level, i.e. as a piece of unconscious knowledge that can be elevated to a conscious status and, as such, it may acquire surface realization in special communicative situations (e.g. in spelling).

The level of (a) of PRs generally, and in the above example /la:ʃʃa/ for *lássa*, must be invariant, i.e. discrete and of a constant form, whereas the corresponding word form in actual speech production is not. A set of interface rules must therefore be assumed to mediate between PRs and implementations. In particular, I will assume that two types of interface rules, viz. 'levelling' rules and '*gestalt*' rules, will operate on phonological base forms.

'Levelling' rules will effect transformations like /la:ʃʃa/ ⟹ la:ʃ:a (where ⟹ indicates level shift, and the omission of / / is meant to reflect the fact that the form right of the arrow is neither phonological nor phonetic: rather, as a realizational program, it is an independent category constituting an intermediate level between those two). So we here will have to accept the assumption that morphemes are psychologically real, even if we cannot actually specify to what extent linguistic elements can be taken to be isomorphic with psychological facts (cf., e.g. [5], 10–12). In the above example, levelling rules turn a type (a) pre-implementational, intermediate phonological representation into the corresponding next-to-phonemic phonetic representation by removing the morpheme boundary feature from between /t/+/j/ and replacing /ʃʃ/ by ʃ: via a pronunciation subroutine, in a way corresponding to the mechanism involved in the notion of Kiparsky's [4] Bracket Erasure.

(ii) However, the form *lássa* ⟶ la:ʃ:a will also undergo further operations, including the relativization of the [+long] component of /a:/. This follows from one of the *gestalt* rules – that of temporal organization –, a set of rules whose common property is that they always involve a portion of an utterance as a whole. A correspondence like /a:/ – [a·] or [a], in fact, cannot be interpreted in terms of isolated segments if we wish to maintain the criterion of biuniqueness. The motivation for a derivation /a:/ ⟶ [a]~[a·] can only be found in the structural effect of a word form as a whole, in the present case most immediately in the architecture -*V:C:*-, in particular, the occurrence of ʃ: after a:, i.e. a (temporal) foot organization factor. The main properties of *gestalt* rules (omitting details) are as follows. (ii/a) *Gestalt* rules determine the utterance unit in speech production in a global way. This is unambiguously shown, in terms of my own experimental results, by 'sequence reduction' and 'sequence size truncation'. Another type of evidence comes from the stage of a child's first language acquisition where non-adult, "crude" programming with respect to a given word form results in a disorderly arrangement of the articulatory components involved, one that does not match the order imposed by the phonological base form. For instance, Smith's [10] data include *squat* surfacing as [ɡap], *queen* as [ɡi:m], etc. by transposition of the bilabial component (cf. also [17], 411). (ii/b) The units undergoing *gestalt* rules may be of various sizes. They may involve single morphemes, but also several, semantically connected word forms (the latter case is observable primarily in sequence size truncation). (ii/c) In lenition processes, *gestalt* rules may exhibit varying effectiveness in modifying individual articulatory elements within a global articulatory program. For instance, of several units within a single word form, all of which are underlyingly specified as the same phoneme, e.g. /k/$_{1-3}$ in *gyerekeknek* 'children+Dat.', some will, and others will not, lose the element involved in the lenition process, in this case the stop component. This depends on the phonotactic position of the unit in question, the genre and the tempo of speech, the degree of lenition, the phonetic makeup of the segment, and so forth. In addition, it also depends on the component itself; in vowel substitution errors, according to Shattuck-Hufnagel's data [9], esp. 124), the standard deviation of the feature [+tense] in erroneously substituted items exceeds the expected probability values several times more than that of [+back].

GP is made up by the totality of *gestalt* rules. The question of what sort of a cortical equivalent might be ascribed to GP is difficult to answer. Anyway, linguistic signs and processes are still considered to be best described, in terms of the functional hierarchy of the operation of language, by the

model first proposed by Wernicke [16]. In essence, psycholinguistics also traditionally accepts this three-step mediation model as one that confirms the authenticity of *gestalt* theories exactly "in the realm of perceptual organization" (see e.g. [7], 146). In Wernicke's model, the levels wedged in between sensorium and cognitive representation are a bilaterally-connected "representation of specific 'gestalt' elements" and, on the speech production side, a "representation of motor commands (concepts of movements)" (cf. Creutzfeldt [1], 5).

(iii) The domain of application of GP including *gestalt* rules is, obviously, phonetic implementation, especially that of lenition processes. (Remark: the concept of 'lenition' as exposed in its classical form in Natural Phonology, cf. [12], [2], etc. does in fact not cover the whole typology of phenomena occurring in spontaneous speech production. On the basis of a collection of data taken from Hungarian a larger-scale typology may be established when also lenition processes manifesting themselves in sequence size portions of speech are taken into consideration, cf. [15].)

In the rest of this paper my main concern will be the way *gestalt* rules fit into the rule hierarchy (P-rules, MP-rules, MS-rules, etc.) that is amply discussed in post-*SPE* phonologies (cf., e.g. [11], [3] [6]). In terms of a typology of lenition processes observable in Hungarian, *gestalt* rules fit into this traditional classificatory pattern rather badly. The facts are as follows. (iii/a) One particular lenition type, covering a set of essentially identical changes, may equally embody rules of diverse categories. 'Reduction', for instance, may simply be a change that we normally classify as a phonetic rule: the slight delabialization of *a* in *változása* 'its change' calls for that label. In other cases, reduction results in a change that can be characterized as a rule of phonological nature in that, by deleting a phonologically relevant feature, it alters the phonological status (e.g., class membership) of a segment as in m $\longrightarrow$ w̃ (*mondták* 'say' Past3PPl). By eliminating a major classificatory feature, the realization may turn into the phonological base form of another lexemic alternant: by devoicing *u* in *azután* 'then' we get a result like *az*V*tán* which appears to be the 'fortis' version of *aztán* 'idem' (cf. [13]). (iii/b) There is not a complete and mutual overlap in that all types of lenition permit the occurrence of all possible rule categories. 'Truncation', for instance, is by definition a phonological category, not a phonetic one; indeed, there are clear examples (e.g. *szóval* 'in other words' — [so]) to show that truncated forms may fail to exhibit any further phonetic change (the omission of suffix being obviously not an instance of reduction). In other cases, it must be admitted, truncation and phonetic change may simultaneously occur within a single sequence, e.g. *valami ilyesmi* 'something like that' — [vɔ̃mĩjeʃmĩ] where final *i* undergoes reduction by centralization and changes in height and degree of illabiality. Consequently, the notions of truncation and phonetic rule are mutually exclusive. As for 'deletion' and 'loss', both lenition process types destroy a complete segment at the actual point in PR. The rules effecting these processes are undoubtedly of a non-phonetic character; but they may either be phonological like in cases of *t*-elision, e.g. in *ezt* 'this+Acc.', or result in morpholexemic switch as in the various versions of *miért* 'why' (cf. [13], 182). (iii/c) Scope properties are also non-relevant for the classification of lenition rules. Larger-scope processes, i.e. those involving a sequence of adjacent segments, can be realised by phonetic rules (such as sequence reduction) as well as by morphophonemic or morpholexemic ones (as detailed above for cases of truncation). On the other hand, lenition phenomena involving single segments can also qualify as instances of any of these three rule types. (iii/d) Finally, it is appropriate to point out that rules responsible for lenition processes may also lead to results that do not lend themselves to a neat interpretation in terms of a linguistic system-oriented classification. Whenever sequence size truncation yields a realization that further undergoes elimination of backness contrast in a vowel – as in *ôtkor* $\longrightarrow$ *ôtkôr* '5 o'clock' with o $\longrightarrow$ *ɸ* – the speaker in fact (over)applies vowel harmony in a way that, in terms of various lines of reasoning, can be taken to be of a phonetic, or morphophonemic, or (potentially) morpholexemic character.

The lack of correspondence between phonetic, morphophonemic and morpholexemic rules on the one hand and the set of *gestalt* rules on the other is conspicuous enough to make one wonder if those two systems of rules actually occupy different levels within the total system. However, the source of such mismatch is not that their structural descriptions reveal rule-governed phenomena of different depth: it is not the case that the former set of rules refer to phenomena restricted to underlying form and the latter account for events at some level intermediate between underlying and surface representation. (Aphasiacs' errors, in particular cases of syllable elision as in *catholicize* – /kæθəlāyz/, *solidification* – /saləfəkēyšən/, demonstrate that syncope applies to underlying form, not (some level of) surface representation, cf. [8], 24–29.). Rather, the difference actually lies in the fact that the rules categorized as above and *gestalt* rules can be stated for (a typologically diverse range of) allegro phenomena, whereas rules of the former type cover lento forms only. All that this distinction entails in itself, however, is that the number of *gestalt* rules is larger. But the *punctum saliens* of the comparison is that *gestalt* rules refer to sequences (utterance units) as wholes, whereas traditional types of rules refer to segments or concatenations of segments appearing between boundary features, even if their structural descriptions involve boundary features themselves as well. So, *gestalt* rules represent an independent category of rules; cover a set of phenomena exhibiting higher variability; and, consequently, phonetic, phonological and morpholexemic rules can, to a significant extent, be logically subordinated to them.

REFERENCES

[1] CREUTZFELDT, O. (1987), "Inevitable deadlocks of the brain–mind discussion." GULYÁS, B. (ed.): *The brain–mind problem*, Leuven, 1–27.

[2] DRESSLER, W (1984), "Explaining Natural Phonology", *Phonology Yearbook, 1*, 29–50.

[3] DRESSLER, W. (1985), *"Morphonology. The dynamics of derivation"*, Ann Arbor, Michigan.

[4] KIPARSKY, P. (1982), "Lexical morphology and phonology", *Linguistics in the Morning Calm*, Seoul, 3–91.

[5] LINELL, P. (1979), *"Psychological reality in phonology"* Cambridge, London, New York, Melbourne.

[6] MOHANAN, K. (1986), *"The theory of Lexical Phonology"*, Dordrecht, Boston, Lancaster, Tokyo.

[7] OSGOOD, C. (1963), "On understanding and creating sentences", *American Psychologist 18*, 735–751.

[8] SCHNITZER, M. (1972), *"Generative Phonology – evidence from aphasia"*, University Park, Pennsylvania.

[9] SHATTUCK-HUFNAGEL, S. (1986), "The representation of phonological information during speech production planning: Evidence from vowel errors in spontaneous speech", *Phonology Yearbook 3*, 117–149.

[10] SMITH, N. (1973), *"The acquisition of phonology: A case study"*, London–Cambridge.

[11] SOMMERSTEIN, A. (1977), *"Modern phonology"*, London.

[12] STAMPE, D. (1973), *"A dissertation on Natural Phonology"*, New York, London.

[13] SZENDE, T. (1988), "A note on morphophonological alternations in Hungarian", *UAJb–Ural-Altaic Yearbook 60*, 177–182.

[14] SZENDE, T. (1989), *"Phonological representation and 'Global Programming'"*, *Magyar Fonetikai Füzetek, 21*, 132–135.

[15] SZENDE, T. (1992), *"Alapalak és lazítási folyamatok"* [Phonological representation and lenition processes], Budapest. (Forthcoming.)

[16] WERNICKE, C. (1894), *"Grundriss der Psychiatrie"*, Leipzig.

[17] WILBUR, R. (1981) "Theoretical phonology and child phonology: Argumentation and implication", GOYVAERTS, D. (ed.): *Phonology in the 1980's*, Ghent, 403–429.

METAPHONOLOGY OF ENGLISH PARONOMASIC PUNS

W. Sobkowiak

Institute of English, Adam Mickiewicz University,
Poznań, Poland.

ABSTRACT

Phonostatistic differences between English paronomasic (heterophonic) puns on the one hand and malapropisms and running text on the other are shown to be due to speakers' metaphonological control over the former. It is hypothesized that this control results from the action of metalinguistic subcomponent of functional competence, which, together with structural competence, forms human language faculty.

1. INTRODUCTION

Speech play: puns, 'secret languages', tongue-twisters, rhyming, impersonations, etc. are usually regarded as but providers of external evidence in phonology, exclusively used to assess plausibility of theoretical claims concerning rules and representations. They are hardly ever linguistically studied in their own right, as exponents of what has been referred to as pragmatic or functional competence. The reasons for this neglect have been variously stated in the pertinent literature: speech play is volatile, variable, literary, deliberate, artificial, nonreferential, hence extralinguistic. The common view of science as necessarily dealing with *serious* subjects has not been irrelevant in "excluding scholarship from this realm where lightness is all" ([4]:5). My aim in this paper is to show that speech play - puns in particular - can no more be treated as a 'mere performance phenomenon' than e.g. code and style switching, simplified registers, 'baby talk', and dozens of other phenomena routinely studied by socio- and psycholinguistics.

2. THE MODEL

For the purposes of this presentation I adopt the following model of human language faculty. Linguistic perfomance is normally driven by two types of competence. One is structural (grammatical) competence à la Chomsky, which provides the necessary substratum of representations and rules on various levels of language structure: phonological, morphological, syntactic, semantic. The other, by and large ignored in the standard generative tradition, is *functional* or *pragmatic* competence, which is responsible for how the knowledge of language structure is actually put to use in a communicative setting. Halliday and Hymes were the first to attempt a coordination of the two - so far disparate - views of language competence in the early 1970's.

Functional competence itself is far from being a compositional monolith. One of the most influential views of the many language functions has been that of Jakobson [6]. Jakobson relates functional modes of language to the components of a communicative situation: expressive function is focused on the speaker, impressive - on the listener, fatic - on the channel, etc. In the context of

this paper, it is the metalingual (henceforth: metalinguistic) function which is of most interest. Functioning metalinguistically speakers/listeners concentrate on the language itself, deliberately inspecting and manipulating it 'from the outside'. This I call *metalinguistic competence*. This is not only involved in scholarly discussions of grammar or philosophy, as most authors would have us believe. It lies at the very foundation of the human ability to play with language, and – in particular – to indulge in punning. More specifically, it is the *metaphonological competence* which is predominantly implicated in paronomasic (heterophonic) punning, which is the subject of this paper.

My hypothesis is the following: if punning (and other types of speech play) crucially involves metaphonological control over and above other types of functional indexing normally encountered in communication, this fact should have statistical ramifications in some phonological aspects of performance so controlled. Thus, if puns are phonologically different from 'ordinary' texts or speech errors – both of which are presumably *not* controlled metalinguistically – the argument that there is a dedicated metalinguistic subcomponent of functional competence would be corroborated. The view of performance as essentially a statistical reflection of competence goes back to Cedergren & Sankoff's [3] approach.

As this presentation is part of a larger project [7], it will be possible to present only some of the relevant results.

3. DATA AND RESULTS

Paronomasic puns (e.g. *Freud <---- afraid, sanctuary <---- thank you very*) appear in an amazing variety of playful genres: conundrums, knock-knocks, fake book titles, alphabet games, 'daffynitions', fractured French, graffiti. They are put to commercial use in advertisements and to jocular use in conversation. They are ubiquitous. I have collected – from about ninety printed sources – a corpus of 3850 items (types) like those at the top of this paragraph, transcribed them phonemically (American accent, fast/casual speech, stress ignored) and entered them in a computer database for further processing. To allow calculation of segmental identity, puns (intrusions) and sources were segment-wise aligned, Vitz & Winkler [8] style, e.g.

intrusion: /sæŋkču.eri/.
source : /Θæŋkjuveri/

This corpus was phonostatistically compared with Fay & Cutler's [5] collection of malapropisms (the type of speech error showing closest structural affinity to puns) and with Carterette & Jones's [2] data on phoneme frequencies in running English speech. The results of this comparison are as follows.

3.1. Overall similarity

Intrusions appear to be significantly more alike their sources, in terms of segmental identity, in puns than in malapropisms, as seen in the following table, which is arranged by proportion of nonidentities to the segmental length of source/ intrusion, the VITZ index:

TABLE 1. Overall segmental identity of sources and intrusions

VITZ	PUNS– %	MALAPROPS %
mean=	33.17	50.13
<=10%	1.5	0.0
<=20%	25.6	0.6
<=30%	21.4	14.8
<=40%	27.2	23.5
<=50%	14.4	30.1
<=60%	4.2	7.1
<=70%	3.1	9.3
<=80%	2.0	8.2
<=90%	0.3	1.1
<=100	0.3	5.5
N=	3850	183

The difference between frequency distributions of puns and malapropisms, relative to VITZ, is significant by χ^2 test. There is a mismatch of one phoneme in three in puns, as opposed to one in two in malapropisms, on the average. And this despite the fact that the mean length of malapropisms exceeds that of puns by 1.5 segment (7.1 vs. 5.6), which could favour low-VITZ figures by boosting the denominator of the proportion.

Punning intrusions are also more alike their sources in terms of featural similarity. Using an ad hoc system of 14 distinctive features (SYLL, CONS, SONO, CONT, VOIC, LABI, APIC, CORO, HIGH, MID, LOW, BACK, ROUN, GLID) I calculated frequency distributions of puns and malapropisms relative to DF difference. The results are presented in TABLE 2. Thus, for example, the proportion of cases where the two corresponding phonemes of the source and of the intrusion differ by only one DF equals 36.1% in puns, and 24.1% in malapropisms.

TABLE 2. Overall featural similarity of sources and intrusions

DF DIFF.	PUNS %	MALAPROPS %
1	36.1	24.1
2	34.4	30.7
3	17.4	27.0
4	8.4	12.3
5	3.3	3.4
6	0.4	2.4
C vs. C	44.7	61.7
V vs. V	55.3	38.3
N=	4620	381

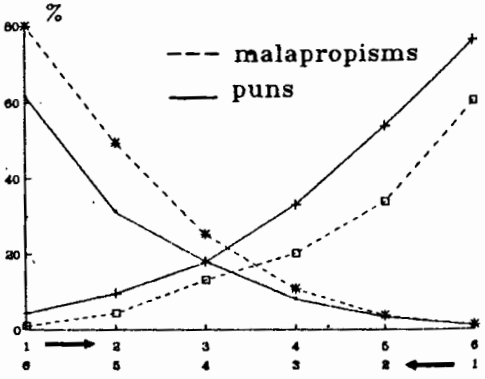
Similarly, at the first position of segmental nonidentity, counting from the left, puns exhibit more featural similarity than malapropisms. Failing to append pertinent tabulation for reasons of brevity, let me add that in about 50% of puns the first diverging segments differ by one or two features, with the figure for malapropisms being 32%.

Another interesting effect transpiring from TABLE 2 is the relative preponderance of vocalic oppositions in puns, as opposed to malapropisms. Relative to the latter value, which is close to that of a running English text (C=58.7%, V=41.3% according to [2]), the figures for puns are almost completely reversed: punsters apparently prefer to alter a vowel than a consonant in the source.

3.2. Lefward and Rightward Segmental Identity

Another significant point of difference between puns and malapropisms is the amount of segmental identity counted from both ends sources/intrusions. As is seen in DIAGRAM 1, puns behave like a near-perfect mirror-image reflection of malapropisms in this respect. In puns, sources/intrusions are more alike at the end than at the beginning, which means that punsters are more apt to change word-onset segments, whereas victims of malapropisms tend to mess up the offsets, keeping the onsets constant (which is also significantly the case in tip-of-the-tongue states; cf.[1]).

DIAGRAM 1.



Segmental identity from extremities of sources/intrusions

3.3. Phoneme Frequencies

Finally, puns differ from malapropisms and running texts in terms of phoneme frequencies, relative both to static distributions in sources/intrusions and to distributions of phoneme ousting, i.e. those cases where a phoneme is changed in transition from source to intrusion.

Thus, in malapropisms the phoneme frequency distributions of sources and intrusions are not significantly different from each other (χ^2=15.5 at 33 df, about 1200 phonemes), while in puns they are (χ^2=391 at 44 df, about 20,000 phonemes). Similarly, there is a significant discrepancy between the phoneme frequency distributions of intrusions in puns on the one hand and both malapropisms and text data ([2], about 48,000 phonemes) on the other.

Similar results are obtained in calculating frequency distributions of phonemes which oust (substitute) or are ousted. Malapropisms do not differ significantly from Carterette & Jones's [2] data in this respect, but puns do: both ousted and ousting distributions differ from those of running text, and from each other. Further interesting effects in grouped data are that: (1)stops are more frequent as ousting sounds in puns, (2)sonorants in puns are more stable (less amenable to ousting) than they are in malapropisms, (3)as are consonants as a class.

3. CONCLUSIONS

If puns are phonostatistically different from both malapropisms and running text, we can reductively infer that the differences are due to the additional factor of metaphonological control exercised by a subcomponent of speakers' functional competence over the performance mechanisms. Such control is absent from 'ordinary' speech encoding in a communicative situation where referential functions of language predominate. Puns, and - by extension - speech play of all kinds, are thus shown to be respectable areas of language study, properly belonging to - *sensu largo* - linguistic competence.

4. REFERENCES

[1]Brown,R.W. & D.McNeill.(1966), "The 'tip-of-the-tongue' phenomenon", *JVLVB, 5*, 325-37.

[2]Carterette,E.C. & M.H.Jones. (1974), *"Informal speech"*, Berkeley: University of California Press.

[3]Cedergren,H.J. & D.Sankoff. (1974), "Variable rules: performance as statistical reflection of competence", *Language, 50.2*, 333-55.

[4]Culler,J. (1988), *"On puns: the foundation of letters"*, Oxford: Blackwell.

[5]Fay,D. & A.Cutler. (1977), "Malapropisms and the structure of the mental lexicon", *Linguistic Inquiry, 8.3*,505-20.

[6]Jakobson,R. (1960), "Linguistics and poetics", In T.A.Sebeok (ed.), *"Style in language"*, Cambridge, Mass.:The MIT Press, 350-77.

[7]Sobkowiak,W. (in print), *"Metaphonology of English paronomasic puns"*, Franfurt: Peter Lang Verlag.

[8]Vitz,P.C. & B.S.Winkler. (1973), "Predicting the judged 'similarity of sound' of English words", *JVLVB, 12*, 373-88.

L'ACCENT DE L'ARABE PARLE A CASABLANCA ET A TUNIS ETUDE PHONETIQUE ET PHONOLOGIQUE

R. Bouziri, H. Nejmi & M.Taki

E.H.E.S.S., Université Paris 3, Université Paris 8

ABSTRACT

A phonetical-phonological comparative study of accent in the dialects of CASABLANCA and TUNISIA reveals a certain homogeneity in the phonological and the perceptual level. It shows also a difference in the parametric level.

BREF RAPPEL SOCIOLINGUISTIQUE

Dans les deux villes CASABLANCA et TUNIS, plusieurs systèmes linguistiques sont présents: L'ARABE, LE FRANCAIS, LE BERBERE. Le système arabe présente une diversité de niveaux: l'arabe classique, l'arabe standard, l'arabe formel et l'arabe dialectal. Il ne s'agit pas ici d'une discussion sur le plan sociolinguistique de ces différents niveaux .Notre étude porte sur le parler de CASABLANCA et de TUNIS, langue maternelle des Casablancais et des Tunisois. Elle est constituée de deux parties :

- Une analyse phonétique qui traitera de l'accent du point de vue perceptif et instrumental.
- Une analyse phonologique qui permettra d'interpréter les données phonétiques .

1- ANALYSE PHONETIQUE

Le corpus est composé de mots du lexique et de syntagmes (nominaux et verbaux) répondant à des critères doublement organisés:

a. Selon des structures syllabiques des deux parlers.

b. Selon les catégories syllabiques que sont les bissyllabiques et les trisyllabiques .

L'analyse a été effectuée au laboratoire de phonétique de l'UFR de Linguistique de l'université Paris 7.

1.1 TEST DE PERCEPTION

Les auditeurs de chaque parler ont noté intuitivement la syllabe qu'ils percevaient comme accentuée.

RESULTATS DU TEST

Nous constatons que l'accent est perçu sur la pénultième dans la structure CV-CV dans les deux parlers, en revanche dans la structure CəC-CəC la place de l'accent diffère ,sur la pénultième dans le parler de Tunis et sur la dernière dans le parler de Casa. Pour la structure CVC-CVC c'est la pénultième qui est accentuée dans les deux parlers. Quant aux autres structures c'est toujours la syllabe dite lourde qui porte d'une façon générale l'accent. Si l'hypothèse du poids syllabique s'avère valable en général, cela posera un problème majeur pour la structure CV-CəC car dans les deux parlers l'accent est sur la syllabe légère CV alors que l'on s'attend à ce qu'il soit sur la dernière CəC. Ce résultat nous mène à poser la question sur la nature et le statut de la syllabe CəC : peut-on lui attribuer le même statut que la syllabe lourde ou bien lui en attribuer un autre ?

Nous pensons que la syllabe CəC ne doit être considérée comme lourde qu'en deuxième degré alors que les autres CVC le sont en premier degré. Pour les trisyllabiques CV-CV-CV l'accent est sur la pénultième.

Les résultats du test sont donnés dans les tableaux suivants:

s/s	CV	CV	s/s	CəC	CV
Casa	61,5	38,4	Casa	62,2	37,5
Tunis	73,5	26,4	Tunis	75	25

s/s	CV	CəC	s/s	CVC	CVC
Casa	78,5	21,4	Casa	42,2	57,7
Tunis	80	20	Tunis	43,6	56,3

s/s	CəC	CəC	s/s	CVC	CV
Casa	32	67	Casa	65,3	34,6
Tunis	78,7	21,2	Tunis	65,2	34,7

s/s	CV	CVC	s/s	CV	CV	CV
Casa	27,6	72,3	C.	39,6	49,7	10,7
Tunis	26,4	73,5	T.	35,4	50	14,5

1.2. ANALYSE INSTRUMENTALE

Le corpus a été analysé essentiellement sur l'analyseur de mélodie de P. Martin qui nous a permis de relever les valeurs de Fo , I (sommets) et de durée. L'analyse a été effectuée principalement sur la syllabe ouverte CV et la syllabe fermée CVC en position accentuée et inaccentuée. Le contexte où sont étudiées ces syllabes est neutre excluant toute influence de l'intonation ou des processus de focalisation et d'emphatisation.

RESULTATS

Le résultat indique en général que la syllabe accentuée possède une Fo plus élevée, une intensité forte et une durée plus longue par rapport à la syllabe inaccentuée. Dans les deux parlers la fréquence fondamentale est le paramètre le plus important dans la mise en relief, cette prééminence n'a pas d'autonomie car elle se combine toujours avec l'un des deux autres paramètres, avec le paramètre I chez tous les sujets dans le parler de Tunis et seulement chez les sujets féminins dans le parler de Casa. En revanche chez les sujets masculins la Fo se combine avec la durée. Quant au caractère spécifique des deux parlers on remarque que pour la syllabe CV le parler de Tunis possède une Fo moyenne plus élevée et qui est de quatre quarts de ton chez les sujets masculins et de deux chez les sujets féminins. Quant au paramètre durée il semble que les tunisois allongent plus que les casablancais. En ce qui concerne le paramètre I on remarque peu de différence. Pour la syllabe CVC on constate qu'il y a une différence au niveau de Fo de quatre quarts de ton entre les sujets masculins des deux parlers alors que chez les sujets féminins il n'y a pas de différence . Pour le paramètre Durée on enregistre les mêmes résultats que la syllabe cv c'est-à-dire l'allongement de la syllabe accentuée des sujets tunisois. Le paramètre I ne présente que peu de différence.

CV parler de Casablanca

	Homme		Femme	
	acc.	n/acc.	acc.	n/acc.
Fo Hz	125	114	247	255
I db	39	37	40	38
D. cs	15	10	17	11

CV parler de Tunis

	Homme		Femme	
	acc.	n/acc.	acc.	n/acc.
Fo Hz	156	142	246	184
I db	38	36	39	38
D. cs	22	14	25	16

CVC parler de Casablanca

	Homme		femme	
	acc.	n/acc.	acc.	n/acc.
Fo Hz	126	106	241	216
I db	39	35	40	36
D. cs	8	5	9	7

CVC parler de Tunis

	Homme		Femme	
	acc.	n/acc.	acc.	n/acc.
Fo Hz	158	132	263	224
I db	38	34	39	35
D. cs	11	7	14	11

2. ANALYSE PHONOLOGIQUE

Le traitement phonologique que nous proposons pour le processus d'accentuation entre dans le cadre de la phonologie métrique élaborée par Prince (1983) et B. Laks (1988). En effet ce cadre est fondé sur une structure rythmique reflétée par une grille métrique représentant "le seul instrument formel d'assignation des proéminences" (B.Laks. 1988, 141). Dans ce modèle les structures accentuelles se manifestent directement au niveau de la grille; celle-ci est construite sur la base d'éléments accentuables qui sont de simples positions métriques pures organisées temporellement. Ainsi l'organisation de la grille ne fait-elle référence ni au contenu phonétique, ni à la substance phonologique de ces éléments. Tous les éléments susceptibles de porter un accent sont designés par un astérisque au niveau 0 de la grille. Le plan syllabique détermine les distinctions qualitatives que la grille doit prendre en considération dans l'organisation des séquences. Les deux mécanismes formels qui sont à la base de la construction de la grille sont :

a. le principe de la grille parfaite (GP), défini par le paramètre de la directionnalité (de droite à gauche ou de gauche à droite) et par la nature de l'élément de son point de départ (fort ou faible).

b. le principe de l'augmentation des extrémités:

RE (la règle d'extrémité) augmente un temps fort d'une extrémité, cette dernière est paramétrique. Ces deux principes sont contraints par le principe d'évitement d'antagonisme (EA). En outre une position accentuable peut être considérée comme extramétrique soit au début du mot (extra,I) soit à la fin (extra, F), (cf.B.Laks. 1988 , 164 - 5).

Partant des résultats de l'étude phonétique des deux parlers nous proposons l'organisation suivante:

les syllabes qui comportent une voyelle périphérique sont marquées par un astérisque au niveau σ de la grille .

Les voyelles qui se trouvent à la fin du domaine métrique sont considérées comme extramétriques.

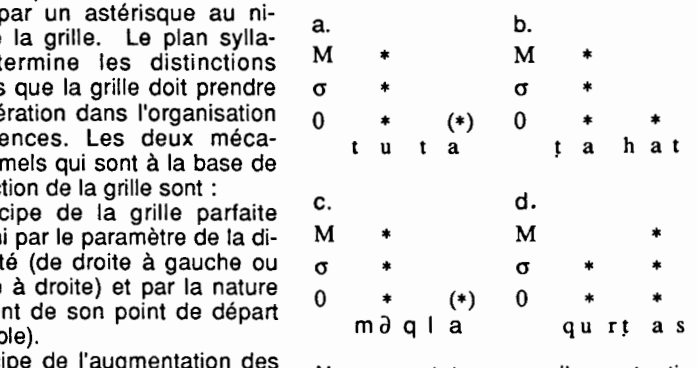
La règle d'augmentation s'applique au niveau M qui est l'étage supérieur par rapport à celui de σ .

L'unique distinction que nous soulignons entre les deux parlers concerne la construction de GP; bien que les deux parlers soient caractérisés par la même directionnalité ils se distinguent néanmoins quant à la nature du temps initial. Nous pouvons ainsi résumer ces paramètres de la manière suivante :

Parler de Casablanca : extra (O, F), RE (M,F), GP (,GD,f).

Parler de Tunis : extra(O,F), RE (M,F), GP (,GD,F)

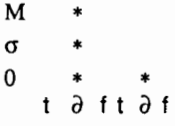
Les séquences bissyllabiques:



Nous constatons que l'accentuation de ces séquences est similaire pour les deux parlers; quant à la forme CəC-CəC qui se caractérise par une accentuation qui diffère dans les deux parlers, nous lui réservons le traitement suivant :

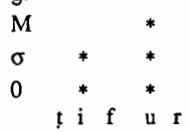
-Parler de Casablanca
e.1

Suivant le paramètre qui définit GP comme allant de gauche à droite avec un temps initial faible, nous obtenons au niveau σ un temps fort sur la deuxième syllabe qui sera augmenté au niveau M .
e.2



Pour le parler de Tunis GP est défini ayant comme temps initial un temps fort d'où la mise en relief de la première syllabe dans cette séquence .
f.

g.



Les structures trisyllabiques:
L'étude phonétique nous a montré que c'est la pénultième qui est accentuée dans les deux parlers. Le mécanisme proposé est le suivant :
h.

REFERENCES BIBLIOGRAPHIQUES:

ANGOUJARD, J-P., 1984. *Aspects d'une micro-prosodie: le modèle arabe.*
Thèse d'état, Université Paris 8.

BOHAS,G & al, 1989: Accentuation et effacement dans le parler de Tanger. *Langues orientales anciennes, philologie et linguistique,2*

KOULOUGHLI, D.E., 1978. *Contribution à la phonologie générative de l'arabe: le système verbal du SRA (Nord Constantinois, Algérie)* Thèse de troisième cycle, Paris 7.

LAKS, B., 1988. "Des Grilles et des arbres", *Recherches Linguistiques,17* Université Paris 8.

PRINCE, A., 1983. "Relating to the Grid", *Linguistic Inquiry, Vol. 14,.1.*

TAKI, M., 1988. *"L'alternance vocalique et la structure syllabique de l'arabe marocain"*, ms. inédit.

TAKI, M., 1989. *"Les verbes défectueux en arabe marocain: essai d'un traitement phonologique tridimensionnel"*, ms. inédit.

TAKI, M., 1990. *Syllabation, Association et Variation: Approche phonologique tridimensionnelle de l'arabe.* Thèse de Doctorat de l'Ecole des Hautes Etudes en Sciences Sociales, Paris.

INTERRELATION OF PERCEPTION AND PRODUCTION IN INITIAL LEARNING OF SECOND-LANGUAGE LEXICAL TONE

Jonathan Leather

University of Amsterdam

ABSTRACT

One group of native Dutch-speaking subjects were trained to perceive the lexical tones of standard Chinese (Putonghua) and were then tested on their ability to produce them, while a second group were trained (using a visual display, and without auditory exemplification) to produce the tones, and subsequently tested on their ability to perceive them. From longitudinal records kept of subjects' perceptual decisions and of acoustic parameters of their productions, the interrelation of learners' evolving perceptual and productive abilities was examined - and found for both groups to be significantly correlated.

1. INTRODUCTION

This paper reports on experiments designed to explore the interrelation between perceptual and productive abilities in the initial learning of a new sound (sub-) system. The phonetic proficiency goal was the citation-form tone system of Putonghua - i.e. standard Chinese or "Mandarin". These tones are phonetically realized as time-varying patterns of voice fundamental frequency: level, rising, dipping and falling contours respectively (see [1] for a review).

2. METHOD

The subjects were two groups of native Dutch speakers aged between 19 and 28, and with no knowledge of any tone language. One group underwent computer-managed tone perception training involving: (i) presentation of (digitized) tokens of the four tonally-distinguished words /ȳ/, /ý/, /y̌/ and /ỳ/ ('mud', 'fish', 'rain' and 'jade') produced by one Beijing native; (ii) presentation of tone tokens of first one female, then two male and two female speakers, for tone labelling. All labelling responses (collected through keyboard input) were logged under program control. Information feedback was provided in L1 in the form "Yes, correct." or "No, it was X" (X being the intended word). Trials continued until a proficiency criterion was satisfied of at least 80% correct identification of randomly-ordered tone tokens from all four speakers, at which time production ability was tested. In the production test the subject was asked to say the Putonghua word (again one of the four minimal quadruplets) corresponding to the L1 gloss displayed on the screen. The tones thus elicited were in a fixed random order (24 trials in all). In both this production test, and the production training of the second group of learners described below, larynx period data were recorded with a Laryngograph and phonetically assessed using a speaker normalization procedure and assessment algorithm described elsewhere [2, 1].

To enable the second group of learners to acquire tone production ability without prior experience in tone perception, it was necessary to use alternative means of exemplifying for them the F0 contour shapes of the tones, and to provide them with external information feedback on the phonetic consequences of their production attempts. The training system designed for this purpose collected larynx period data and provided on the screen visual and verbal feedback on the learner's F0 contours (the signal processing and task control routines used are detailed in [1]). A real-time plot of smoothed, speaker-normalized learner F0 was displayed beneath, or (if the learner preferred) superimposed upon, a similar plot of the exemplar F0 contour. The learner could thus make a visual appraisal of the match between his/her own tone production attempt and the model. Each production was phonetically assessed, and a record kept -again under program control - of all assessment parameters: the pitch level at 20%, 50% and 85% of the F0 contour, and its duration. Screen messages were provided for each parameter that in any trial was assessed as unsatisfactory. In the first stage of training, learners were presented on the screen with the F0 contours of the same single-speaker exemplars as were heard by the perceptually-trained learners. Subsequent stages of the training required subjects to produce the tone words in response to L1 glosses presented (on the screen) embedded in a question frame meaning "What is the word for ...?" No target F0 contour was displayed in this stage of the training until a learner's production attempt for a word elicited was unsatisfactory. Subjects continued this training until, if they satisfied the proficiency criterion by producing two consecutive "good" tokens for each of twenty randomly-ordered tone type elicitations, their ability to perceive the tones was tested. The tone stimuli in this perception test were the same digitized natural speech tokens of four Beijing natives (2 male and 2 female) used with the perceptually-trained group. Subjects were first presented with tone tokens of one (female) speaker, and then heard all four speakers; throughout, the tones were varied in a fixed random order.

3. RESULTS

3.1 Perceptually-trained group

Of the 17 learners who attained proficiency in tone perception, 9 were able without any productive training to produce tones with acceptable F0 contours in a fully contrastive system. Since these subjects' only experience of the tones was auditory, their targets for tone production must have been derived from representations developed for tone perception. Moreover, a consistency in subjects' productions suggests that these representations were fairly stable -as might be expected of the hypothetical tone prototypes they had established by the end of the training. Detailed analysis of the production assessment records reveals a fairly high degree of acoustic-phonetic invariance: a particular error type was recorded significantly more often in all, or only one of, the five test productions per tone than in two, three or four of them. In other words, those productions not assessed as "good" tended to be characterized, for individual subjects and tones, by certain recurrent patterns of deviance.

To investigate the correlation of learners' perceptual abilities with their production performances, both were quantified. A general measure of perceptual ability, M1, over the duration of the perceptual training was provided by multiplying the number of trials required to reach criterion by the mean number of misidentifications per tone type. The total number of detail errors in the production test was taken as a measure of production ability, M2. Over all 17 subjects who satisfied the proficiency criterion there is a significant positive correlation of M1 with M2 ($r = .65$, $p < .01$), indicating that subjects' productive skill was generally commensurate with their ability in perceptual learning. In greater detail: by subject, and by tone, the correlation of (i) the proportion of tokens misidentified in the final stage of the perceptual training with (ii) the proportion of production trials characterized by errors other than of duration is moderate ($r = .42$) but again significant ($p < .01$). To a noteworthy extent, then, learners produced more accurately the tones which they could more accurately label.

3.2 Productively-trained group

Some subjects were able after only production training to make correct identifications of all or some of the tone tokens in the perception test. To investigate the correlation of these subjects' productive abilities and perceptual performances, the number of production trials resulting in an 'unsatisfactory' assessment was taken as an inverse measure of productive learning, and the number of types of identification error recorded in the perception test as a measure of perceptual ability. The correlation of these measures is noteworthy among the 15 subjects who satisfied the production proficiency criterion: $r = .70$ ($p < .05$). Among these subjects there is a similar correlation between the number of types of error observed in production attempts on the one hand, and in perceptual decisions on the other ($r = .69$, $p < .05$). These positive correlations between measures of tone-productive ability and tone-perceptual accuracy would suggest that F0 patterns learned for the direction of tone production were referred to for the perceptual categorization of tone tokens heard from other speakers.

In a comparison of the performances of the perceptually-trained and productively-trained groups, it appears that the perceptually-trained learners enjoyed some overall advantage. By the sign test, they made fewer errors in tone perception yet no more in tone production (in each case $p < .05$). This is perhaps not surprising, considering the comparatively unnatural learning conditions of the productively-trained group. Secondly, for each of the respective tones the performances of the two groups prove to be significantly correlated in both the perceptual and productive modalities (for perceptual confusions Pearson's $r = .72$, $p < .01$, and for production errors, $r = .76$, $p < .001$).

4. DISCUSSION

The present learners did not, it appears, need to be trained in production to be able to produce, or in perception to be able to perceive, the F0 patterns of the target phonetic system: training in one modality tended to be sufficient to enable a learner to perform in the other. A learner who, after perceptual training, was able - in some cases from the first attempt - to produce the tones correctly, must have correctly inferred the requisite acoustic targets, and drawn upon his knowledge of the articulatory-acoustic characteristics of his own speech to attain them in production. A learner who, after production training, was able to identify correctly the tokens of other speakers, presumably exercised the ability to map the acoustic output of others into the phonetic space of his own. Theoretically, this could have been accomplished by fitting time- and range-normalized candidate F0 contours to learned contour prototypes -by means, for instance, of fuzzy logical pattern-matching (see e.g. [3, 4]). While the outcomes of an experimental study should not be too freely generalized to natural learning situations, the present findings would support a model of L2 speech pattern learning in which the learner's primary goal is the construction of phonetic prototypes to which the operations of both perception and production may be geared. These prototypes capture the central acoustic tendencies of 'good' tokens, and serve both perceptual decisions (cf. [5, 6]) and production activity (cf. [7, 8]). However, since there may be no simple correspondence between learners' perceptual and productive values for acoustic-phonetic parameters (e.g. [9, 10]), these prototypes, it may be postulated, are operationalized by means of schemata (i.e. structured plans) which define the serial and hierarchical orderings of cognitive and motor activities, providing the algorithmic bases for psycho-acoustic decisions (in perception) and feedback-based adjustments (in production). An association of independent perceptual and productive schemata with each phonetic prototype will account for any divergences

that may be observed between a learner's perception and production of a phone (see [9, 10, 11]). With operational schemata mediating between it and the phonetic events in relation to which it is defined, the prototype would theoretically satisfy one of the classic requirements of a phonological unit ([12]): that it be neutral with respect to the activities of production and perception.

REFERENCES

[1] LEATHER J. H. (1988) Speech pattern elements in second-language acquisition. Doctoral dissertation, University of London.

[2] LEATHER J.H. (1987) Automatic recognition of Chinese word tone from F0, with and without amplitude and speaker information. Proceedings of the European Conference on Speech Technology (Edinburgh) 1, 327-330.

[3] MASSARO D.W, COHEN M.M. and TSENG C-Y. (1985) The evaluation and integration of pitch height and contour in lexical tone in Mandarin Chinese. Journal of Chinese Linguistics 13, 267-289.

[4] TSENG C-Y, MASSARO D.W. and COHEN M.M. (1986) Lexical tone perception in Mandarin Chinese: evaluation and integration of acoustic features. In H.S.R. Kao and R. Hoosain (eds), Linguistics, Psychology and the Chinese Language. Centre of Asian Studies, University of Hong Kong, 91-104.

[5] KLATT D. (1979) Speech perception: a model of acoustic-phonetic variability and lexical access. Journal of Phonetics 7, 279 -312.

[6] BLUMSTEIN S. E. and STEVENS K. N. (1979) Acoustic invariance in speech production: evidence from measurements of the spectral characteristics of stop consonants. JASA 66, 1001-1017.

[7] LINDBLOM B. E. F., LUBKER J. and GAY T. (1979) Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation. Journal of Phonetics 7, 147-161.

[8] LADEFOGED P., DECLERK J., LINDAU M. and PAPCUN G. (1972) An auditory-motor theory of speech production. Working Papers Phonetics (UCLA) 22, 48-75.

[9] SHELDON A. (1985) The relationship between production and perception of the /r/-/l/ contrast in Korean adults learning English: A reply to Borden, Gerber, and Milsark. Language Learning 35, 107-113.

[10] BOHN O. and FLEGE J.E. (1990) Perception and production of a new vowel category by adult second-language learners. In J. Leather & A. James (eds) Proceedings of NEW SOUNDS 90, University of Amsterdam.

[11] SHELDON A. and STRANGE W. (1982) The acquisition of /r/ and /l/ by Japanese learners of English: evidence that speech production can precede speech perception. Applied Psycholinguistics 3, 243-261.

[12] LINELL P. (1982) The concept of phonological form and the activities of speech production and perception. Journal of Phonetics 10, 37-72.

FUNDAMENTAL FREQUENCY RANGE AND THE DEVELOPMENT OF INTONATION IN A GROUP OF PROFOUNDLY DEAF CHILDREN

Evelyn Abberton, Adrian Fourcin and Valerie Hazan

Department of Phonetics & Linguistics,
University College London, England

ABSTRACT

Results are presented from a 4-year developmental study of intonation in an unselected group of deaf children educated in an oral environment.

1.INTRODUCTION

Although intonational competence is still developing at 10 years of age in normal children [4], basic patterns are appropriately used by about 3. The speech perceptual and productive abilities of profoundly deaf children are delayed compared with their normally hearing peers [2], but some profoundly deaf children develop good conversational ability and highly acceptable and intelligible speech [3]. Clearly, pragmatic, syntactic, phonetic and phonological features are all important, and, in the area of pronunciation, both segmental and prosodic control have to be achieved.

2. SUBJECTS, TESTS & MEASURES

In this paper we illustrate certain findings relating to intonation development from a 4-year study of aspects of the speech perceptual and productive abilities of a group of 16 severely-profoundly deaf children (from age 7 - 8) educated orally. Results have already been presented for some of the children's intonation after 2 years [1]. Here, we now discuss the fundamental frequency and intonation development of the 2 children with the least impaired hearing and the 2 with the most impaired hearing (according to pure tone audiometry) after 4 years. All four are congenitally deaf. Their pure tone average losses in the better ear at 0.5, 1.0 and 2KHz are

Child 1 - 83dB HL; Child 2 - 90dB HL; Child 15 - 112dB HL; Child 3 - 115dB HL.

Their speech has been recorded at regular intervals and analysed using laryngographic and acoustic techniques. Fundamental frequency measures obtained are compared with qualitative auditory analyses: larynx frequency histograms (Dx plots) derived from recordings made during story telling sessions can be related to perceived pitch range, and modal values related to judgments of perceived high or low voices. Scattergrams (Cx plots) of all pairs of successive vocal fold vibrations in the recording correlate with perceived vocal roughness or smoothness. Figs 1-4 show Dx and Cx plots for Children 1 & 16 in March 1985 and May 1989.

3. RESULTS - QUANTITATIVE

Table 1 shows the frequency ranges and main frequency modes in the children's speech. Second order histograms are used to eliminate creaky voice contribution from the measurements, and the range estimates are from 90% of the digram occurrences. This is our standard procedure in UCL work with deaf adults as well as with children.

Table 1. Fx ranges and modal values

Child	Mode (Hz) of 2nd order range			Range in octaves of 2nd order Dx		
	'85	'87	'89	'85	'87	'89
1.	265	272	225	0.52	0.47	0.44
2.	321	312	157	0.55	0.47	0.53
15.	296	348	312	0.47	0.63	0.36
16.	511	484	304	1.44	0.91	0.51

Note: The 1987 distributions for Children 15 & 16 are multimodal.

4. RESULTS - QUALITATIVE

Apart from Child 15, the other children all

show a decrease in modal frequency as one would expect with age, and Child 2's voice has broken in a normal manner. Except for Child 2, the other three still have perceptually high voices.

There is no simple relationship between Fx range measures and perceived narrowness of pitch range and monotonous voice [1]. Apart from Child 16 whose 1985 Fx values represent his physiological rather than his speech pitch range, these 4 deaf children have largely normal octave widths compared with Hunt's normally hearing group [6], but they differ markedly in their control of voice pitch contours to organise their speech in terms of an intonation system.

Over the four years of study (during which time they had no speech-language therapy) the four children show different patterns of intonation development. In 1985 Child 1 was already using pitch to organise her speech into word groups, and using major pitch changes for focus, although most of these nuclear tones were falling and often over-long. Perceptually she had a narrow pitch range. In 1989 nuclear lengthening has disappeared and she is using a full range of nuclear tones in syntactically and attitudinally appropriate ways. Her pitch range no longer seems narrow although the octave width is, in fact, smaller than 4 years earlier. Child 2, similarly, in 1985 showed the demarcative and focussing functions of pitch control but with nuclear lengthening and almost exclusive use of rise-fall tones. (His segmental phonology was much less mature than Child 1's.) He has also made progress by 1989, despite still showing some nuclear lengthening. Most of his tones are falls, but contrast appropriately with rise falls, and some rises are correctly used. With less variety of nuclear tones, his pitch range still sounds narrow. Children 15 and 16 are much more delayed but have made some progress: in 1985 Child 15 had a narrow pitch range but was using pitch for demarcation and focussing. In 1989 his syntax is still very immature but nuclear placement is usually correct. Most tones are falls but rises are beginning to appear. Over the years Child 16, despite his profound hearing loss, has successfully reduced his wide physiological Fx range to one which is speech-like, and has learned to use his vocal output in the give-and-take of conversation; his speech is not very intelligible but is organised in terms of pitch control: he has clear tones, often utterance-final, and nearly always falling. All the children have improved voice quality in terms of regularity of vocal fold vibration.

5. CONCLUSION

There is a dearth of normative data on the pattern of intonation development but these four hearing-impaired children, using conventional amplifying hearing aids from an early age show that even profoundly deaf children can acquire facility with linguistic pitch control in several ways. Nevertheless, their progress is slow and delayed, and it remains to be seen whether gains in phonation quality and in pitch control and use could be obtained at the right moment in development through systematic visual feedback therapy and speech-processing hearing aids that focus attention on the low frequency elements of speech [5].

6. REFERENCES

[1] ABBERTON,E., FOURCIN,A. & HAZAN,V. (1988) "Analytic assessment of speech development in deaf children: production." *Proc. SPEECH '88, 7th FASE Symposium* (Edinburgh: Institute of Acoustics) 1077-1084.

[2] ABBERTON,E., FOURCIN,A & HAZAN,V. (1990) "The development of contrastiveness in profoundly deaf children's speech." *Clin. Ling. & Phon.* 4,209-220.

[3] CLARK,M. (1989) *"Language through living for hearing-impaired children."* London: Hodder & Stoughton.

[4] CRUTTENDEN,A. (1974) "An experiment involving comprehension of intonation in children from 7 to 10." *J.Ch.L.* 1:221-231.

[5] FOURCIN,A. (1990) "Prospects for speech pattern element aids." *Acta Otol. Suppl.* 469:257-267.

[6] HUNT,L. (1988) "An introduction to normative voice measurement in children using the laryngograph." Unpublished BSc Speech Sciences project, University College London.

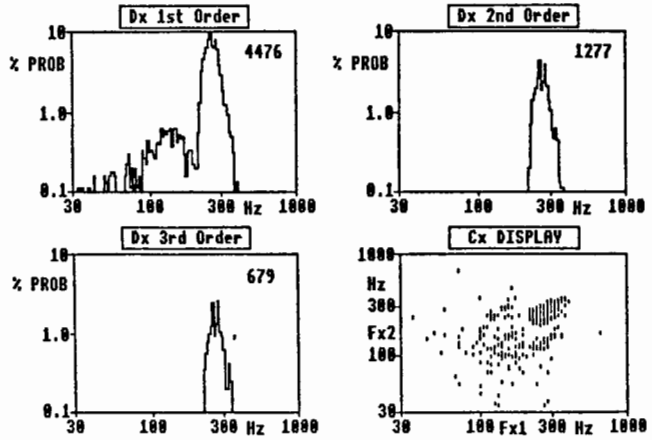


Figure 1 Fundamental Frequency histograms (Dx) and Scattergrams (Cx) for Child 1 in March '85

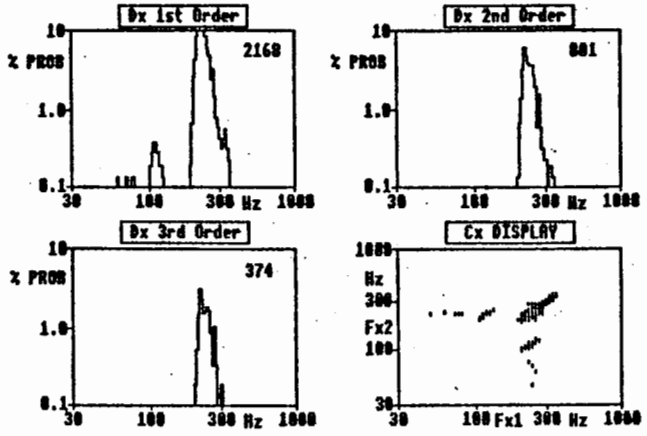


Figure 2 Fundamental Frequency histograms (Dx) and Scattergram (Cx) for Child 1 in May '89 {analyses based on period by period measurements}

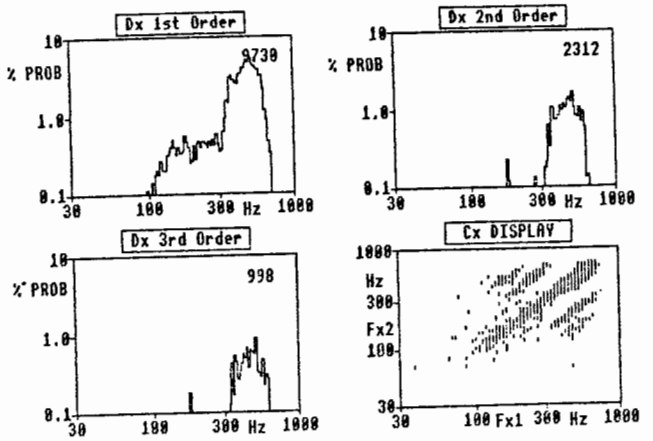


Figure 3 Fundamental Frequency histograms (Dx) and Scattergrams (Cx) for Child 16 in March '85

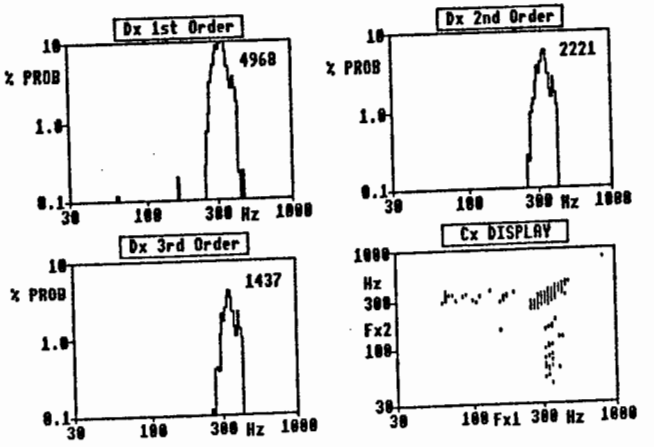


Figure 4 Fundamental Frequency histograms (Dx) and Scattergram (Cx) for Child 16 in May '89 {analyses based on period by period measurements}

THE ROLE OF LANGUAGE FORMULATION IN DEVELOPMENTAL DISFLUENCY

Frank Wijnen

University of Utrecht, The Netherlands

ABSTRACT

The disfluency patterns of two 2-year-olds are compared. In both children, disfluency shows an increase and a subsequent decline. In one of the children, disfluency is mild, in the other it is excessive. The excessively disfluent child shows many word part repetitions, and relatively few sentence incompletions. Moreover, most of his self repairs are phonologically motivated. In the other child, word- and word-string repetitions as well as sentence incompletions are more frequent, and a relatively large number of self repairs involve syntactic alterations. It is concluded that the disfluencies are related to phonological encoding in the excessively disfluent child, and to sentence planning in the mildly disfluent child.

1. INTRODUCTION

In most children speech fluency deteriorates temporarily between ages 2 and 3 [9], although there is considerable inter-individual variation in the extent of the problem. In some cases the child becomes a stutterer. Several studies have pointed at a connection between developmental disfluency and language development [4]. It is often argued that fluency decreases as a result of the increasing grammatical complexity of utterances, which poses progressive demands on the child's language production ability. In [6] I developed a specific version of this hypothesis: The *Development of the Formulator Hypothesis* (DFH).

The DFH starts from the observation that language development around age 2.5 is characterized by the transition from *telegraphic speech*, which lacks almost all function words and morphosyntactic elements, to a morphosyntactically more mature level of language competence. The acquisition of closed-class elements and morpho-syntax necessitates the development of a component of the speech production mechanism that is dedicated to morpho-syntactic processing and serial order planning, which can be identified with the *positional planner* in Garrett's speech production model [1]. Due to the positional planner's initial lack of automaticity and imperfect co-ordination with other components in the speech production mechanism, speech planning will start to break down more often, which produces an increase of disfluency. Usually, however, speech fluency will be restored as the new system gets settled. More importantly, the DFH predicts that as the rate of disfluency rises, its distribution over sentence positions will change. Disfluencies will start to concentrate at loci in speech that coincide with moments at which the positional planner is highly active, viz. the onset of clauses and major constituents. This prediction was confirmed in a longitudinal case study. The subject in this study showed a disfluency peak at age 2;8. Before this age, disfluencies were distributed randomly over sentence positions. As of 2;8, however, they occurred predominantly at sentence onsets and phrase-initial function words.

The preliminary results of a second longitudinal case study, however, showed a different pattern [7]. Again, a significant and quite dramatic increase of disfluency was observed, followed by a decrease. However, the disfluencies were concentrated at sentence onsets from the beginning of the observation period onwards. Moreover, the subject appeared to be more advanced linguistically than would have been expected on the basis of

the DFH. These results agree with the general observation that there are considerable inter-individual differences in the rate of language development. Furthermore they suggest that the developmental process underlying the disfluency episode in the second child cannot be the one described by the DFH. It is unclear as yet what other process may be responsible for the increase of disfluency in this child.

The primary aim of this paper is to contribute to the solution of this problem. To achieve this aim, I will present some new data with respect to differences between the patterns of disfluency in the two children mentioned before. I will try to corroborate the conclusion of [7], viz. that the disfluencies in the two subjects have different sources. Particularly, I will argue that disfluency in the second subject mentioned is primarily related to another component of language formulation, viz. *phonological encoding*, i.e., the unpacking of word form information from the mental lexicon [3]. In order to do so, I will make two assumptions. First, I will assume that a disfluency always results from a disturbance in the planning of an utterance segment that is yet to be uttered. The second assumption, which is based on Levelt's work on self-repairs [3], states that speakers avoid interrupting a word, unless it is a source of trouble itself. The corollary of these assumptions is that different types of disfluency may signal utterance planning problems at different levels. In particular, the repetition of an initial word fragment will predominantly signal problems in preparing the remaining parts of the word for articulation [8]. A word repetition, by contrast, would point at a planning difficulty with regard to some aspect of the subsequent sentence fragment.

2. METHOD

2.1. Subjects

The data in this study are derived from longitudinal language corpora of two Dutch boys, T and H. Both children were observed between ages 2;4 (years; months) and 3;0. Language development was assessed with the aid of *TARSP*, a Dutch adaptation of Crystal's *Language Acquisition, Remediation and Screening Procedure* [5]. TARSP divides the course of grammatical development into 7 phases. At age 2;4, T appeared to be a relatively backward Phase 3 child, whereas H was at an advanced Phase 4 level. This implies that T could produce sentences containing up to 3 constituents; he was not yet able to expand constituents into word groups and he had very limited morphology. H, on the other hand, could produce 4-constituent sentences, expand constituents into word groups and use some verbal and nominal inflections productively. To advance from Phase 3 to Phase 4, the average child needs about 5 to 6 months.

2.2. Recording and Transcription

The children's speech was recorded at home while interacting with their mothers. Roughly one hour of conversation was recorded per week. Apart from the literal content of the children's utterances, their phonetic structure was transcribed in places where this would clarify the interpretation of speech. The types of disfluency that were transcribed are *repetitions* (of word parts, words and word strings), *revisions*, *incomplete sentences*, *blocks* and *prolongations*, *word breaks*, and *senseless sound insertions*.

3. RESULTS

3.1. Disfluency Rates

At 2;4, T produces an average of 2.7 repetitions per 100 words. H is slightly more disfluent with an average of 4.2 repetitions per 100 words. In both children the repetition rate increases in the subsequent months. When disfluency is at its peak, at age 2;8, T is still very mildly disfluent, with an average of 4.5 repetitions per 100 words. H's repetition rate reaches a maximum of 29.5 repetitions per 100 words at age 2;7, which amply exceeds the normal limits. It may not come as a surprise that H's mother consulted a speech therapist, who nevertheless advised not to interfere. In both children disfluency rapidly declined. At age 3;0, T had 2.7 repetitions, and H 5.9 repetitions per 100 words. Both children are now normally fluent speakers.

3.2. Disfluency Types

In the remainder of the Results section, two segments of the observation period will be singled out, viz. the first month, around age 2;4, and a period of roughly one month around the time when disfluency was at its peak, corresponding to

age 2;8 in T and 2;7 in H.

Table I shows a breakdown of the *repetitions* according to the size of the utterance fragment involved. Collapsed over both periods, T appears to have much more word and word-string repetitions than H. In H, on the other hand, the word-part repetitions are predominant. Ignoring the category of indeterminate repetitions, this difference reaches significance ($\chi^2 = 27.29$, df = 2, $p < .001$).

TABLE I. Distribution of repetition types in T and H. WST = Word string repetitions; WRD = word repetitions; W-P = word part repetitions; IND = indeterminate. Percentages in parentheses.

Corpus	*WST*	*WRD*	*W-P*	*IND*	*Total*
H 2;4	3	22	18	-	43
	(7)	(51.2)	(41.9)	(-)	(100)
H 2;7	4	48	165	16	233
	(1.7)	(20.6)	(70.8)	(6.9)	(100)
H Tot	7	70	183	16	276
	(2.5)	(25.4)	(66.3)	(5.8)	(100)
T 2;4	3	17	38	1	59
	(5.1)	(28.8)	(64.4)	(1.7)	(100)
T 2;8	14	61	46	-	121
	(11.6)	(50.4)	(38)	-	(100)
T Tot	17	78	84	1	180
	(9.4)	(43.3)	(46.7)	(0.6)	(100)

Note however that the developmental pattern differs between the two children. T shows a transition from a predominance of word-part repetitions to a predominance of word and word-string repetitions. By contrast, an opposite development can be witnessed in H. H's pattern accords with the 'classical' observation that repeated elements are progressively truncated in the developmental course of stuttering [4].

Under the assumptions made above, this finding suggests a difference in character of the planning difficulties underlying the observed discontinuities. In particular, it may be expected that H experiences more problems in constructing the phonological shape of words than T. A first, indirect piece of supportive evidence for this conjecture can be derived from a quantitative analysis of *sentence incompletions*. These disturbances can be considered to result from a failure at the level of sentence planning. If H's disfluencies are reflective of phonological encoding processes at word level, whereas T's discontinuities reflect sentence planning, one would expect less sentence incompletions in H than in T. This is precisely what Table II indicates. Collapsed over both periods, the ratio of incomplete to complete sentences is significantly lower in T than in H ($\chi^2 = 9$, df = 1, $p < .005$). An inspection of the figures in Table II suggests that this difference is primarily determined by the figures relating to the late periods.

TABLE II. Sentence incompletions (SI) and fully interpretable, non-interrupted sentences (NS)(≥ 1 word; *yes*'s and *no*'s excluded). Percentages in parentheses.

Corpus	*SI*	*NS*	*Total*
H 2;4	8	457	465
	(1.7)	(98.3)	(100)
H 2;7	16	295	311
	(5.1)	(94.9)	(100)
H Tot	24	752	776
	(3.1)	(96.9)	(100)
T 2;4	24	1284	1308
	(1.8)	(98.2)	(100)
T 2;8	127	1162	1289
	(9.9)	(90.1)	(100)
T Tot	151	2446	2597
	(5.8)	(94.2)	(100)

3.3. Self-Repairs

A final piece of evidence can be derived from an analysis of the kinds of *speech repairs* that are made by the subjects. Speech repairs involve the interruption of an ongoing utterance, some delay, and a retracing that encompasses an alteration of the original utterance [3]. According to the nature of the alteration the repairs were classified as *phonological*, *lexical*, or *syntactic*. It seems reasonable to expect that if a particular type of planning problem is predominant, the number of errors related to this problem that penetrates into overt speech, where they may be monitored and repaired, should also be relatively large. Table III shows the number of different types of repairs in T and H. It is clear that the distribution of repair types differs between subjects ($\chi^2 = 13.85$, df = 2, $p < .001$, disregarding the 'other' category). The difference is concentrated in the categories of phonological and syntactic repairs. Proportionally, H has approximately twice as many phonological repairs as T, whereas T has almost 9 times as many syntactic repairs as H. This outcome supports the previous conjecture that the sources of planning trouble underlying disfluency differ between T and

H.

TABLE III. Self repairs involving phonological (PHO), lexical (LEX), syntactic (SYN), and other (OTH) alterations. Percentages in parentheses.

Corpus	*PHO*	*LEX*	*SYN*	*OTH*	*Total*
H 2;4	12	4	-	-	16
	(75)	(25)	-	-	(100)
H 2;7	13	3	1	-	17
	(76.5)	(17.6)	(5.9)	-	(100)
H Tot	25	7	1	-	33
	(75.8)	(21.2)	(3)	-	(100)
T 2;4	12	5	4	1	22
	(54.5)	(22.7)	(18.2)	(4.5)	(100)
T 2;8	16	16	17	4	53
	(30.2)	(30.2)	(32.1)	(7.5)	(100)
T Tot	28	21	21	5	75
	(37.3)	(28)	(28)	(6.7)	(100)

4. DISCUSSION

The results presented here support the interpretations in [6] and [7]. There appears to be a difference between T and H regarding the origin of speech disfluency. T shows a prevalence of word and word string repetitions, a relatively large amount of sentence incompletions and a relatively high number of syntactic self-repairs. H, by contrast, shows mainly word-part repetitions; he has relatively few sentence incompletions and his repairs mainly involve phonological alterations. Consequently, T's disfluency seems to be mainly related to planning operations at sentence level, whereas H's disfluencies appear to be associated with the programming of word forms.

Of course it is sensible to entertain some reserve with respect to this interpretation, in view of the fact that it is to some extent based on assumptions which are, although they appear quite plausible, in need of external validation. This will be an issue in further research.

According to one of these assumptions, disfluency results from a breakdown of planning processes. Alternatively, it has been suggested that disfluencies reflect *covert* repair operations, i.e. self-repairs which precede articulation, by virtue of the speaker's ability to monitor so-called 'internal speech' [2, 3]. This hypothesis suggests an even closer relation between repetitions and (overt) self-repairs than is proposed here. The confrontation of these opposing views should also be a topic in further research.

It seems fair to conclude that the DFH is too narrow an explanation of developmental disfluency. Apart from sentence planning, phonological encoding may also be associated with childhood fluency problems, which accords with certain views on adult stuttering [8]. It remains to be clarified, however, what developmental process affecting phonological encoding is responsible for the *reduction* of fluency.

5. REFERENCES

[1] GARRETT, M.F. (1982), "Production of speech: observations from normal and pathological language use", in A.W. Ellis (ed.), *Normality and pathology in cognitive functions*. London: Academic Press.

[2] KOLK, H. (1991), "Is stuttering a symptom of adaptation or of impairment?", in H.F.M. Peters et al. (eds.), *Proceedings of the 2nd conference on Speech Motor Control and Stuttering*, to appear.

[3] LEVELT, W.J.M. (1989), *"Speaking: from intention to articulation"*, Cambridge (MA): MIT Press.

[4] STARKWEATHER, C.W. (1987), *"Fluency and stuttering"*, Englewood Cliffs: Prentice-Hall.

[5] VERHULST-SCHLICHTING, L. (1987), *"TARSP: Taalontwikkelingsschaal voor Nederlandstalige kinderen van 1-4 jaar"*, Lisse: Swets & Zeitlinger.

[6] WIJNEN, F. (1990), "The development of sentence planning", *Journal of Child Language, 17*, 651-675.

[7] WIJNEN, F. (1991), "The role of sentence formulation in developmental stuttering", in H.F.M. Peters et al. (eds.) *Proceedings of the 2nd conference on Speech Motor Control and Stuttering*, to appear.

[8] WINGATE, M. (1989) *"The structure of stuttering"*, New York: Springer.

[9] YAIRI, E. (1982), "Longitudinal studies of disfluencies in two-year-old children", *Journal of Speech and Hearing Research, 24*, 490-495.

THE ACQUISITION OF VOICING CONTRAST IN NORMAL AND AT-RISK INFANTS

U. Bortolini*, C. Zmarich, S. Bonifacio**.**

*** Centro di Fonetica del CNR, Padova (Italy)**
**** Div. ORL, Istituto per l'Infanzia "Burlo", Trieste (Italy)**

ABSTRACT

This paper reports an acoustical investigation of the development of the voicing contrast in Italian word initial stops produced by three groups of infants: premature, low-birth and controls. The purposes of the study were to compare the patterns of acquisition of the acoustic-phonetic cues for voicing in the speech of at-risk infants and controls and to discuss the inter-group differences in relation to phonological proficiency. The cues investigated were VOT values for stops in initial word position. The productions of the subjects were recorded at the ages of 18,21,24,27 months. The results are discussed in terms of similarities and differences among the three subject groups, of the rate of changes in the acoustic-phonetic cues across ages, and in terms of the differences existing at each age level.

1. INTRODUCTION

This research is a part of a larger investigation concerning possible effects of different handicapping conditions present at birth on language learning. It is well known from the literature that newborn children at-risk have inferior maturation level as compared with controls, and that the alterations of language learning and phonology proficiency could be due to these causes. Speech timing and its variability as source of information about speech motor control development in children have been object of much recent interest. Developmental studies indicate that it takes children several years to establish the motor-control skills needed to realize phones in mature fashion.

However, the relationship between development of accuracy in the control of acoustic-phonetic cues and phonological development has not been investigated extensively. The purposes of this paper were: to compare the patterns of acquisition of the acoustic-phonetic cues for voicing in the speech of at-risk infants and controls; to discuss intersubject different results in relation to phonological proficiency. The measure studied was initial stop-consonant Voice Onset Time (VOT), which is known to be the most reliable acoustic cue separating voiced-voiceless stops. Apparent differences in the ages at which similar phonemic voicing distinctions are made across languages may actually be the result of the different phonetic categories employed in those languages.

The establishment of mature VOT does not arise from a single principle: both speech motor control capabilities and auditory factors are important in this regard.

The discrepancy in VOT acquisition between at-risk subjects and controls can be seen as an adaptation to perceptual constraints as well as to production factors.

2. PROCEDURE

The total population of this study consisted of 4 infants born at less than 37 weeks gestation and 4 full-term weighing less than 2500 grams. A control group of 4 children born full term at normal weight and 4 adult aged from 24 to 26 years also participated. The test was administered to small (S) and normal infants (N) at 18,21,24,27

recorded under standard recording conditions (using Uher model 4200 portable tape recorder with Electrovoice model 635A microphone) saying each of 12 test words at least three times. The test items were the following minimal pair pseudo-words, contrasting labial, dental and velar voiced and voiceless stops: *'papa, 'baba, 'pipi, 'bibi, 'tata, 'dada, 'titi, 'didi, 'kaka, 'gaga, 'kiki, 'gigi.* The infant and adult productions were collected in randomized order. VOT values of each initial stop were measured using Sygnalize 1.2 (1988-90 by Eric Keller). All VOT values reported here are from tokens produced with oral and velopharingeal complete closure.
Means and standard deviations were computed for each syllable for each subject group.

RESULTS

A pair-wise comparison of the VOT means in subject groups for age level is given below. As shown in Fig. 1 the adults' VOTs for voiceless and voiced stops are greater before high vowel /*i*/ than before low vowel /*a*/. Taken together the results shown in Fig. 2.3.4. indicate that: 1. the distinction between voiced and voiceless stops in word initial position emerges relatively late as measured by differences in mean VOT within each age group: 27 month-old normal children have acquired all voicing contrasts except /*k*/ vs. /*g*/, however as measured by differences in mean VOT they still do not produce /*d*/ or /*g*/ in an adult-like manner. 2. VOT values differed significantly between Normal and Preterm infants but not between Normal and Small infants. 3. the children's standard deviations are generally greater than the adults, and inter-Preterm subject's variability was greater than controls. We interpreted these findings as further evidence that long lead voicing is more difficult to produce than zero or short lag voicing. The ability to control relative timing of the articulatory and laryngeal events necessary for the production of voiced italian stops (long lead) appears to be physiologically more complex than initiating both events simultaneously and thereby producing phonemically voiceless stops. Thus, the frequent substitution of /*p*/, /*t*/, /*k*/ for /*b*/, /*d*/, /*g*/, in young children's meaningful speech are most likely a result of speech motor-control factors than of linguistically based rules.

REFERENCES

[1] ALLEN, G. D. (1985), "How the young French child avoids the pre-voicing problem for word-initial voiced stops", *Journal of child language*, 12, 37-46.

[2] DAVIS, K. (1990), "The acquisition of VOT: is it language-dependent?", *Papers and Reports on Child language Development*, 29, 28-37.

[3] GANDOUR, J., HOLASUIT PETTY, S. (1986), "The acquisition of the voicing contrast in Thai: a study of voice onset time in word-initial stop consonants", *Journal of Child Language*, 13, 561-72.

[4] JENSEN, S.T., BOGGILD-ANDERSEN, B., SCHMIDT J., ANKERHUS J., HANSEN E. (19899, "Perinatal risk factors and first-year vocalizations: influence on preschool language and motor performance", *Developmental Medicine and Child Neurology*, 30, 153-161.

[5] KEWLEY-PORT, D., PRESTON, M.S. (1974), "Early apical stop production: A voice onset time analysis", *Journal of Phonetics*, 2, 195-210.

[6] LISKER, L., ABRAMSON, A.S. (1964), "A cross-language study of Voicing in Initial Stops: Acoustical Measurements", *Word*, 20, 384-422.

[7] MACKEN, M.A., BARTON D. (1980), "The acquisition of the voicing contrast in English; a study of voice onset time in word-initial stop consonants", *Journal of Child Language*, 7, 41-74.

[8] ROTHENBERG, M. (1968), "The breath-stream dynamics of simple-released-plosive production", *Biblioteca Phonetica*, 6, 68-109.

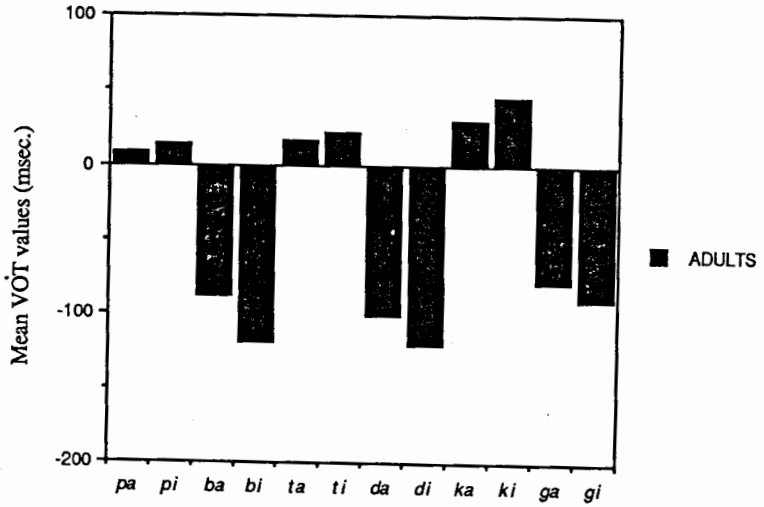


Fig.1. Mean VOT values for Adults

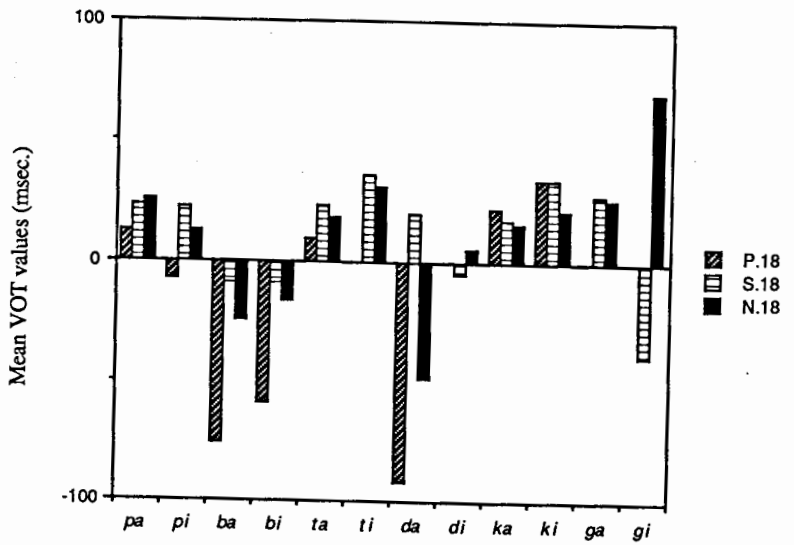


Fig.2. Mean VOT values at 18 months

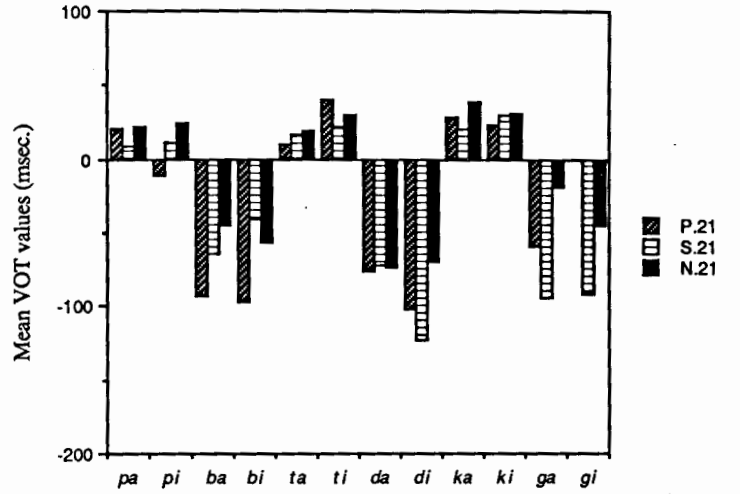


Fig.3. Mean VOT values at 21 months

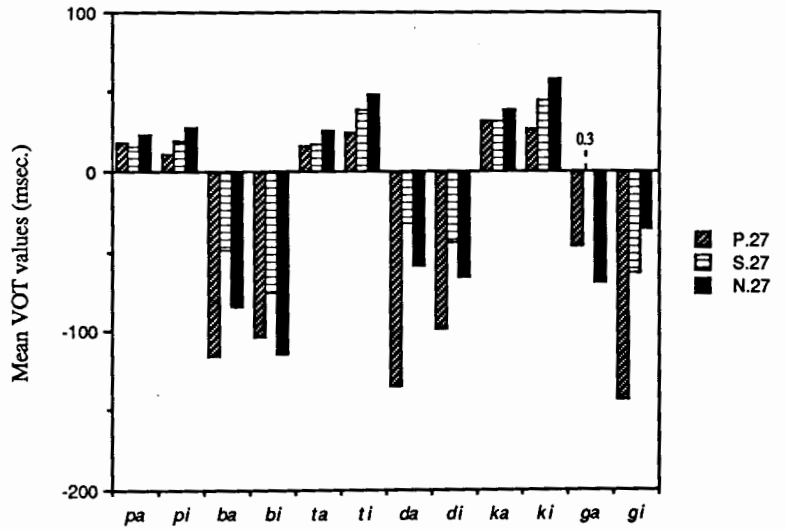


Fig.4. Mean VOT values at 27 months

VOWEL ACQUISITION IN FRENCH AND ITALIAN

P. Bonaventura

Department of Linguistics, University of Texas at Austin.

ABSTRACT

Disyllabic babbling and speech of 3 Italian and 3 French children has been transcribed and analyzed in order to evaluate claims for both basic sound-making propensities in babbling (including coarticulatory constraints [12]), and trends towards target language properties.

The consistencies within the groups and the differences between groups showed the influence of the target languages on the set of vocalic productions prior to the acquisition of the first words.

The CVCV forms were more consistent with the predictions regarding coarticulation [12] in Italian than in French, perhaps because of different relations between target language patterns and basic infant propensities in the two languages.

Vowels produced in disyllabic utterances by 3 Italian and 3 French monolingual children, in babbling and speech, have been analyzed.

The data have been obtained by monthly recordings of all productions of the children in their home environment; phonetic transcription of disyllabic utterances have been used for a distributional analysis.

The consistencies of individual patterns within groups and the differences between groups showed the influence of the target languages on the set of vocalic productions prior to the acquisition of the first words.

A general tendency in CVCV patterns was for cooccurrence of Front vowels with Palatal consonants, of Central vowels with Labials and of Back vowels with Velars in Italian babies, that seem to fit with MacNeilage and Davis predictions about early coarticulatory patterns. The French children, though, depart from this schema in Back vowels which appear mostly after Palatal consonants.

Also, Italian children productions are very dissimilar from target-language coarticulatory preferences whereas French show a closer fit and maybe evidence for an earlier shift toward language preferred complex articulatory abilities.

1. INTRODUCTION: GOALS OF THE STUDY

The goal of this paper is to describe the development of use of the vowel space in French and Italian babies from the babbling stage through use of the first words. The two groups have been investigated in order to see if they reflect target-language influences in their babbling.

An additional question beyond the sheer frequency of occurrence of vowels in babbling and speech, concerns consonant-vowel relationships: patterns of cooccurrences of vowels and consonants in disyllabic utterances have been analyzed in order to test an aspect of MacNeilage and Davis [10] "frame/content" theory of speech production: they predicted systematic coarticulatory constraints between the C-V segments within syllabic 'frames' of early babbling (see below).

This work was sponsored by a scholarship from the Fyssen Foundation, Paris, 1989-90.

1.1. Background

Recent studies on babbling and speech ([2], [5], [10]) have confirmed prior studies of babbling (e.g. [8]) showing that there exists a favorite set of vowels in the front central-middle /low part of the articulatory space, that are the first to be produced by infants: these patterns may be universal in babbling. In addition, every language has at least one vowel in this region.

In addition to some consensus on the preference of lower left quadrant vowel-like sounds, there is some general tendency to favor different areas of expansion on the vowel space (i.e. toward high-front area for English and back for Cantonese: See [5]); such trends have also been tested in perceptual experiments [4].

MacNeilage and Davis theory of speech production [10] addresses the problem of coarticulation in acquisition of speech : an elementary unit of speech production is postulated, a pure "syllabic frame", observable in single or repeated episodes of mandible oscillation; when these episodes are accompanied by vocalization, the basic sequence of Labial consonant + Central vowel is produced.

Elementary movements of the tongue can cooccur with the frame provided by mandible oscillation: pre-fronting or "consistently-held" tongue fronting would result in a sequence of Alveolar consonant + Front vowel and a pre-backing, or "consistently held" tongue backing, would result in a Velar + Back vowel sequence. The C and V segments, at this stage ("nonvariegated babbling"), would not be independent, but produced with the "frame", as a whole unit.

In variegated babbling, local modifications due to tongue positioning on the front-back/low-high axes, can appear: in these forms, real differentiation between segments begins to emerge and the segments start to be produced independently within the frame.

This view of early babbling differs basically from the classical hypothesis on coarticulation in adult speech, that assumes "a) discrete and invariant units serving as input to the system of speech production " and b) eventual obscurations of the boundaries between units at the articulatory or acoustic levels"[7].

2. METHOD

2.1. Subjects

The subjects studied were 3 French monolingual children, Camille, Louis and Myrtille, and 3 Italian monolingual children, Luca, Francesco and Evelina. The age range was 0;9 to 1;5.

An average of 4 sessions has been considered for each child.

2.2. Data collection

The French material has been kindly provided by the Experimental Psychology Lab.,C.N.R.S., Paris.

Two of the Italian children have been recorded in Rome, by a procedure similar to the one used for the French children in Paris: the sessions took place at home, every 15-20 days, in the presence of at least one parent and one or two experimenters; Luca's recordings have been kindly provided by the Phonetics Lab., C.N.R., Padua.

2.3. Data analysis

IPA Transcriptions of the disyllabic utterances by the babies have been stored on MacIntosh computer by IPAPlus fonts, kindly made available by Prof. G. Boulakia, of the Institute of Phonetics of the Charles V University, Paris.

A distributional analysis has been performed on the database by the software "Quatrième Dimension": ad hoc formats and procedures were created by Mme C. Carcassonne of the Center of Mathematics applied to Humanities, C.N.R.S., Paris.

Two analysis have been performed, separately on babbling and speech:

I) Computation of total number of vowels by class (nine classes are considered :BackHigh, BackMid,BackLow,CentralHigh,CentralMid, CentralLow, FrontHigh, FrontMid, FrontLow).

II) Computation of child vowels in first syllable vs. second syllable, with respect to the consonant preceding every vowel (four consonant classes have been considered: Labials, Alveolars/Dentals, Palatals, Velars).

3. RESULTS

Results of the vowel frequency analysis show an overall preference for the MF, LC and MB vowels (Fig.1) by French and Italian babies, in both babbling and speech: LC appear to be more frequent, both in babbling and in speech; higher numbers of LC and MB, though, are found in Italian than in French.

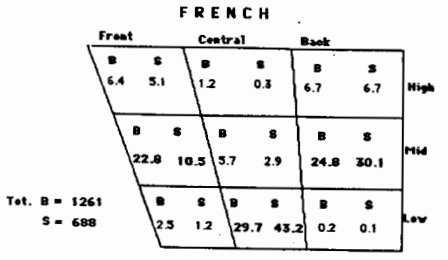
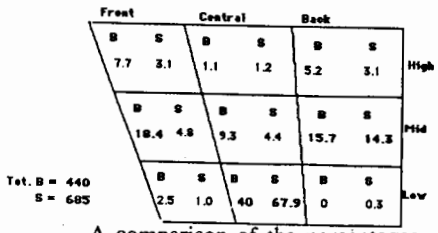


Fig. 1: Total distribution of vowels in babbling (B) and speech (S) of Italian and French children (expressed as percentages).



A comparison of the percentages shown above (Fig.1) with the frequency of occurrence of the phoneme classes in each language (from [6], [1]) shows that the LC presence in the data reflects the situation of the adult languages: [a] has a frequency of 31% in Italian and of 17% in French; actually, in Italian this vowel appears twice as often as in French.

MF vowels, the second preferred set, have 25% frequency altogether in Italian and 31% in French, although, according to my classification, the French MF space contains a higher concentration of phonemes than the Italian one (see. Fig.2).

Overall French and Italian patterns are very similar, although Italian babies have significantly less MF in speech with respect to the French ones.

The CVCV results (Table 1) show highest frequency of cooccurrence of Front vowels with Palatal consonants in Italian, whereas in French Front vowels tend to be articulated after Palatal and Dental consonants.

Central vowels cooccur consistently with Labials in Italian, but they are equally frequent with Labials and Velars in the French data.

Finally, Back vowels cooccur with Velar consonants in two Italian subjects and with Labials in Luca, whereas in French they show a different tendency to be coarticulated with Palatals.

A comparison with the frequencies of vowels in CV syllables from the most frequent 200 disyllables in Italian (from [3]) and from the most frequent 100 words in French ([9]), shows some correspondence between the French babies preferences for Front Vowels to occur with A/D consonants, and the frequency of this constraint in the language (20%); also, Central vowels in French show high frequencies with Labial and Velar consonants (19- 14%), as well as in the babies productions. In French, though, occurrence of Central vowels is also high after A/D consonants (19%). Finally, Back vowels appear most frequently in an A/D environment in French (13%), but they are preferred after Palatals in the data.

The Italian language frequency pattern favors A/D consonants in the environment of all classes of vowels (F:29%, C:15%; B:22%): this tendency is not reflected by the Italian children.

4. DISCUSSION

The differences that have emerged between the French and Italian patterns and the English patterns reported in MacNeilage and Davis [12] can be interpreted as follows:

1) The higher number of LC found in Italian with respect to French/English can be attributed to a target-language influence.

2) The drop in MF vowels from babbling to speech, stronger in Italian than in French, reflects different properties of the target-vowel spaces, as well: French children are drifting toward a space where four phonemes are concentrated in the MF area (see Fig.2), whereas the Italian space is more [a]-centered, and MF vowels are represented only by two phonemes ([e]-[ɛ]).

Overall French and Italian patterns differ from English in the following: a) MF are not present in high percentages in English babbling; accordingly, MF have a low frequency (11%) in the language.
b) The greater number of LF vowels reported by MacNeilage and Davis [11] reflects the high frequency of [ae] in English; the result could also be due to the classificatory system adopted in this study, where both French and Italian [a]'s are included in the LC category, even if the French articulation is intermediate between the English and the Italian one (see Fig. 2).

3) The Italian CVCV data reflect the scenario postulated by the 'frame/content' theory; French data, on the other hand, show an overall preference for Front and Back vowels to be produced in Palatal/Dental context, and for Central vowels to occur in Labial/Velar context.
The question therefore arises as to whether there exists a progressive shift towards coarticulatory patterns preferred in the target language, as has been shown for single vowels.

The comparison with the frequencies of vowels in the most frequent CV syllables in the language shows some evidence for a drift towards target coarticulatory patterns for Front and Central vowels in French children; this trend is absent in Italian children.

This effect might be due to a slower rate of transition from infant to adult articulatory patterns. It could be argued that the acquisition of coarticulatory constraints develops after the ability to produce independent segments is acquired: in this view, acquisition of speech production consists in separating segments from a holistic production 'frame' and consequently reassemble them as independent units in the speech chain.

REFERENCES

[1] ANTONETTI, P. and ROSSI, M. (1970), "*Précis de phonétique de l'italien*", in Publications des Annales de la Faculté de Lettres Aix-en-Provence. Série: Travaux et Mémoires, N. LVIII, Aix-en-Provence.

[2] BICKLEY, C. (1983), "*Acoustic evidence for phonological development of vowels in young children*", Utrecht , Tenth International Congress of Phonetic Sciences.

[3] BORTOLINI, U., Tagliavini, C. and ZAMPOLLI, M. (1971), "*Lessico di frequenza della lingua italiana contemporanea*", Milano.

[4] DeBOYSSON BARDIES, B. et al. (1984), "Discernible differences in the babbling of infants according to the target language", *Journal of Child Language*, 11, 1-15.

[5] DeBOYSSON BARDIES, B. et al. (1989), "A cross-linguistic investigations of vowel formants in babbling", *Journal of Child Language*, 16, 1-17.

[6] DELATTRE, P. (1965), "*Comparing the phonetic features of English, German, Spanish and French*", Heidelberg.

[7] KENT,R.D. and MINIFIE, F.D. (1977), "Coarticulation in recent speech productin models", *Journal of Phonetics*, 5, 115-133.

[8] KENT, R.D. and MURRAY, A. (1982), Acoustic features of infant vocalic utterances at 3, 6, and 9 months", *Journal of Acoustical Society of America*, 72, 353-365.

[9] JUILLAND, A. (1970), "*Frequency dictionary of French words*", The Hague, Mouton.

[10] MACNEILAGE, P.F. and DAVIS, B.L. (1989), "*Acquisition of speech: frames, then content*", in: M. Jeannerod (Ed.) Attention and Performance, XIII, Motor Representation and Control, Hillsdale, N.J., Lawrence Erlbaum.

[11] MACNEILAGE, P.F. and DAVIS, B.L. (1990), "*Acquisition of speech production: the achievement of segmental independence.*" In Hardcastle, W.J. and Marchal, A. (Eds.) Speech Production and Speech Modelling. Kluwer, Dordrecht (in press).

[12] MACNEILAGE, P.F. and DAVIS, B.L. (1991), "Vowel-Consonant relations in babbling", this volume.

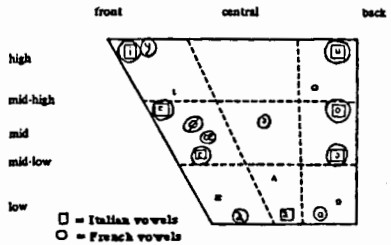


Fig. 2 : A vowel chart of the American, French and Italian vowels (Adapted from : P. Ladefoged: "A course in Phonetics")

Table I : Pattern of cooccurrence of Vowels and Consonants in CV syllables produced by French and Italian babies.

Italian				French			
Vowels	L	F	E	Vowels	M	L	C
C	Labial	Labial	Labial	C	Velar/ Labial	Labial	Velar
F	Palatal	Palatal	Palatal	F	Palatal	Dental	Labial/ Dental
B	Velar	Velar	Labial / Velar	B	Dental/ Palatal	Palatal	Palatal

Italian: L = Luca, F = Francesco, E = Evelina
French: L = Louis, M = Myrtille, C = Camille
Vowels = Front, Central, Back

CONTOURS MELODIQUES DANS LA REDUPLICATION SYLLABIQUE: ETAPES-CLES DANS L'ACQUISITION DE LA PAROLE.

M.M Vidal-Petit

Université René-Descartes de Paris V.

ABSTRACT

Our studies on logorrhea and dyslalia throw into light the systematic presence of an intonation pathology.Its origin is to be found probably in the babbling genesis and above all during the key-period of syllabic reduplication development. The syllabic reduplication acquisition can be viewed through 3 specific steps being marked by the development of specific melodic shapings. The longitudinal study of some infants from 7 to 28 months of age without disorders and of 2 infants having heavy psychiatric disorders have pointed out that the coming out of the first words seems to be conditioned by the control of the last step described in the syllabic reduplication acquisition.

1. LOGORRHEE ET DYSLALIE

Nos études sur des pôles fluants et non fluants de la parole chez l'enfant (logorrhée, dyslalie) ont mis en évidence la présence systématique d'une pathologie de l'intonation.

1.1.Dans le cadre de la logorrhée

Nous notons [1]:

- une courbe intonative présentant un caractère répétitif et musical. Si le caractère musical de cette courbe se révèle différent dans son expression, le caractère répétitif est marqué par l'aspect systématique du schéma mélodique, et cela, très souvent, quelque soit le type d'énoncés (interrogatifs, assertifs, etc.).
-la valeur musicale de la parole peut selon les enfants correspondre soit au principe de la mélodie en musique, soit appartenir plus au genre de la musique tonale où l'harmonie et la mélodie sont réglées par l'obligation de respecter un ton principal. Dans ce dernier cas, la notion de chant vient de la capacité de l'enfant à parler tout d'abord en maintenant une même fréquence ou un ton durant une longue durée, mais surtout de son aptitude à réaliser des pas mélodiques, c'est à dire de passer brutalement d'un ton à un autre sans transition à l'inverse de la parole usuelle.

1.2. Dans le cadre de la dyslalie. Nous notons [2]:

-une courbe mélodique pauvre marquée par une double absence de relief intonatif due à la faible variation de la fréquence fondamentale le long de l'énoncé et à la petite variation intrasyllabique du fondamental. L'impression de recto-tono donnée par la faible variation du Fo se trouve renforcée par le fréquent maintien d'une note où fréquence dans une syllabe durant une longue durée ou/et par la production de pas mélodiques.

1.3.Discussion

Nous relevons étonnament, dans les deux pôles fluants et non fluants, la présence de traits mélodiques communs qui sont le maintien d'une même fréquence ou note et les pas mélodiques.

Nous constatons entre ces deux pôles un degré quantitatif dans l'utilisation de ces traits spécifiques. Le pôle fluant diffère du pôle non fluant par un usage répétitif systématique de ces schémas mélodiques lesquels ne se rencontrent ni dans la parole de l'adulte, ni dans celle de l'enfant sans troubles âgés de plus de 12-15 mois.

Chez les enfants sans troubles plus jeunes ces schémas mélodiques sont retrouvés lors de la reduplication syllabique dans le babillage, mais ils constituent alors une étape normale.

2- LA REDUPLICATION SYLLABIQUE DANS LE BABILLAGE.

Une étude longitudinale d'enfants âgés de 7 à 28 mois nous permet d'envisager l'acquisition de la reduplication syllabique à travers la maîtrise de trois étapes clefs.

2-1. La première étape se singularise dès l'âge de 7 mois par une succession de syllabes redupliquées produites sur la même note que nous pouvons représenter selon le schéma suivant:

da da da da da da

2-2. La deuxième étape (8-9 mois) se manifeste par une rupture tonale au sein de la reduplication concrétisée par le passage d'une note à l'autre correspondant au principe des pas mélodiques que l'on peut schématiser ainsi:

da – da da – da da –
da da da

2-3. La troisième étape (9-12 mois) se distingue des deux premières par l'apparition de contours mélodiques descendants pouvant être visualisés ainsi :

ta ta ta ta ta ta

Le dernier contour d'une série de "ta ta" est souvent marqué par un allongement sur la dernière syllabe pouvant indiquer l'assertion, la résolution, la finalité etc...

La maîtrise de cette étape conditionne et annonce dans la parole des enfants l'apparition massive et rapide de contours complexes syllabiques ainsi que celle des premiers mots ou de leur ébauche au niveau du signifiant attendu en regard du signifié désigné.

Progressivement la production des schémas mélodiques de la première et de la deuxième étape va diminuer puis disparaitre définitivement au fur et à mesure de l'augmentation de la fréquence d'apparition des contours ascendants,descendants, complexes et des premiers mots.

La prise en compte et la confirmation de l'importance de cette troisième étape nous sont apparues à travers l'étude de 2 enfants présentant des troubles psychiatriques graves.

3- CAS DE 2 ENFANTS AUTISTES

3-1. Etude du premier cas

- L'apparition de la première étape a eu lieu vers 9 mois et s'est maintenue très longtemps.
- L'apparition de la deuxième étape se situe à 18 mois coïncidant parfaitement avec celle tardive de la marche à quatre pattes. Il est à noter que l'enfant se déplace sans regarder devant lui la tête tournée de côté. La mère signale sans que nous ayons pu le vérifier que l'enfant prononce [ma ma] pour maman, [pa pa] pour papa et [abwa] pour à boire.

- L'apparition de la troisième étape intervient seulement à 28 mois après l'acquisition de la marche. Un mois plus tard nous relevons fréquemment des contours variés, montants et descendants, de types exclamatifs, assertifs, interrogatifs, ainsi que l'apparition des premiers mots comme [gato], [bato], [mama].

3-2. Etude du deuxième cas

Nous n'avons pas pu mettre en évidence aussi nettement une corrélation entre les étapes motrices de cet enfant et le passage d'une étape de la reduplication à une autre. Cet enfant a été par ailleurs moins bien suivi en raison des contraintes données par la famille. Néanmoins la troisième étape intervenue à 30 mois va amener comme chez les enfants sans troubles et chez le premier enfant autiste, une explosion langagière.

4- CONCLUSION

Nos études sur des enfants sans troubles et sur 2 enfants autistes semblent bien montrer que l'apparition des premiers mots est conditionnée par la maîtrise de la troisième étape décrite dans l'acquisition de la reduplication syllabique.

5- REFERENCES

[1] VIDAL, M-M. et PRENERON, C. (1985), " *Prosodie et mode d'insertion dialogique chez 3 enfants logorrhéiques*", Etudes de Linguistique Appliquée, *56*, 96-113.

[2] VIDAL, M-M.(1985), " La dyslalie en question", Folia Phoniatrica, 37, 139-151.

HOW SLAVIC PHONETIC TYPOLOGY CHANGED THROUGH CONTACT

Herbert Galton

Institut für Slawistik der Universität Wien

The most important changes characterizing Slavic such as the many palatalizations and the vocalic opposition contrasting back and front vowels especially in endings are not independent developments, tho' some of them are perfectly compatible with the I.-E. character of the language, but are frankly due to an imitation of Altaic (more precisely, probably Proto-Turkic) speech habits where they can be explained by the agglutinative morphology. This latter was not imitated by the Slavs, though, as being entirely alien to an inflexional I.-E. idiom. Huns and Avars dominated the Slavs for the four critical centuries ca. 400–800 A.D.

In this talk we will proceed from the assumption, which I have tried to justify elsewhere, that there was indeed a Balto-Slavic language spoken in a dialect continuum in Eastern Europe, roughly from the shores of the Baltic to somewhere north of the Black Sea, still in the first half of the first millennium A.D.

I cannot, of course, supply a phonological system of Balto-Slavic any more than anybody else, but have to proceed from the one attributed to the parent tongue in [1]51, 31–64, with due allowance for some changes which may be assumed to have intervened in the formation of Balto-Slavic. [12], 61 and [10], 22, have characterized I.-E. by and large as a consonantal type of language, and an inspection of the inventories given both by Szemerényi and Gamkrelidze-Ivanov would seem to bear out such a judgment. Thy type of I.-E. from which Blt.-Sl developed had at least three rows of velars (gutturals), whether we adopt those postulated by the former or the latter authors. In either case we arrive at about a dozen velars ([16], 64, [5], 34) and six to eight alveolare and labials; furthermore, there are spirants, of which the latter authors have three, all of the hissing sibilant kind, the former one, plus two nasals, liquids and glides each, for a total of 25–29 consonants, as opposed to the five cardinal vowels, long or short, plus diphthongs (none in Slavic any more than Altaic). There are no hushing sibilants or affricates in either I.-E. system.

It is of particular interest to us that although palatal(ized) consonants may be attributed at least to a part of Proto-I.-E., their reflexes in Slavic certainly show no palatalization, thus prase 'pig' and zima 'winter' with Satəm-. I.E. *k', *g'h are not in any way to be considered palatalized in Proto-Slavic. No incentive seems, therefore, to have come for palatalization from the parent tongue, although many of its daughter idioms have undergone such a process, so that it cannot have been in conflict with its evolutionary tendencies. There is, of course, a physiologi-

cal foundation for such a tendency, and that is the broad adaptability of velars to the influence of ensuing as well as preceding vowels, because their acoustic locus is espevially apt to adjust itself to its environment(I 3 1,67, 114, 124,133).

To this day, Lith. has e.g. kimštas 'stuffed' against O.C.S. čęstъ, skýstas 'liquid' vs. čistъ, from the same root also Slavic čęsta 'clearing', or, with the voiced counterpart, gélti 'to prick' against Com.Sl. *žędlo 'sting', or gelà 'pain' against O.C.S. žalь. It is, therefore, in my opinion not enough to say that the velars simply were palatalized in Slavic, and that this feature distinguishes it from Baltic, but we have to find the cause of this difference, which cannot lie in the initial system. For the mere anticipation of the vowel articulation must originally have been the same in both branches, but eventually led to different phonemes in Slavic presumably many centuries before what palatalization there evolved in Baltic. There cannot have been many more "empty slots" in the South than in the North. The effect was both earlier and much stronger in Slavic, although it eventually also did percolate to the North, especially to Latvian.

However, Blt.-Sl. had already added to the stock of I.-E. spirants in the shape of a /š/, as a result of the change of /s/following the so-called r-u-k-i formula, consisting in an adjustmentof its upper formant to that of these sounds if they immedtiately preceded (I 10 1, 30). Also here the physiological/ acoustic conditioning can be explained, but does not suffice for a statement of causality. Since it is a very old change and fell still within the period of Blt.-Sl. neighbourhood with Indo-Iranian tribes, which likewise carried it out, Burrows (I 2 1,79)excludes a coincidence in the change s > š, which means that the link is causal; it is significant that the more northern Baltic has carried it out much less systematically, and not shared in the subsequent Slavic development š > x (ch) before back vowels, This also goes to show that the Slavs were much more sensitive to the division of vowels into back and front, and that the /š/ in Slavic became a palatal which was in some sort of harmony with its environment.

Again, the physiological mechanism is not far to seek. In his still unrivaled "Slavische Phonetik", Olaf Broch (I 1 1,59)states explicitly that the boundary line between /š/ and /x/ (ch) is fleeting, and that a small change in the positionof the articulating organs can bring it about. If an ever greater part of the tongue tip is bent down from the area of the teeth and alveolae, he says, involving a bigger concentration of the bulk of the tongue in the posterior area, the /š/ will be seen to gradually change into the velar fricative. The question remains for us to be tackled as to why the Slavs should have carried out such a shift, which their Baltic cousins did not. It was carried out before back vowels, which apparently did not combine very well with palatals at a certain stage in the genesis of Slavic. This change, then, shows a strong degree of adaptation between the consonant and the tautosyllabically following vowel.

Nor is this the only example of its kind. The three Slavic palatalizations, for which we will follow the traditional order, show the same sensibility, only to front vowels, in the case of the third even extending to preceding /i/, with due allowance for the labialized character of the following one which prevented it. We notice that the first is carried out in a true neogrammarian spirit all over the Slavic territory, while the second and the third seem to fade in the Far North of the Slavic world, as becomes more and more

clear from the birch-bark writs of the Novgorod and Pskov area[191,118].

Another sound which is susceptible to the effects of its vocalic environment is the /l/, which in Slavic split into two phonemes /l/ and /l'/, the first with a hard allophone [ɫ], the second palatal and due to a merger of l + j. It is important to realize that say lista 'leaf' differs by this consonant and not by the vowel from the second syllable of vol'i 'to the will'. I make a special point of this, because there is a teaching of an alleged Slavic synharmony of the syllable about, according to which entire syllables in Proto-Slavic were either hard or soft (labio-velarized vs. palatalized), so that their symbols can precede the notation of the whole syllable. How does this theory account for such facts? Are there different degrees of syllabic harmony, greater in /l'i/ than in /ɫi/ with its neutral phoneme? Besides, under the auspices of this theory we are always treated to theoretical examples like say ta - t'a, ny - ńi, which, if put together, would yield Japanese rather than Slavic words. Slavic remained true to the I.-E. type in that, however much it may have opened its syllables at the coda, it permitted very respectable sequences ("clusters") at their beginning, where no reduction occurred; do we have say in ra-zdruši-ti 'to destroy' a labio-velarized syllable °zdru- as against a palatalized 'zdra- in razdrašiti 'to solve'? The great Dutch slavicist N.van Wijk ([18]91,45) explicitly mentions cases like O.C.S. bregъ 'shore', where there is absolutely no reason to attribute even a mere phonetic palatalizatiom to the initial /b/, while fully admitting, of course, that the effect of front vowels also on directly preceding non-velars must have been stronger in Proto-Slavic than in the other I.-E. languages, without changing their phonemic status.

The palatal consonants arose from palatalizations, as in the case of velars, or sequences of alveolars plus /j/, and for this, I might add, there was no ready pattern in I.-E., all these moves were Slavic innovations which set off that idiom from Baltic and were essential in constituting Slavic as such. With the different results of some of these sequences we cannot concern ourselves here, suffice it to say that the results were at first all palatal in all subgroups, in cluding the /št/, /žd/ of O.C.S., and the question now remains as to where this strong effect of front vowels and /j/ has come from. About this, van Wijk says [18]1 that we do not knew whence such a strong effect of vowels on preceding consonants has come; Roman Jakobson [5]1 in a way answered this question by placing Slavic within a wider Eurasian setting, and in his turn, P. Ivić ([1]71,51) has taken up this suggestion, but would like to know when, where, and under what historical circumstances such an influence has taken place.

The answer to Ivić's very pertinent question can presumably be supplied by a reference to the historical circumstances under which the Slavs lived in the critical period, roughly from 400 to 800 A.D. A first answer has been supplied by Schelesniker [1]4, who believes, though, in the synharmonism of the Slavic syllable, but i.a. correctly appreciates the importance of the change *u > /y/ (for which there was no "case vide"), as well as of the progressive (Altaic) direction of the third palatalization. Again in the case of the former change, it cannot be sufficiently stressed that the term "delabialization" explains absolutely nothing, but is a mere label.

In the period in question, the Slavs were dominated by various Altaic tribes, foremost the Avars, whose empire came to an abrupt end shortly before 800, but before that by Huns (who also roped in the Slavs for military service, cf. [1]51, p.230) Bulgars and Khazars. This was not a matter of mere neighborhood or some -stratum, but certainly at least in

the case of the Avars an interpenetration affecting the Slavic anthropological type and - most importantly- involving the language of command under which these "sqalāb" (III1, 225, 238; 'slaves' later 'border guards') of the Avars were sent into battle for their masters all over the borders of their vast empire, which as a result brought about a largely unitary lingua franca - Slavic.

We can proceed from the assumption that all those peoples were Turkic (thus Abaev in [6],141), however, the picture would not be changed in its overall outlines if they had been Mongols in view of their languages' phonetic nearness at that time ([13], 91). Now the morphological structure of these languages is dominated by agglutination, one of whose consequences is that the vowels of the morphemes attached to the stem must share its back vs. front character. This results in vocalic oppositions a : ä, ɤ (Slavic value) : i, e : ö, u: ü, plus an /e/ about which there is some argument. The consonants of these morphemes underwent a strong assimilatory effect of the vowels, which comes to the fore in the eldest, Runic, alphabet of the Old Turkic inscriptions in the Orkhon and Yenisei valleys of the VIII.ct., where we find two letters each representing b -b', g - g', d - d', k - k', l - l', n - n', r - r', s - s', t -t' etc. There is a respectable array of sibilants and affricates s, z, š, ž, č, dž ([17], 78), which should in my opinion make it clear where the model which the Slavs sought to imitate came from.

However, I still maintain that the imitation was not absolute and was limited to the phonetic inventory plus phonotactic rules, but extended neither to the agglutinative nature of the Altaic languages nor affected Indo-European syllable structure in the initial part, nor introduced a synharmonism of the syllable. The correlation of palatalization constitutes a later development.

REFERENCES

[1] BROCH,O.(1911), "Slavische Phonetik", Heidelberg.

[2] BURROWS,T.(1963), "The Sanskrit Language", London.

[3] FANT G.(1973), "Speech Sounds and Features", Cambridge, Mass.

[4] GABAIN, A.von (1974), "Alttürkische Grammatik", Wiesbaden.

[5] GAMKRELIDZE T.V. - IVANOV V.V. (1984), "Indoevropejskij jazyk i indoevropejcy", Tbilisi.

[6] ISAEV M.I. - TENIŠEV E.R.(1990), "Ossetica-Turcica", Voprosy Jazykoznanija, No. 6, p.140-143.

[7] IVIČ,P.(1965), "Roman Jakobson and the Growth of Phonology", Linguistics 18, p. 35-78.

[8] JAKOBSON, R. (1971), "Remarques sur l'évolution du russe", Selected Writings I, The Hague.

[9] JANIN V.L. - ZALIZNJAK A.A.,(1986) "Novgorodskie gramoty na bereste", Moskva.

[10] LAMPRECHT, A.(1987), "Praslovanština, Brno.

[11] LEWICKI,T.(1956), "Źródła arabskie do dziejów słow." Wrocław.

[12], POHL, H.D. (1986), "Zur Typologie des Altbulg."Die slav.Sprachen 10, p. 61-70.

[13] POPPE,N. (1960), "Vergleichende Grammatik der altaischen Sprachen" Teil I: Lautlehre, Wiesbaden.

[14] SCHELESNIKER,H.(1975),"Turanische Einflüsse im urslav.Sprachsystem" Wiener Slav.Jahrbuch 21,237-41.

[15] SŁOWNIK STAROŻYTNOŚCI SŁOWIAN, (1964), article on Huns by Żak.

[16] SZEMERÉNYI, O. (1980), "Einführung in die vergleichende Sprachwissenschaft", 2d ed., Darmstadt.

[17] ŠČERBAK, A.M., (1970), "Sravnitel'naja fonetika tjurkskix jazykov", Leningrad.

[18] WIJK, N. van (1941), "Zum urslavischen sogenannten Synharmonismus der Silben", Linguistica slovaca 3, p. 41 - 48.

AN EXPERIMENTAL STUDY OF PRONUNCIATION OF STANDARD RUSSIAN

L.A.Verbitskaya

Leningrad State University, USSR

ABSTRACT

The study produced on the basis of a computer version of Russian Derivational Dictionary [1]. Analysis of variation in the pronunciation of borrowings that contain hard or soft consonants followed by the vowel /e/ is presented. Some theoretical implications relevant to the linguistic system of Russian are cosidered.

Standard pronunciation in spite of its tendency to remain stable cannot but react to a constant change of the sound system. As a result, pronuncia- tions formerly considered colloquial or dialectal, become fully accepted today. This constant sound change presents a problem for the description of the system, especially for prescribing a standard.

In recent decades we have observed the process of the emergence of a unified pronunciation standard. A number of its features owe their origin to the linguistic system.

Russian, as well as other languages with a long-standing tradition, has many borrowings from foreign languages; they amount to 10-20% of the wohle lexicon. Many of them constitute an integral part of the lexicon and are judged foreign only by origin. Examples: метр (metre), адрес (address), культура (culture). Most of them follow the sound patterns of the words of Russian origin but there are some whose pronunciation is somewhat different.

Of great interest to this study is the pronunciation of borrowed words containing hard or soft consonants before /e/. In words of Russian origin only soft cognates can occur in this position. Examples: /v'era/ вера (belief), /na stol'e/ на столе (on the table) etc. Hard consonants before /e/ in such borrowings as термос (flask). кафе (cafe) are, therefore, a new feature of Russian pronunciation, but they are, at the same time, the result of the process caused by the linguistic system of Russian. A combination of a hard consonant with /e/ is not an alien feature of Russian, on the cont- rary, it is a potential feature of the Russian sound system which is evidenced by such existing words as жесть /žes't'/ (tin-plate), шесть /šes't'/ (six), шест /šest/ (pole) etc.

As any linguistic phenomenon, pronunciation of hard

or soft cognates bofore /e/ follows some rules. There exist factors that influence the choice of hard or soft consonants. These factors can be divided into three groups: 1) phonetic factors, among which the type of the consonant in question seems to be the most important. The position of this consonant as regards stress is also relevant; 2) morphologicial factors, i.e. whether the borrowed word has acquired declension paradigms in Russian; 3) lexical factors, i.e. the time of the borrrowing and the degree of assimilation of the word in the Russian language.

As has been shown in special studies, the choice between hard and soft cognates before /e/ does not depend on the language from which the word has been borrowed. The choice seems to depend on such individual features as the level of eucation, age and place of residence of the speaker.

There are grounds to believe that pronunciation of borrowings with the vowel /e/ follows the rules different from those applied to the basic vocabulary. Great variation in their pronunciation makes the formulation of the rules especially difficult. They are more complicated than those applied for the words of Russian origin because they are statistically determined and require a lot of data on the pronunciation of each word. Therefore, the first stage of the study was to get an exhaustive list of the borrowings with hard and soft consonants followed by the vowel /e/.

The material for the study has been taken from Russian Derivational Dictionary(2) that contains 110 000 words segmented into morphemes. The dictionary has a great number of borrowings. With the help of a computer a list of all entries with rhe vowel /e/ in the room has been made. It consists of 25 214 words, including borrowings. For further analysis only the latter have been selected. This final list includes 8 275 words in which 2 295 roots with the vowel /e/ occur; all the items have been arranged in alphabetical order. 1 057 roots are represented by one word in the dictionary, 804 roots occur in 2 to 5 words; one root (метр) is represented by 355 words.

The list has been arranged as a card-index: it gives information on the time of borrowing, the language from which the word has been borrowed, the pronunciation of the word, the degree of its assimilation in the Russian language etc.

During the experiment each subject was asked to give his own variant of the pronunciation of every item in the list. All these pronunciations have been analysed and compared with the data in the card-index and with the information about the subjects: their age, level of education etc., as well as with the pronunciation of these words in the pronouncing dictionary [2].

Results

1) The choice of a hard or soft cognate before /e/ depends on the articulatory type of the consonant: about 80% of all the dentals before /e/ were hard consonants.

2) The frequency of occu-

rance of hard cognates is greater in stressed syliab·les.
3) Hard /s/, /r/, /n/, /m/, /f/ occur in words having no declension para- digms. Examples: фрикасе (fricas-see), амбре (scent), кашне (scarf), консоме (consom-mee), кафе (cafe).
4) More often a hard cognate occurs in words known to the subjects but seldom used by them in their own speech.

Thus, the experiment has demonstrated that the tendency for the occurence of hard rather than soft consonants before /e/ in borrowings in present-day Russian seems to be very strong, and, I believe, it cannot be ignored in teaching standard pronunciation.

References:

[1] Worth, D.S., Kozak, A.S., Johnson, D.B. Russian Derivational Dictionary. New York, 1970

[2] A Pronouncing Dictionary of Russian. Ed. by R.I.Avanesov. Moscow, 1987

AUDITORY ANALYSIS OF COMMUNICATIVE MEANINGS IN PREVERBAL VOCALIZATIONS

Y. Isenina

State University of Ivanovo, USSR

ABSTRACT

The aim of the research was to investigate the communicative meanings defined in the acts of communication (CM) and intonation of 70 preverbal vocalizations of five Russian 14-22 months old children by means of auditory analysis. The determination of CM in context and in isolation significantly differed for PV of negation, agreement, request but not for emotional PV of displeasure, joy, anger, admiration. There are two types of PV:diffuse and clear-cut.

1. INTRODUCTION

Vocalizations are sounds uttered by a child in the preverbal period. The acoustic approach to the investigation of preverbal vocalizations is being replaced by functional and semiotic approaches (2,5,6). The experiments showed that preverbal babbling is a means of perceived speech prosody mastering (8). When the child is 7-8 months old the parents correctly perceive the vocalizations of request, hunger and surprise (5). These works gave us the opportunity of putting forward the hypothesis that in the preverbal period the child masters in vocalizations a number of CMs corresponding to the CMs of verbal utterances. The aim of our research were the following: 1. by auditory analysis to investigate the correspondence between the child vocalization CMs and the CMs of some Russian utterances; 2. to investigate the intonation of the determined PV and compare it with the intonation of the corresponding Russian utterances. In order to achieve the first goal the independent experimentators had to determine the CMs of PVs using the full context and then Russian informants had to recognize different types of PVs in context and in isolation. In order to achieve the second aim the intonation parameters marked by the informants were compared with the intonation of the corresponding types of Russian sentences.

2. PROCEDURE

At child's home the experimentor described in a low voice the situation of the communication, the child's gestures, the expression of his face, the direction of the gaze, his actions and the actions of the adults before and after PVs. Every session lasted 1,5-2 hours. The child's PVs, the speech of the adults and the experimentor's words were

tape-recorded. 400 communicative acts of 5 children aged 14-22 months were recorded. After the analysis of all the contexts of the communicative acts by two independent experimenters five types of CMs in PVs were determined: agreement or positive answer to somebody's question or request; demand or request; the request to label an object; negative answer to somebody's question; the request to repeat the just spoken sentence.

The vocalizations expressing the following emotions were also singled out by the experimentors who took the full context into consideration: Joy, admiration, displeasure, anger. The experimental corpus consisted of 20 vocalizations with emotional meanings (4,5,6, 5 of every type of the emotional meanings) and 50 vocalizations of the other above mentioned meanings (10 PVs of every type). Both types of PVs were recorded on two tapes in random order.

Forty nine Russian informants took part in the experiment. Twenty one informants listened to 50 PVs and the description of the communicative situation contexts, 25 informants listened to these PVs in isolation. In comparison with the independent experimentors who judged the meanings of PVs using the full context, the context produced for the informants consisted only of the description of the situation, and the words of the adult. The behaviour of the mother and the child after the communicative act and also the child's mimics and gestures when they could point directly to the type of vocalization were not given (for ex.-nods and shakes, angry expression of the face and so on).

The instruction was as follows: - Listen to every vocalization twice and define its CM. The informants were supplied with a list of meanings determined by experimentors. The number of every PV (produced on a card by the experimentor when the PV was perceived) was to be written opposite its meaning in the list if this meaning was determined by an informant. The same instruction and procedure were carried out when the informants listened to the emotional vocalizations.

Three another informants (the graduates of the philological department specializing in phonetics) listened to the PVs in isolation and graphically represented the changes in voice-pitch (rise, fall, rise-fall, fall-rise); tenseness (tense, lax); loudness (loud, soft); the voice register (high, middle, low).

3. DATA PROCESSING AND RESULTS

There were 33 tables made dealing with the number of PVs in context and in isolation defined "correctly" - in the same way as experimentors. At first we had to verify whether the informants defined the meanings of the vocalizations at random or according to their stable perceived qualities. If the PVs of every type were guessed at random then not more than two of ten could be guessed correctly because the relation of the number of PVs of every type to the number of PVs of all the types would be 1:5. Statistical testing of the hypothesis about the part of

the variants (7) showed that both in context and in isolation communicative meanings were not guessed at random. Statistical analysis (according to T-White criterion (3) showed that the results of determinig PV meanings in context (except emotional PV significantly (P-0,05) differed from the results of determing these meanings in isolation. The results of determining the emotional meanings in isolation did not differ significantly from the results of the perception in context. The mean number of correct guesses of PV of displeasure in isolation $\bar{x}$-81%;in context-$\bar{x}$-82%; anger-$\bar{x}$-56%: (in isolation); $\bar{x}$-66% (in context); joy-$\bar{x}$-84% (in isolation); $\bar{x}$-74% (in context); admiration-$\bar{x}$-68% (in isolation), $\bar{x}$-80% (in context).

4. DISCUSSION

Perceived in isolation the request to repeat the partner's words is the most easily determined PV (the mean number of correct gueses $\bar{x}$-81%).The request to label an object ($\bar{x}$-36%) was being mixed with other requests and demands.Agreement or positive answer ($\bar{x}$-40%) was being mixed with a denial or a negative answer ($\bar{x}$-33%).But the PVs of agreement pronounced with the intonation of Russian utterance "aha"and the PVs of refusal pronounced like Russian utterance "ne-a" were perceived without failures. In order to answer the question why though not being determined at random these kinds of PVs were yet being mixed up and perceived significantly worse than in context one should consider their graphic analysis.

As for emotional PVs in isolation those expressing anger ($\bar{x}$-56%) were being mixed with another negative emotion similar to it but a milder one-displeasure ($\bar{x}$-81%), the same can be said about admiration ($\bar{x}$-68%) and pleasure ($\bar{x}$-68%).
In graphic representations of 3 informants the coincidence in guessing all the parametres was 80% (loudness - $\bar{x}$-80%, tenseness-$\bar{x}$-83%, the changes in voice-pitch-$\bar{x}$-86%, the voice register-$\bar{x}$-56%).
In order to understand our data we turned to the representation of the intonation of the corresponding types of meanings in Russian utterances (1). The request to repeat the words in Russian is conveyed with the help of a high level rising tone. The same tone was represented in the informants' analysis of PVs. That is why this kind of PV was correctly perceived in isolation.
Short negative answer or refusal in Russian is expressed by a falling tone or a rising-falling tone corresponding to the Russian word /n e ə/ which means "no". The informants represented the intonation of all the negative answers as falling but in case of /e ə/ corresponding to the Russian word /n e ə/ as rise-fall.
The intonation of the request to label an object can be compared with the intonation of Russian questions: And this one? And you?which have a falling tune. The informants also represented this kind of request PV as falling. Requests and demands in Russian have a falling tone for demands and a rising tone-for requests.In the same way they were intoned in PVs. As both the

requests to label an object and the general request have a falling tone they were mixed up.
The tone of agreement or a positive answer in Russian can be falling and rising. The same kinds of tones were defined by the informants in PVs and that is why they were mixed with PVs - negative answers or refusals. In three cases the intonation of PV corresponded to the Russian word /ə'h ʌ/ with the meaning "yes" and had a rise-fall tone.
In the Russian language differentiating features of different types of intonational structures of an utterance are the direction of the vowel tone and the distributions of tone levels of the precentral part centre and postcentral parts. In PVs neither precentral or postcentral parts are observed. That is why the context is necessary to define the meaning of many PVs.

6. CONCLUSION

The results of the data analysis made it possible to draw the conclusion that PVs have diffuse and clear-cut meanings. Diffuse meanings have PVs expressing answer to a question including agreement, positive answer and denial. They have a rising and a falling intonation and are easily mixed up.
Such CMs of PVs as "a request to repeat the words", PVs with the meaning of agreement (Russian /ə'h ʌ/) and with the meaning of denial (Russian /h e ə/) could be determined without any context. Their sensory patterns had been formed and were informative enough to be recognized without a precentral or postcentral parts of the utterance. Requests, demands and requests to label should be included into one diffuse group. As for the emotional meanings of the PVs (anger, joy, displeasure, adoration) it appears that their sensory patterns were formed as they were successfully perceived both in isolation and in context.
In this work PVs were analysed and compared with the Russian intonation mostly according to voice pitch. The role of tensness, voice register and loudness may be the topic of further investigation.

7. REFERENCES

1. Brizgunova, E.A. (1980) "Intonatsija". In: "Russkaya grammatica", Moskva: Akademia Nauk.
2. Esenina, E. (1986), "Doslovesniy period razvitiya rechi u detey", Saratov: SGU.
3. Lakin, G.F. (1973), "Biometria", Moskva: Vishaya Shkola.
4. Marcos, H. (1987), "Communicative functions of pitch range and pitch direction in infants", J. of Child Lang. 14, 255-268.
5. Ricks, D.M. (1975), "Verbal communication in preverbal normal and autistic children". - In: Language and cognition: deficits and retardation. Ed. N.O'Connor, L.: Butterworth, 75-80.
6. Roberts, K., Horowitz F. (1986), "Basic level categorization in 7 and 9 months old infants", J. of Child Lang., 13, 191-208.
7. Urbach, V.Y. (1975), "Statisticheskiy analiz v biologichesich i medizinskich issledovanijach", Moskva: Medisina.
8. Vinarskaya, E.N. (1987), "Rannee rechevoye razvitje rebenka i problemy defectologiy". Moskva: Prosveshenje.

PHONETIC AND LINGUISTIC ASPECTS OF PITCH MOVEMENTS IN FAST SPEECH IN DUTCH[1]

J. Caspers and V.J. van Heuven

Dept. Linguistics/Phonetics Laboratory,
Leyden University, The Netherlands.

ABSTRACT

Assuming that speakers tend to preserve the communicatively important aspects of speech, time pressure seems to be a promising experimental tool for isolating the important aspects of intonation. Linguistic hypotheses concerning the optionality of accent and boundary marking pitch movements were tested by having subjects read aloud stimuli in a normal and fast rate. Speakers did not economize on accent lending pitch movements, but 40% of the boundary marking pitch movements disappeared under time pressure, reflecting the linguistic hierarchy in obligatory and optional intonation phrases.

1. INTRODUCTION

We assume that speakers under time pressure will keep unimpaired those parts of the speech signal that are the most important. By comparing normal and fast (read aloud) speech we hope to isolate the more important aspects of intonation. In the present experiment we concentrate on the question if less important accent or boundary marking pitch movements disappear sooner under time pressure than important pitch movements.

2. LINGUISTIC BACKGROUND

2.1. Optionality of Pitch Accent Movements

The notion of integrative accent [1,2] offers an opportunity to distinguish between more or less important accent positions. For example, in the sentence:

(1) There is a tear in your trousers

presenting new information, the most important accent lies on tear (the 'exponent' [2], the constituent on which the integrative accent is placed). The complete utterance can be put into focus, i.e. made important, by just this one accent. However, speakers can choose to highlight other parts of the sentence separately, by placing additional pitch accents on embedded exponents (here on trousers). We formulated the hypothesis that speakers under time pressure can omit pitch accents that correspond to focus domains that can be incorporated into a higher-order focus domain (hypothesis 1).

It is known that a strong correspondence exists between the distribution of new and given information and accent placement: pitch accents generally highlight parts of the sentence containing new information. However, under certain circumstances it is acceptable to put given information into focus by a pitch accent [5]. Assuming that pitch accents highlighting new information are more important than pitch accents focussing given information, we expect speakers to economize on the latter (hypothesis 2).

2.2. Optionality of Boundary Marking Pitch Movements

Speakers use boundary marking pitch movements to highlight communicatively important breaks

in the speech stream. We adopted the phonological theory of prosodic domains [4] to get a grip on differences in importance of prosodic boundaries. The theory presents a range of hierarchical prosodic domains (from "Syllable" to "Phonological Utterance"), of which the "Intonational Phrase" (henceforth 'I'), the domain of intonation contours, is likely to be marked off with a phonetic boundary. The I is a relatively free domain; "root sentences" and "obligatory I's" (cf. [4], p 188ff.) obligatorily form I's, but the I can be restructured, i.e. split up in a number of smaller domains, as a consequence of - for instance - lowering of the speaking rate. This restructuring process is optional, but not completely free; it is limited to positions with a certain syntactic structure. Generally, the optional I-boundaries can occur after a noun phrase (but one cannot separate an obligatory argument from its head) or before an embedded sentence (but an NP may not be broken up). The higher the speaking rate, the smaller the opportunity to restructure an I. From this theory of I-domains we derived the hypothesis that boundary marking pitch movements can dissappear under time pressure when located at an optional I-boundary (hypothesis 3).

3. METHOD

To test hypothesis 1, eight stimuli were constructed of the form:

(2) (Weet je wat die gekke broer van mij heeft gedaan? Hij heeft) een ou[5] de Citroën[4] met voor[1]-wielaandrijving voor z'n vriendin[2] gekocht[3]. (Know you what that crazy brother of mine has done? He has an old Citroen with front-wheel drive for his girlfriend bought.)

The superscript numbers indicate the degree of 'embeddedness' ('DEMB') of the possible pitch accents[2]. A pitch accent on 1 (the exponent) can not be omitted, 2 to 5 are regarded as optional (hierarchically, 5 is considered the easiest to omit). The context sentence has the function of presenting part of the stimulus sentence (not parenthetical) as new information.

To test hypothesis 2, another four sentence pairs were made:

(3) (Salman Rushdie is na lange tijd weer in de openbaarheid verschenen.) De schrijver bood in een televisie-interview zijn excuses aan./ In een televisie-interview bood de schrijver zijn excuses aan. (Salman Rushdie has after a long time again a public appearance made. The writer offered in a television interview his apologies).

The underlined parts of the stimulus have the same referent as the subject of the context sentence (in parentheses). Because we did not know what the influence of the sentence initial place of the given information in the test sentence would be, the stimulus was repeated with the given information in sentence medial position[3].

To test hypothesis 3, six small texts were constructed, consisting of one to four rather long sentences. Configurations for obligatory[4] and optional I-boundaries ('IB') were systematically varied. As obligatory I's, appositions and nonrestrictive relative clauses were used (indicated with '[O' for the left boundary and ']O' for the right boundary), next to root sentences (']R'). The end of a noun phrase (']NP') and the beginning of an embedded sentence ('[S'') were regarded as optional I-boundary positions. Two additional syntactic configurations were systematically varied in the stimulus material. In a pilot experiment we found pauses at places which could not be described in terms of optional I-boundary positions, but only as: an S' within a long noun phrase ('([S'])') and the beginning of a prepositional phrase ('[PP'). Both configurations were regarded as optional I-boundary positions.

The stimuli were printed on

separate cards, using only full stops and capitals, refraining from other punctuation marks, in order to avoid guiding the subjects in placing boundary markers as much as possible (a rather complicated text without ãny punctuation marks is virtually impossible to read aloud). Six subjects read the stimuli aloud in a normal and fast speaking rate.

4. ANALYSIS AND RESULTS

Two phonetically trained listeners independently marked pitch accent positions and boundaries in the relevant parts of the material. A third judge gave decisive judgements in those cases were the other two markers did not agree (13%). The first author transcribed the pitch configuration at each boundary in terms of the Dutch intonation grammar [3].

TABLE 1. Frequency of plus and minus accent scores in normal (N) and fast (F) speaking rate for the five grades of predicted optionality (DEMB, cf. section 3).

DEMB	no accent N	no accent F	accent N	accent F
1	-	-	48	48
2	6	1	42	47
3	42	39	6	9
4	1	3	47	45
5	4	5	44	43

TABLE 2. Frequency of plus and minus accent scores in normal (N) and fast (F) speaking rate for words containing new information (INFO 1), sentence-initial (INFO 2) and sentence-medial given information (INFO 3).

INFO	no accent N	no accent F	accent N	accent F
1	27	39	165	153
2	-	-	24	24
3	4	3	20	21

A hierarchical loglinear analysis of the data in table 1 shows that the effect of speaking rate on accent placement is totally insignificant ($z=.034$, $p=.488$). The same type of statistic analysis was performed on the data in table 2. Again, the factor speaking rate proved insignificant ($z=-.024$, $p=.492$).

TABLE 3. Total number of potential boundaries (N), percentage of boundaries realised in a normal speaking rate and percentage thereof deleted under time pressure, broken down by seven types of I-boundary (IB, cf. section 3).

IB	realised	deleted	N
1]R	100 %	2 %	42
2 [O	80 %	13 %	30
3]O	96 %	9 %	24
4]NP	58 %	61 %	114
5 [S'	64 %	57 %	36
6 ([S')	67 %	38 %	12
7 [PP	34 %	77 %	90
8 Ø	3 %	77 %	846

The strongest reduction in boundary marking occurs at ordinary word boundaries (Ø) and at the beginning of prepositional phrases. The end of a root sentence is almost always marked, as are the edges of appositions and nonrestrictive relative clauses. In between lies the group of optional I's, extended with the category 'S' in NP'. ANOVA shows that the effect of I-boundary type is significant, $F(7,231=17.6)$, $p<.001$. Newman-Keuls post hoc analysis shows further that there are no internal differences among the obligatory boundaries (types 1,2,3), nor among the optional boundaries (types 4,5,7,8). However, type 6 ('S' in NP') does not differ from either of these two groups ($p<.05$).

In both normal and fast tempo-conditions approximately 95% of the perceived boundaries received a boundary marking pitch movement. Normal/fast boundary pairs were subjected to a further analysis, exploring the possibility that complex boundary marking pitch movements used in normal speech would be replaced by simpler movements in fast speech. Typical-

ly, the type of pitch movement remains the same in both speaking rates, with the following exceptions:
- If a boundary marking rise ('2', cf. [3] p 73) is followed by a declination reset, generally the reset vanishes in fast speech.
- When the boundary is marked by a late rise plus a non prominence lending fall ('2B'), approximately a third of these boundaries gets the simpler configuration of high declination plus fall ('ØB').

5. CONCLUSIONS

We reject hypothesis 1: when reading aloud fast, speakers do not economize on the number of pitch accents placed on embedded exponents. We also reject the second hypothesis: the accent distribution on given information is the same in both speaking conditions. Simply counting the number of pitch accents realised under time pressure is apparently not sensitive enough a method to bear out differences in communicative importance of pitch accents.

The third hypothesis can be accepted. When a speaker is under time pressure, the number of boundaries dropped is approximately 40%. Boundaries disappear mainly at optional I-boundaries, i.e. F0-markers on optional boundaries are more likely to disappear under time pressure than markers of obligatory boundaries. Of the PP-boundaries one third are marked, which indicates that prepositional phrases can play a role in restructuring I's. Two thirds of the positions with the structure '[S' in NP' are marked by a boundary in the normal speaking condition, forcing us to abandon the linguistic restriction that disallowes the formation of I-boundaries at the beginning of an embedded sentence, that interrupt a noun phrase.

Finally, boundary marking pitch configurations tend to be simplified when the boundary remains in fast speech. Changes in shape of accent and boundary marking pitch movements will be the objects of our future research; as a first approximation we shall examine differences in excursion size of pitch accents in relation to our linguistic hypotheses.

6. REFERENCES

[1] Baart, J.L.G. (1987) "Focus, syntax, and accent placement", Diss.: Rijksuniversiteit Leiden.
[2] Fuchs, A. (1980) "Accented subjects in 'all-new' sentences", Wege zur Universalienforschung, Tuebingen: Gunter Narr, 449-461.
[3] Hart, J 't, R. Collier & A. Cohen (1990) "A perceptual study of intonaton", Cambridge University Press.
[4] Nespor, M. & I. Vogel (1986) "Prosodic Phonology", Dordrecht: Foris.
[5] Nooteboom, S.G. & J.G. Kruyt (1987) "Accents, focus distribution, and the perceived distribution of given and new information: An experiment", Journal of the Acoustical Society of America 82, 1512-1524.

---

1. This research was partly supported by the Foundation for Linguistic Research, which is funded by the Netherlands Organisation for Research, NWO, under project # 300-173-005.
2. We abstract from the probability of a pitch accent on the indicated positions.
3. In sentence final position it is not possible to accent a constituent containing given information [5, p 1521].
4. We use this term for both root sentences and "obligatory I's".

APPROFONDISSEMENTS SUR LA CO-VARIATION ENTRE F0 ET DOUBLEMENT CONSONANTIQUE DANS CERTAINS DIALECTES ITALIENS

Amedeo De Dominicis

Université de Pise, Italie

ABSTRACT

In [1] and [2] I showed that in some Italian dialects the occurrence of the phonological phenomenon called Raddoppiamento Fonosintattico (+RF) is marked by an F0 contour; it differs from the one which is superimposed upon the contexts where the RF lacks (-RF). These tonal features are not of redundant kind. For this purpose, I propose an experiment dealing with synthetic speech. The goal is to verify if, by modifying F0 upon +RF contexts, speakers recognize a -RF and vice versa.

1. HYPOTHESES

Dans [1] et [2] je propose l'hypothèse suivante: dans de nombreux dialectes, le RF serait conditionné par le contour intonatif relatif au contexte phonologique immédiatement contigu à la consonne affectée par le phénomène (contexte V#C:V). La nature de ces traits intonatifs est contrastive, c'est à dire qu'il n'existe pas un seul contour tonal du RF qui soit valable pour tous les dialectes: dans chaque dialecte, la manifestation du RF (en contextes V#C:V) est accompagnée d'un contour intonatif différent; toutefois à l'intérieur d'un même dialecte, ce contour est tout à fait différent de celui qui accompagne les contextes phonologiques où le RF ne se manifeste pas (contextes V#CV).

Les travaux cités laissaient en suspens une question: celle qui concerne le statut théorique soit du contraste tonal entre contextes V#C:V et V#CV soit de la relation entre ce contraste et le phénomène du RF à l'intérieur d'un même dialecte. Il s'agit d'établir si le contour intonatif est simplement un trait redondant du RF, ou bien s'il est - comme je le crois - un trait distinctif et une contrainte de la manifestation du RF dans le cadre d'un dialecte donné. Ici je me propose de rendre compte de la vérification que j'ai conduite à ce sujet.

2. MATERIAUX

Les matériaux utilisés sont trois phrases extraites de l'enregistrement du récit populaire de "Saint Pierre et le jambon" (de [3]: 275). Le récit est produit par la voix d'un locuteur parlant le dialecte d'Introdacqua (AQ), dans les Abruzzes. Introdacqua est un village où le dialecte est encore fort enraciné.

A la suite d'un travail d'analyse spectrographique que j'avais conduit dans ma recherche précédente sur ce dialecte [1], j'avais remarqué que dans la plupart des cas la situation est la suivante. Le type V#CV est lié à un contour de F0 plat, tandis que le type V#C:V l'est au ton descendant.

Le corpus se base sur un exemple du type V#CV ('a#pə): /ʃtu prəsottə ne j 'a pərdutə prOprjə nəʃunə/ ("ce jambon, vraiment personne ne l'a perdu"); et un exemple du type V#C:V (a#'ʧe): /rij'ett a 'ʧesə 'krəʃtə/ ("il retourna à Jésus-Christ").

3. EXPERIENCE

30 sujets ont participé à l'expérience. C'étaient tous des habitants d'Introdacqua, âgés de 27 à 80 ans. Ils ont été classés dans trois groupes d'âge, établis par certaines conventions: 10 "jeunes" (5 hommes et 5 femmes), de 27 à 35 ans; 10 "adultes" (5 hommes et 5 femmes), de 36 à 45 ans; 10 "personnes âgés" (5 hommes et 5 femmes), de 46 à 80 ans.

Chaque sujet a écouté 3 couples de phrases. Les phrases avaient été partiellement manipulées, à l'aide d'un éditeur de parole synthétique, en ce qui concerne les valeurs de durée et de F0. Ces valeurs ont été opposées du point de vue de leurs effets perceptifs.

Le couple 1 se base sur le matériel: /rij'ett a 'ʧesə 'krəʃtə/. La phrase 1 est le signal original à ton descendant sur /a#'ʧe/, donc un exemple du type V#C:V (cf.[1]); la phrase 2 présente une modification du ton sur /a#'ʧe/ qui est ascendant, tandis que la durée de la consonne affriquée (ʧ) reste inchangée.

Le couple 2 se base toujours sur /rij'ett a 'ʧesə 'krəʃtə/. Dans la phrase 1 le ton sur /a#'ʧe/ est inchangé: il est descendant; par contre, la durée de la consonne affriquée (ʧ) est modifiée: elle est réduite de moitié. Dans la phrase 2 le ton sur /a#'ʧe/ est modifié: il est ascendant; la durée de la consonne affriquée (ʧ) est inchangée.

Le couple 3 se base sur le materiel /ʃtu prəsottə ne j 'a pərdutə proprjə nəʃunə/. Dans la phrase 1 le ton sur /'a#pərdutə/ est modifié: descendant; la durée de la consonne occlusive (p) est inchangée. Dans la phrase 2 le ton sur /'a#pərdutə/ est inchangé: il est plat; en revanche, la durée de la consonne occlusive (p) est modifiée: elle a été doublée.

Aprés avoir écouté le couple 1, les sujets devaient indiquer la phrase où /ʧ/ dans /a 'ʧesə/ est plus long. Ils devaient en faire autan avec le couple 2. Par contre, aprés avoir écouté le couple 3, ils devaient indiquer la phrase où /p/ dans /'a pərdutə/ est plus long.

4. RESULTATS ATTENDUS

Le but de cette expérience est de vérifier si, par suite d'une modification de l'intonation sur les contextes +RF, les sujets arrivent à percevoir un -RF et vice versa.

La réponse attendue est toujours la phrase 1 dans les trois couples.

5. RESULTATS

Tableau 1: TOTAL.

	Réponses		
	Phrase 1	Phrase 2	Abstention
Couple 1	17 (57%)	2 (6%)	11 (37%)
Couple 2	11 (37%)	4 (13%)	15 (50%)
Couple 3	14 (47%)	7 (23%)	9 (30%)

Tableau 2: JEUNES.

	Réponses		
	Phrase 1	Phrase 2	Abstention
Couple 1	4 (40%) 2h 2f	0	6 (60%) 3h 3f
Couple 2	4 (40%) 2h 2f	0	6 (60%) 3h 3f
Couple 3	2 (20%) 2h	4 (40%) 2h 2f	4 (40%) 1h 3f

Tableau 3: ADULTES.

	Réponses		
	Phrase 1	Phrase 2	Abstention
Couple 1	5 (50%) 5f	0	5 (50%) 5h
Couple 2	0	2 (20%) 1h 1f	8 (80%) 4h 4f
Couple 3	5 (50%) 5f	2 (20%) 2h	3 (30%) 3f

Tableau 4: PERSONNES AGÉES.

	Réponses		
	Phrase 1	Phrase 2	Abstention
Couple 1	8 (80%) 4h 4f	2 (20%) 1h 1f	0
Couple 2	7 (70%) 3h 4f	2 (20%) 1h 1f	1 (10%) 1h
Couple 3	7 (70%) 4h 3f	1 (10%) 1h	2 (20%) 2f

Tableau 5: HOMMES.

	Réponses		
	Phrase 1	Phrase 2	Abstention
Couple 1	6 (40%)	1 (6%)	8 (54%)
Couple 2	5 (34%)	2 (12%)	8 (54%)
Couple 3	6 (40%)	5 (34%)	4 (26%)

Tableau 6: FEMMES.

	Réponses		
	Phrase 1	Phrase 2	Abstention
Couple 1	11 (73%)	1 (6%)	3 (21%)
Couple 2	6 (40%)	2 (12%)	7 (48%)
Couple 3	8 (54%)	2 (12%)	5 (34%)

On formulera les remarques suivantes:
- Une forte abstention surtout chez les "jeunes" et les "adultes" (en particulier "hommes" plutôt que "femmmes").
- Un fort *matching* des résultats obtenus avec ceux qu'on attendait, en ce qui concerne la catégorie des "personnes âgées".
- Moindre abstention et meilleur *matching* avec les résultats attendus, en ce qui concerne les réponses de la catégorie "femmes" par rapport à celle des "hommes", mais seulement dans la catégorie d'âge "adultes".

6. CONCLUSIONS

- L'hypothèse est validée du moins dans le groupe "personnes âgées".
- Naturellement les résultats obtenus sont plus eloignés de ceux qu'on attendait dans le cas du couple 2 et encore plus dans le cas du couple 3. Mais tous ceci ne constitue pas un problème, car la complexité de la tâche croît dans le couple 2 et surtout dans le couple 3.
- Il faudrait se demander pourquoi les résultats sont aussi différents entre "personnes âgées" d'un côté et "jeunes-adultes" de l'autre.

En outre ces données montrent une différence frappante entre les réponses des hommes et celles des femmes du groupe "adultes".
Probablement, il s'agit simplement d'une issue classique des études dialectologiques et sociolinguistiques:
a) la norme dialectale est gardée plutôt chez les locuteurs plus âgés; les autres, plus jeunes, présentent des interférences avec la langue standard.
b) les femmes montrent généralement une attention et un comportement linguistique plus soigné que les hommes, qu'il s'agisse de vérifications ayant pour objet la compétence linguistique standard, ou qu'il s'agisse de semblables vérifications perceptives ayant toutefois pour objet le dialecte local (cf.[4]: 91). En particulier, dans un contexte social dominé par l'émigration masculine, comme celui d'Introdacqua, les femmes "adultes" sont les seules qui restent et gardent la norme dialectale, que les hommes "jeunes" et "adultes" ne maîtrisent plus car, en général, ils ne rentrent au village qu'à l'occasion des vacances.

7. REFERENCES

[1] DE DOMINICIS, A. (1990), "Fenomeni di 'cadenza' melodica e *raddoppiamento fonosintattico* in alcuni dialetti di area italiana", *L'Italia dialettale*.

[2] DE DOMINICIS, A. (1991), "*Raddoppiamento fonosintattico* and F0 contours in some italian dialects (Calabria and Umbria)", *Papers from the 1990 Cortona Phonology Meeting*, Turin: Rosenberg & Sellier.

[3] GIAMMARCO, E. (1979), "*Abruzzo*", vol. 13 de "*Profilo dei dialetti italiani*" (par les soins de M. Cortelazzo). Pise: Pacini.

[4] TRUDGILL, P. (1974), "*Sociolinguistics*", Harmondsworth: Penguin.

LA VALEUR ABSOLUE DU GRADIENT DE FO : DEFINITION, LOCALISATION ET DISTRIBUTION EN LECTURE DE TEXTE FRANCAIS SOUS TROIS CONSIGNES DIFFERENTES

G. Caelen-Haumont

Laboratoire de la Communication Parlée
UA n° 368 INPG / ENSERG, Grenoble , France.

ABSTRACT

This paper aims at putting forward a new pitch parameter, which is the absolute value of the Fo gradient in the lexical word. Three sets of reading instructions which become more and more exacting with regard to discourse intelligibility, do not upset this parameter prevalence, but nevertheless exert a significant modulation on the distribution of the three parameters analysed in this study.

1. INTRODUCTION

Généralement les études qui portent sur l'analyse du pitch (ou Fo) s'appuient sur les valeurs moyennes calculées sur l'ensemble de la voyelle, sur la partie stable centrale [5], ou aux deux-tiers [4]. Parfois encore 3 références sont prises, aux frontières et au centre de la voyelle [3].

Dans l'étude que nous avons menée sur les relations de coïncidence numériques entre 6 modèles prédictifs (2 syntaxiques, 3 sémantiques, 1 pragmatique) et les paramètres prosodiques [2], nous proposons outre les paramètres de l'énergie et de la durée, et les paramètres mélodiques "classiques" du maximum de Fo (ou FoM) et Fo moyen (ou Fom), un nouveau paramètre mélodique qui s'est révélé très efficace, à savoir la valeur absolue du gradient de Fo (ou |ΔFo|). Ce nouveau paramètre est dans cette étude relatif aux unités lexicales.

L'expérimentation porte sur un texte[1] de 30 mots lexicaux composé de 3 phrases et de 11 "groupes minimaux". Nous rappelons [1] que ces groupes minimaux sont définis syntaxiquement et prosodiquement comme les groupes syntaxiques de plus bas niveau, immédiatement supérieurs à la structure superficielle, éventuellement associés avec le groupe syntaxique suivant de manière à former une structure de 5 syllabes, nécessaire et suffisante pour l'autonomie prosodique du groupe. Trois consignes de lecture ont été présentées à 12 locuteurs (1° lecture naturelle et intelligible 2° lecture très intelligible 3° lecture très très intelligible pour un ordinateur). Avant l'enregistrement, les mots les plus spécialisés ont été explicités, si besoin était, car l'expérimentation ne portait pas sur la compréhension du texte, mais sur la communication de cette compréhension, autrement dit, sur le "faire-comprendre".

Les 36 enregistrements organisés en base de données, ont été segmentés et étiquetés par un expert-phonéticien.

[1] Le texte est le suivant : "D'éminents biologistes et d'éminents zoologistes américains ont créé pour des vers géants un nouveau phylum dans l'actuelle classification des nombreuses espèces vivantes. Ces longs vers prospèrent sur le plancher marin des zones sous-marines profondes. Des sources thermales chaudes y maintiennent une température moyenne élevée."

2. CHOIX DES PARAMETRES MELODIQUES

L'étude dans son ensemble analyse 14 paramètres mélodiques qui se subdivisent en 3 types (valeur absolue du gradient de Fo, maximum de Fo et Fo moyen), en trois localisations (ensemble du mot, syllabe finale, "contour"), et deux contextes de référence (le texte et la phrase). Ces contextes de référence définissent en fait deux espaces de réduction des valeurs numériques, réduction qui concerne aussi bien les modèles que les paramètres mélodiques. Compte-tenu de la petite part de connaissance que véhicule chaque modèle pris isolément, et de toutes les sources de variabilité tant linguistiques qu'extra-linguistiques, un espace à quatre niveaux nous a semblé correspondre à un juste compromis.

L'ensemble de ces combinaisons aboutit à 18 paramètres, mais le nombre a été réduit à 14 par suppression de Fo maximum et Fo moyen dans le contour et dans les deux contextes de référence.

En outre de manière à rendre égales les conditions de sélection de ces divers paramètres en vue d'une comparaison inter-locuteurs, les frontières des items ont été localisés à l'écran en prenant soin de ne pas relever aux bornes de ceux-ci, — qui sont souvent le lieu des valeurs numériques extrêmes—, les unités phonétiques réputées non voisées, de même que les /ə/ qui leur sont postérieurs, le voisement du premier ou l'existence du second relevant de la variabilité locuteurs.

Dans cette communication nécessairement réduite par rapport à l'autre [2], nous n'envisagerons que les 3 types de paramètres (|ΔFol, FoM, Fom) en neutralisant contextes de référence et localisations.

3. CRITERES DE REALISATION DES PARAMETRES. CONTEXTE DES PHRASES

Le paramètre ΔFo est certainement le paramètre le plus délicat à réaliser pour le locuteur dans la mesure où il exige de positionner au sein des pentes mélodiques croissantes et décroissantes le temps de quelques ms deux cibles qui sont à la fois par rapport au mot lexical des extrema absolus (et relativement inverses), et par rapport à la chaîne mélodique de la phrase et du texte, des extrema relatifs. Lorsque pour une raison ou une autre, l'effort est trop grand, les locuteurs positionnent une des deux cibles, en l'occurrence le maximum de Fo, en une position clé du mot lexical. Lorsque ces conditions sont encore trop difficiles, il suffit alors d'ajuster au besoin par approximations successives grâce au feed-back, les valeurs mélodiques moyennes du registre voulu pendant l'énonciation des unités phonétiques voisées du mot, soit un temps considérablement plus long. Ces 3 paramètres |ΔFol, FoM et Fom semblent en fait se comporter comme les avatars progressivement détériorés d'un même processus.

En ce qui concerne les phrases du texte, on remarque que la phrase 1 a la propriété à la fois d'être la plus longue (de l'ordre de deux fois) et de détenir les mots les plus spécialisés. La phrase 2 possède le lexique d'accès le plus facile. La phrase 3 est la plus courte mais présente une information inattendue.

4. METHODOLOGIE D'ANALYSE

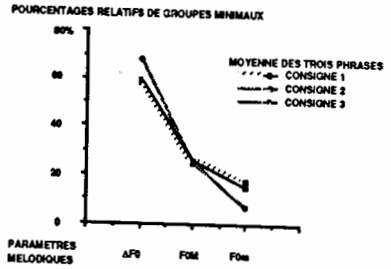
Les combinaisons modèle/paramètre offrant le plus grand nombre de coïncdences ayant été retenues, le principe d'analyse consiste à suivre d'aussi près que possible l'évolution de la distribution des modèles et des paramètres qui leur sont liés. Le groupe minimal est l'élément de base, nécessaire et suffisant, pour être la cible d'un changement de stratégie, mais dans la majeure partie des cas, il se combine à d'autres pour former des macro-structures signifiantes qui fournissent précisément le support à l'expression de la stratégie mise en oeuvre par le locuteur.

Dans ces conditions, la méthode de travail consiste à sélectionner pour le premier groupe minimal, la meilleure combinaison modèle prédictif / paramètre mélodique, —meilleure au sens numérique—, et ensuite à trouver pour le ou les groupes suivants, le meilleur compromis entre ces meilleurs taux de coïncidence et les principes de cohésion et de cohérence du système qui poussent à conserver le cadre conceptuel et mélodique, c'est-à-dire le meilleur compromis entre la dynamique et l'économie du système.

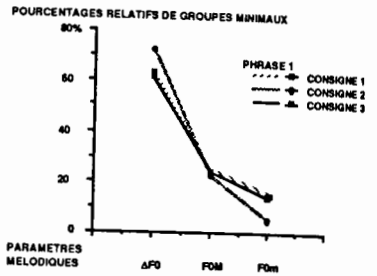
4. RESULTATS

L'étude présente se limite à l'analyse générale de la distribution des paramètres mélodiques tous locuteurs confondus, en fonction des consignes de lecture et des phrases. Nous comparerons les paramètres sous l'angle de leurs pouvoirs explicatifs.

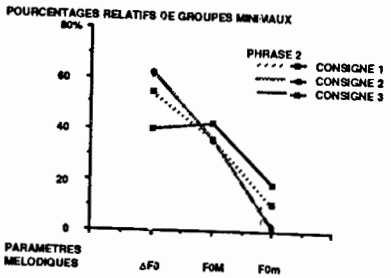
Le graphique 1 ci-dessous présente les pourcentages moyens relatifs des trois paramètres, toutes phrases confondues, en fonction des 3 consignes. Alors que les débits moyens ralentissent très sensiblement, nous constatons que les effectifs des consignes 1 et 3 restent voisins. Les débits de parole (plus pauses) varient en effet tous locuteurs confondus, respectivement de la consigne 1 à la 3, de 2.23 à 1.82 puis à 1.05 mots / seconde. Il ressort des pourcentages que le paramètre |ΔFol correspond en consignes 1 et 3, à 58 ou 59% des observations totales des 3 paramètres, alors que Fom ne compte que 15 à 17% de celles-ci. FoM quant à lui, reste très stable puisqu'il recueille 25 et 26% effectifs, score invariant en consigne 2



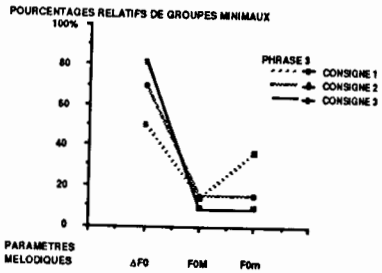
Graphique n° 1



Graphique n° 2



Graphique n° 3



Graphique n° 4

Graphique n° 1 : Pourcentages moyens relatifs des groupes minimaux tous locuteurs confondus, toutes phrases confondues, en fonction des trois types de paramètres mélodiques, Valeur absolue du gradient de Fo (ou |ΔFo|), Maximum de Fo (ou F0M) et Fo moyen (ou F0m).

Graphique n° 2 à 4 : Pourcentages moyens relatifs des groupes minimaux tous locuteurs confondus en phrase 1 (graphique 2), en phrase 2 (graphique 3), en phrase 3 (graphique 4) , en fonction des 3 consignes de l'énoncé et des trois types de paramètres mélodiques, Valeur absolue du gradient de Fo (ou |ΔFo|), Maximum de Fo (ou F0M) et Fo moyen (ou F0m).

(25%). Par rapport aux effectifs des autres consignes, la consigne 2 opère donc une augmentation de ceux de |ΔFo| (68%) au dépens exclusif de Fom (7%). L'évolution de la distribution des effectifs des trois paramètres montre donc dans l'ensemble et indépendamment de l'évolution de la distribution des modèles linguistiques, que l'effet des consignes de lecture, s'il joue considérablement sur le débit de parole, ne modifie pas du tout au tout le choix des paramètres. La consigne 2 toutefois représente une réponse (version paramètres mélodiques) à la première exigence d'augmentation de l'intelligibilité en favorisant au prix de la difficulté de réalisation, le paramètre le plus délicat à mettre en oeuvre, |ΔFo|. La consigne 3 apporte une autre solution qui ramenant les effectifs des paramètres à ceux de la consigne 1, privilégie, indépendamment du domaine conceptuel des modèles linguistiques étudié par ailleurs [2], le paramètre durée de réalisation des unités phonétiques et des pauses.

Les graphiques 2 à 4 montrent l'effet des consignes sur chacune des phrases. Le graphique 2, très proche du graphique 1, maximalise les effectifs de |ΔFo| avec respectivement 63 à 61% des effectifs totaux pour les consignes 1 et 3, et 72% pour la consigne 2.

Le graphique 3 montre que la distribution

des effectifs en consigne 1 est globalement intermédiaire des effectifs des consignes 2 et 3. De manière générale les effectifs de |ΔFo| sont, quelle que soit la consigne, toujours inférieurs (la fourchette est comprise entre 40 et 62%) à ceux de la phrase 1. Il est intéressant de constater que cette dégradation se fait au profit du paramètre qui ne requiert pas le moins d'attention puisque Fo maximum connait une nette progression au dépens également de Fo moyen. La consigne 3, la plus contraignante de toutes, révèle une perturbation assez remarquable puisque, exceptionnellement, FoM recueille les pourcentages les plus grands (42%) au détriment de |ΔFo| (40%), évolution négative qui alimente aussi Fom (18%).

Le graphique 4 qui illustre la phrase 3 est caractéristique. Dans les conditions de lecture naturelle et intelligible (consigne 1), les effectifs de |ΔFo| sont toujours prédominants, mais depuis la phrase 1, ils ne cessent de décroître, ce qui peut facilement s'expliquer par les caractéristiques des phrases elles-mêmes ou encore l'effet fatigue (respectivement 63% -> 54% -> 50%). Lorsque les consignes de lecture exigent plus d'intelligibilité, le contrôle de l'utilisation des paramètres se révèle plus ferme. Il s'opère alors un retournement de tendance et les effectifs de |ΔFo| augmentent sensiblement (50% -> 70%) au détriment exclusif de Fom (36% -> 15%). La consigne 3 prolonge l'effet en augmentant encore les effectifs de |ΔFo| (70% -> 81%) au dépens de Fom mais aussi de FoM et dans les mêmes proportions (15% -> 9%).

Si l'on analyse ces résultats en prenant en compte d'une part la succession des phrases et leurs caractéristiques propres, il apparaît que 1° le contenu de l'information de la phrase 1 est communiqué avec une grande attention en utilisant dans la proportion globale de 2 fois sur 3, le paramètre de |ΔFo| 2° dans les conditions de lecture naturelle et intelligible, l'effet d'un contexte moins difficile (et sans doute aussi de la fatigue) se fait progressivement sentir de la phrase 1 à la phrase 3, détériorant progressivement les performances de |ΔFo| 3° une consigne de lecture "moyennement" contraignante (lecture très intelligible) a pour effet de potentialiser en moyenne les ressources des locuteurs et d'augmenter très sensiblement la proportion de |ΔFo| dans les phrases 2 et 3 4° une consigne encore plus stricte (lecture très très intelligible pour un ordinateur) a inversement l'effet de radicaliser les comportements de la consigne 1, en accusant fortement la détérioration attestée en phrase 2, mais inversement le relâchement substantiel de la tension en cette phrase 2 a pour effet de recréer les conditions favorables à une attention plus soutenue en phrase 3, ce qui est effectivement réalisé comme le montre la sélection massive de |ΔFo| (81%).

5. CONCLUSION

Cette communication a révélé l'efficacité d'un nouveau paramètre mélodique, qui se définit dans le cadre du mot lexical, et qui est la valeur absolue du gradient de Fo. Ce paramètre est sélectionné par les locuteurs dans les proportions globales de 2 fois sur 3, alors que le maximum de Fo représente aussi les deux-tiers des effectifs restants. Des consignes de lecture plus contraignantes ne remettent pas généralement en cause sa suprématie mais modulent de manière significative la distribution des effectifs de ces 3 paramètres. Ce paramètre possède la propriété d'exprimer de la manière la plus adéquate, la relation entre d'une part l'organisation cognitive des informations (approximées par les modèles) et d'autre part l'organisation mélodique de la chaîne parlée, mais mobilisant de ce fait fortement les facultés d'attention des locuteurs, il est nécessairement relayé par d'autres paramètres moins exigeants, mais aussi moins expressifs.

REFERENCES

[1] CAELEN-HAUMONT, G. (1989), "Une représentation syntaxique adaptée à la prosodie", *J. d'Acoustique*, 2, 137-146.

[2] CAELEN-HAUMONT, G. (1991), "*Stratégies des locuteurs et consignes de lecture d'un texte: analyse des intéractions entre modèles syntaxiques, sémantiques, pragmatique et paramètres prosodiques*", Thèse d'Etat, Aix-en-Provence.

[3] EMERARD, F., BENOIT, C. (1987) "De la production à l'extraction, l'état d'un chantier", 16èmes JEP, SFA-CNRS, Hamammet, Tunisie 224-226.

[4] ROSSI, M. (1971), "Le seuil de perception des glissandos", *Phonetica, 23*, 129-161.

[5] VAISSIERE, J. (1989) "On Automatic Extraction of Prosodic Information for Automatic Speech Recognition System", EUROSPEECH, Vol. 1, Paris, 202-205.

DYNAMIC MODEL OF PROSODY IN THE SYSTEM OF SPEECH PRODUCTION

A. METLYUK

MINSK STATE PEDAGOGICAL INSTITUTE OF FOREIGN LANGUAGES

ABSTRACT

The units of all the subsystems of language, prosody in particular, are viewed in this paper as subprocesses or operations in the complex speech production mechanism. Each operation is determined by its aim, so the hierarchy of the units follows the hierarchy of the aims. The operations are reproduced repeatedly in the functioning of language, and the mechanism of the relations of the units remains invariable, irrespective of the functional state of the system.

1. INTRODUCTION

Language as a system of speech production is dynamic not only diachronically but synchronically as well: it changes the structural and functional state of its units in accordance with the thought content, communicative purport and the situation in which the intercourse takes place.

The main factor that forms language as a systemic object is its function (aim), that is the production of the text or the utterance, treated as a minimal text. In other words, it is the actualization and materialization of the thought content in a given situation. The aim of the system, coordinated by the more complicated system "man-society - reality "(part of which is language itself), determines the position, structure and functions of all the units of the particular systems and subsystems of language.

In our experimental and theoretical studies of English, Byelorussian and Russian prosodic units the position of the prosodic system and its components - temporal, accentual, rhythmic, tonal and pausal subsystems - is defined as parallel to the system of unilateral semantic (content) units and the system of bilateral, sign units on the horisontal plane of speech production. On the other hand it functions parallel to the system of phonemes on the vertical plane and forms together with the latter the hierarchy of phonological units of language [1].

At present there is no agreement among linguists as to the set of distinguishable prosodic units and the relations between them in the phonological hierarchy. In this paper I make an attempt to give my interpretation of prosodic units as well as a general presentation of the idea of a processual prosodic system in the functioning of language.

2.THE FUNCTIONING OF PROSODIC UNITS

The specific function of prosody in language is to integrate segmental units into larger segments on all the levels of speech production - the level of the syllable, the word level and the utterance level -, to transform a segment into a unit of a higher level and to differentiate segments. But the distinctive function is not characteristic of all the units and levels of prosodic hierarchy. In these three functions prosodic units are qualified as unilateral units of expression.

Unlike these, the prosodic units of the utterance (e.g. tonemes, contouremes) function as linguistic signs and, consequently, take part in the production of both the formal and the semantic structure of the utterance.

The interaction of different subsystems of prosody in fulfilling the common communicative task is determined by their identical structure on the one hand, and by the specific quality and functions of their units, on the other hand. Common to all the subsystems is the presence of two types of units - elements (microprosodemes) and structural complexes of elements or " phonological syntagms " (macroprosodemes) .

The first type includes syllable chronemes, accentemes, tonemes. The second type is represented by temporal, accentual, rhythmic and tonal structures. The structures as syntagmatic units contain the rules of positional and combinatory variation of microprosodemes as well as the variation caused by the interaction of one subsystem with another. At the same time structures serve as patterns (rules) of grouping microprosodemes and when opposed to one another, form a paradygm in each subsystem.

Besides the elements and structures as units of expression, the prosodic subsystem (tonal in particular) contains semantemes i.e. generalized meanings of definiteness/indefiniteness, finality/nonfinality, completeness/incompleteness, etc. The prosodic mechanism of language is aimed at establishing relations between the formal and the semantic units of prosody, i.e. at the production of the prosodic structure of the utterance or that of its meaningful part. Utterance prosody is formed by the interaction of all the prosodic subsystems, starting from the level of the syllable, and stands out as a polycomponental and polyfunctional sign unit.

The processual character of all the prosodic units is conditioned by the syllable - the basic point of prosody. Due to the integrative quality of its temporal component (syllablechroneme) it demonstrates the mechanism of the interaction between phonemes and prosodemes. The syllable chroneme is the elementary unit of the temporal subsystem of prosody [2]. It forms macroprosodemes - temporal structures (of words, rhythmic units, intonation groups, utterances) and indicates the tempo of speech.

The temporal (syllabic) structure of the word is transformed into the minimal unit of the rhythmic subsystem - the rhythmic group - when rhythm is stress-timed. The new, temporal quality of rhythmic groups, as minimal

units of utterance rhythm is their relative isochrony, which does not depend on the number of the syllables, whereas in the temporal structure of words and rhythmic groups viewed as macroprosodemes this factor is significant. When rhythm is syllable-timed the syllable chroneme functions as a minimal unit.
Actually, the rhythmic structure is derived from the interaction of the temporal and accentual structures.
The relations between the accenteme, the two parallel structures and the sign they constitute take the form of a structural frame, or a cell in the general structure of language, thus demonstrating the principle of organization of a particular subsystem on the horizontal plane and the interpenetration of adjacent subsystems on the vertical plane.The number of the cells on each level correlates with the number of the microprosodemes that take part in the production of the sign unit. The higher the level, the larger the number of the prosodic units.
So in the process of organization of the utterance the prosodemes of a lower level (both elements and structures) provide a basis for the prosodemes of a higher level: the syllable chroneme, as a measure of linguistic time, conditions the occurrence of the accenteme; the accenteme initiates the toneme. Similar are the links between the structures of these units, whereas the structures of one subsystem acquire the status of minimal units on each higher level. A particular subsystem on each level can therefore be regarded as a two - level subsystem with dynamic objects (Cf [3]).

3. CONCLUSION

The systemic description of prosody, i.e. the functional, elementaristic and structural analyses combined make it possible to give a more precise presentation of prosodic hierarchy as: 1) a hierarchy of the prosodic subsystems conditioned by the levels of speech production; 2) a hierarchy of the subsystems on each level of speech production; 3) a constitutive hierarchy of micro- and macro-prosodemes; 4) a subordinative hierarchy of microprosodemes within macroprosodemes.
Moreover, it permits to reconsider such linguistic problems as speech segmentation, semantic structures,variant - invariant relations.
The units of the prosodic system as subprocesses in the complex mechanism of speech production can be presented in the form of algorhythms, which can be helpful in dealing with problems of applied linguistics.

4. REFERENCES

[1] METLYUK, A. (1987), "Vzaimodejstvije prosodicheskikh sistem v rechi bilingva", Minsk: Vishejshaja shkola.
[2] METLYUK, A. (1989), "Prosodicheskije jedinitsi kak podprotsessi v jazikovoj sisteme porozhdenija teksta", Problemi dokazatelstva i tipologizatsii v fonetike i fonologii, Moskva:Akademija Nauk SSSR.
[3] MESAROVIC,M.D.,MACKO, D., TAKAHARA, Y. (1970), "Theory of Hierarchical Multilevel Systems",New-York, London.

ON THE DISCOURSE FUNCTION OF INTONATION

Dieter Huber

Chalmers University of Technology
Department of Information Theory
S-412 96 Gothenburg
Sweden

ABSTRACT

This study explores the differences between discourse intonation and the kind of pitch contours typically found in isolated sentences. Three kinds of material are evaluated systematically: (1) orally read lists of semantically unrelated sentences, (2) orally read narrative texts, and (3) dialogues. The material consists of equivalent samples of Swedish, English and Japanese speech, produced by native speakers (both female and male) of the respective languages. It will be shown that discourse intonation differs from intonation in semantically unrelated sentences with respect to practically all F_0 parameters investigated in this study.

1. INTRODUCTION

Human speakers typically associate their verbal speech utterances with intricate patterns of voice fundamental frequency. This phenomenon has been widely attested, and is acknowledged as a universal, innate quality of speech, common to all speakers, in all languages, and in all kinds of spoken utterances. Numerous scientific studies within a variety of disciplines have been undertaken to investigate the form and function of these fundamental frequency patterns, to establish their communicative status, and to disentangle the seemingly infinite variety of linguistic and paralinguistic conditioning factors that human speakers so aptly and without apparent effort combine into one single contour. Most of these studies have been restricted to the domain of the sentence as maximal unit of linguistic processing, thus adhering to the traditional view that larger units like paragraphs, text and discourse are formed by mere juxtaposition of autarchic, independently prefabricated sentences. There is, however, convincing evidence that human speakers use variations in voice fundamental frequency in a systematic way to signal *cohesion*, *structure* and *prominence* in connected speech according to criteria other than purely syntactic, and that listeners at the other end of the speech communication chain are able to detect and to decode these prosodic messages, and to make use of them in order to gain information about the intended meaning of the utterance in its situational and co-textual context. The purpose of this study is to investigate these differences, i.e. between discourse intonation and the kind of pitch contours typically found in isolated sentences.

2. DATA

Three kinds of material are evaluated systematically: (1) orally read lists of semantically unrelated sentences, (2) orally read narrative texts, and (3) dialogues. The material has been selected from the ATR [7],[8] and the CTH [2] speech databases and comprises equivalent samples of Swedish, English and Japanese speech. The English and Japanese dialogues consist of simulated telephone conversations conducted within the applications domain of conference registration, whereas the Swedish dialogues were conducted spontaneously.

Ten native speakers of the respective languages participated in the recordings selected for this study: 3 speakers of Standard Swedish (2 male, 1 female), 3 speakers of American English (2 male, 1 female) and 4 speakers of Standard Japanese (3 male, 1 female). Registration of the speech samples was conducted in anechoic, sound-insulated recording studios both at ATR in Kyoto (Japan) and at CTH in Gothenburg (Sweden), using high-quality digital recording equipment.

3. ANALYSES

Approximately one minute of recorded speech per speaker and speech style was analysed for this study. Pitch extraction was performed using the DWAPIT pitch determination algorithm presented earlier in [3]. Pitch estimates were obtained at 16-ms intervals for both periodic and aperiodic (laryngealized) stretches of speech. Segmentation of the F_0 tracings into *intonation units* (IU) was performed following the approach published in [4]. According to this approach, two global declination lines which approximate the trends in time of the peaks (topline) and valleys (baseline) of F_0 across the utterance, are computed by the linear regression method. Computation is reiterated every time the *Pearson correlation coefficient* drops below a preset level of acceptability. Segmentation is thus performed without prior knowledge of higher level linguistic information, with the termination of one unit being determined by the general resetting of the intonation contour wherever in the utterance it may occur. The F_0 onsets (intercepts) and offsets (endpoints), durations, declination line slopes and key values of these intonation units, as well as their time-alignment with features of linguistic structure were established individually for each of the speakers participating in this study.

4. RESULTS

4.1 Number of Intonation Units

A total of 586 intonation units has been established in the accumulated material for all ten speakers. The distribution of these intonation units per language and speech style is summarized below in figure 1.

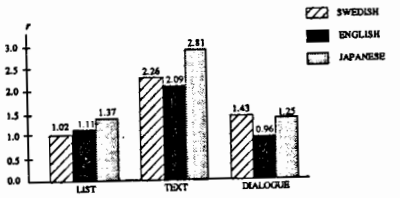


Figure 1. Intonation units per language and speech style. The bar heights *r* depict the ratio between the number of intonation units and the number of sentences contained in the respective material.

These distributions reveal a clear and consistent tendency, observable in each of the three languages, to subdivide orally read texts into a larger number of prosodically cued *chunks* than both the list and the dialogue material. All ten speakers produced predominantly one intonation unit per sentence in the list reading task, as predicted by most studies of sentence intonation, whereas in the text reading task the individual sentences were processed on the average in terms of between 2 and 3 intonation units.

Quite obviously, differences in sentence structure and informational content need to be taken into account for a comprehensive assessment of these ratios. This is particularly relevant with regard to the *r* values obtained for the dialogues which clearly reflect (1) the comparatively larger proportion of short and incomplete sentences included in the material, and (2) the more frequent use of intonation units that stretch over the time extent of several consecutive sentences (cf. [6] for a more detailed discussion). Also, the dialogue material investigated in this study contains significantly less subordination than the texts and sentences, and only few examples of it-clefts and wh-clefts that typically occur as separate, prosodically cued chunks in read narrative.

Considering the higher degree of interlanguage variability found in the

dialogues, it must also be appreciated that the Swedish material consists of spontaneous conversations, i.e. including a larger proportion of hesitations, false starts, fragmentary constructions, etc. than the simulated dialogues in the English and Japanese samples.

4.2 Prosody-Syntax Alignment

The overwhelming majority (84.6%) of intonation units identified by the segmentation algorithm correspond in a clearly defined way with units of syntactic structure. This regular syntax-prosody correspondence, however, is significantly more prevalent in the Japanese (98.2%) than in English (82.2%) and Swedish (79.9%) material. It is also slightly more pronounced in the orally read texts (85.5%) as compared with the dialogues (83.8%).

Most commonly in our accumulated dialogue material, intonation units correspond in a regular fashion with single sentences (40.3%), whereas in the text material the results are more inconsistent between the three languages investigated in this study. In 36.6% of the English and 32.4% of the Swedish texts, intonation units time-align with clauses. In the Japanese text material, on the other hand, only about one tenth (10.1%) of the intonation units pertain to the clause correspondence class, thus indicating a markedly different prosodic processing behaviour.

Larger structures beyond the sentence domain (i.e. stretching over two or three consecutive sentences) are almost exclusively found in the dialogues, with only 1.1% 3-sentence occurrences in the English and 2.1% 2-sentence occurrences in the Swedish texts. Conversely, intonation units corresponding to single constituents in the subsentence domain (i.e. nounphrase-subjects, verbphrases, adverbials, parenthetical constructions, etc) occur more often in the text (41.9%) than in the dialogue (24.9%) material, with a significant prevalence in the Japanese (60.3%) as compared with both the English (35.9%) and Swedish (29.5%) speech samples.

Only the discourse material has been scrutinized at such a detailed level of linguistic analysis. For the speech samples produced in the list reading task, a predominant one-to-one relationship between isolated sentences and single, coherent intonation units has already been established in the previous section.

4.3 Declination Line Parameters

The declination line parameters onset (intercept), offset (endpoint), duration, slope and key were calculated separately for each of the 586 intonation units investigated in this study. Statistical evaluation of these data revealed the following tendencies:

(1) Intonation units aligning with the isolated sentences from the list reading task are on the average shorter, steeper, less varied, and start with higher baseline onsets and substantially lower topline intercepts than in the discourse material;

(2) Important features of prosodic variation such as for instance rising baselines, "bi-modal" toplines, and narrow *versus* wide key (cf. [5]) do not occur in the list material at all, but are frequently used in discourse;

(3) The only parameter for which no statistically significant differences could be established between the different kinds of material is the baseline endpoint, which thus appears to provide a common point of reference, marking the bottom of a speakers voice range for both discourse and isolated sentence production.

Separate investigation of both the IU initial and IU final peaks and valleys, in order to account for the potential status of these points as independently controlled linguistic variables (e.g. [1]) revealed:

(4) significantly higher measures of variability for both the very *first* and the very *last* peaks and valleys in the intonation unit contours of the dialogue as compared with both the sentence and text material;

(5) the consistent use of categorical distinction by all ten speakers with respect to both the first and the last peak/valley of the IU contour in the discourse but not in the list material.

4.4 Laryngealization

Patterns of aperiodic voice vibration (laryngealization) were observed to occur at various kinds of textually, syntactically and prosodically induced boundaries in our material. The acoustical characteristics of these patterns and their function as complementary/compensatory boundary cues have been discussed earlier in [3]. It has also been claimed that female speakers differ in a systematic way from male speakers in their use of laryngealization in connected speech [5]. This claim, based originally on Swedish text material, is further substantiated by the results of the present investigation, which show that the three female speakers participating in this study:

(1) make distinctly more frequent use of laryngealization as a boundary marker than their male counterparts (on the average 13.4% *versus* 8.1%);

(2) apparently prefer to employ *creak* patterns at pre-boundary positions where the men - in as far as they use any laryngealization at all - produce predominantly *creaky voice*.

There are, however, significant differences in the frequency of occurrence of these patterns between the three languages, as reflected in the following percentages:

SWEDISH	26.8%
ENGLISH	33.4%
JAPANESE	39.8%

Even more importantly, the use of laryngealization as a boundary cue differs markedly between the three kinds of material, where it occurs least frequently in the lists of semantically unrelated sentences (Swedish 7.3%; English 10.1%; Japanese 13.5%) and most frequently in the narrative texts (Swedish 60.4%; English 54.2%; Japanese 49.3%). The respective figures for the dialogue material (Swedish 32.3%; English 35.7%; Japanese 37.2%) reveal a somewhat intermediary status for the conversational speaking mode.

In summary, laryngealization as a boundary marker (either alone or together with other juncture cues such as for instance pause, declination resetting, F_0-fall-rise patterns, devoicing, phonological blocking, etc) displays its strongest potential in the highly structured and optimally controlled text reading mode, whereas it is used to a significantly lesser degree in the other two speaking styles, i.e. where the boundaries are signaled by other linguistic (e.g. semantic incoherence between the sentences on the list) or paralinguistic (e.g. changes in voice quality at conversational turn boundaries) means.

REFERENCES

[1] BRUCE,G. (1982), "Textual aspects of prosody in Swedish", *Phonetica* 39, 274-287

[2] HEDELIN,P.& D.HUBER (1990), "The CTH speech database: An integrated multilevel approach", *Speech Communication* 9(4), 365-374

[3] HEDELIN,P.& D.HUBER (1990), "Pitch period determination of aperiodic speech signals", *Proc. ICASSP-90*, 361-364

[4] HUBER,D. (1989), "A statistical approach to the segmentation and broad classification of continuous speech into phrase-sized information units", *Proc. ICASSP-89*, 600-603

[5] HUBER,D. (1989), "Voice characteristics of female speech and their representation in computer speech synthesis and recognition", *Proc. EUROSPEECH-89*, 477-480

[6] HUBER,D. (1990), "Speech style variations of F_0 in a cross-linguistic perspective", *Proc. SST-90*, 186-191

[7] HUBER,D. (1990), "A bilingual dialogue database for automatic spoken language interpretation between Japanese and English", *ATR Technical Report*

[8] KUREMATSU,A. *et al.* (1990), "ATR Japanese speech database as a tool for speech recognition and synthesis", *Speech Communication* 9(4), 357-363

FALLS:
VARIABILITY AND PERCEPTUAL EFFECTS

Anne Wichmann

IBM(UK) Scientific Centre
Winchester, SO23 9DR, England

ABSTRACT

This paper presents an experiment designed to test the effect of final intonation contours on the degree to which an utterance is perceived to be final. The utterances were taken from a corpus [3] of naturally occurring monologue. Each was syntactically complete and semantically unmarked for finality. Keeping the endpoint constant, the starting point of the final fall was sytematically manipulated to create 5 different versions of each sentence. The results of the perception experiment suggest that the higher the starting point of the final fall, the less final that utterance is perceived to be. There is no evidence for any discrete perceptual categories.

1. INTRODUCTION

In abstract representations of intonation, the end of a declarative utterance is generally indicated by assigning a falling contour. Physically, a fall can be any pitch contour that ends at a pitch lower than its starting point. Since both starting point and endpoint are variable within the range of any one speaker, there are any number of falls which that speaker can produce. It is generally assumed, however, that these physical differences are not significant, and that the height of the fall is determined by the declining topline across the utterance. Any significant differences in the resulting overall contour have in the past been related to the slope of the fall, residual fall, and endpoint. This study shows that the starting point also has a systematic perceptual effect.

Other experimental studies of the acoustic correlates of boundaries [1] [2] have compared the physical realisation of contours at the end of syntactically complete and incomplete utterances. In contrast, this experiment uses only syntactically complete utterances. This study also differs from others in that it uses only naturally occurring data. The availability of resynthesis techniques has allowed for at least partial control of the stimuli.

2. EXPERIMENT

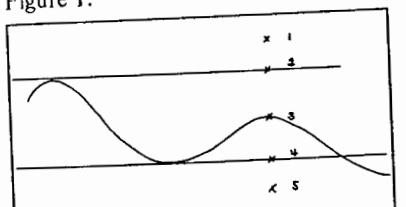
The experiment described here poses two questions:
(i) does a change in the height of the starting point influence the perception of finality?
(ii) how does such an effect relate to f0?

2.1. Method

Ten subjects were presented with five versions of each of 10 naturally-occurring utterances, 50 utterances in all, of which the final contours had been systematically manipulated. The utterances were all syntactically complete, and perceived (in a preliminary experiment) to be semantically unmarked for finality. Leaving the endpoint constant, the final falling contour of each one was assigned five different starting points, varying systematically in height. These five versions of each utterance were LPC-resynthesised. Listeners were asked to judge, on a four point scale, whether the speakers had finished or whether they had more to say.

2.2. Preparation of the stimuli

The ten sentences were digitised at 10 Khz, and normalised for amplitude. Pitch was extracted and the resulting f0 values were checked for octave leaps, and smoothed by hand. The peak of the last accented syllable in each sentence was manipulated to create five different versions of the f0 contour. In each case the f0 peak associated with the accent was adjusted to one of five different positions with relation to the preceding trough and penultimate peak, as illustrated in Figure 1.



Version 1 higher than penultimate peak
Version 2 identical to penultimate peak
Version 3 midway between preceding trough and penultimate peak
Version 4 identical to preceding trough
Version 5 lower than preceding trough

Figure 1. The five versions of each contour

In creating each version of the f0 contour, the f0 values preceding and following the manipulated peaks were adjusted to maintain as far as possible both microprosodic features and the correspondence between f0 and segments.

2.3. Procedure

The resynthesis of 10 sentences, each in five different versions, produced a set of 50 different stimuli. A stimulus sequence file was generated in which each stimulus was repeated five times, thus eliciting 250 responses from each subject. They were preceded by a test sequence of 10 stimuli which were ignored in the analysis. The subjects were asked to judge whether the speaker of each sentence was

- definitely going on,
- probably going on,
- had probably finished, or
- had definitely finished.

For the purpose of the analysis, these responses were converted into ordered data. The response 'definitely going on' became a '1', 'probably going on' became '2', 'probably finished' became '3', and 'definitely finished' became '4'. The lower the score, the less final the utterance was perceived to be.

3. ANALYSIS

An analysis of the results must aim to investigate the significance of all effects : sentences, subjects and f0 contours. It was difficult to do this formally because the responses were ordered categories 1 to 4. This kind of response violates the usual normality assumptions for classical ANOVA. However, the package 'PLUM' [4] was used to fit the appropriate ordered responses category model. The adequacy of fit of each 'treatment' (f0 version) is done by comparing differences in deviance between models with and without the treatment effect (goodness of fit) and comparing these differences with the appropriate chi square value. The results are shown in Table 1.

Table 1. Significance of sentence, subject and f0 effects.

	deviance difference	d.f.	sig. level
treatments (f0 versions)	(2640 - 2199) = 441	4	.01
subjects	(2438 - 2199) = 239	9	.01
sentences	(3113 - 2199) = 914	9	.01

As expected, all effects are highly sig-

nificant at the 1% level.

3.1. F0 effects

If we average the finality scores for all sentences and all subjects we see that there is a systematic gradient difference between the versions. It is clear that the lower the starting point of the final fall, the greater the degree of perceived finality. See Figure 2.

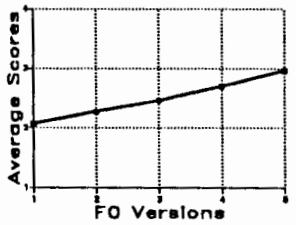


Figure 2. average scores of finality: This graph shows the average scores of finality for each version.

We know, however, that both the sentences and the subjects had a significant effect on the results. In order to see the effects of f0 manipulation without the influence of the significant between-sentence differences, the average scores across replicates for each version were ranked from 1 to 5. The least degree of perceived finality was ranked 1 and the highest score ranked 5. Tied scores were ranked equally.

For each of the versions 1 to 5, the average rank was calculated across all sentences and all subjects.

The results of this ranking are shown in Figure 3.

It can be seen that the different f0 versions have a marked and consistent effect on the perception of finality. A divergence from the predicted ranks is only possible in one direction, since no utterance could be perceived as less final than 'definitely going on' or as more final than 'definitely finished'. The deviations of the endpoints of the line are therefore to be expected. The greater deviance at the non-final end of the scale is also predictable, since we can assume that the presence of any falling contour which falls to the speaker's base line will indicate at least some degree of finality. There are minor differences in the degree of slope between any two consecutive points, but the overall trend is a straight line. The f0 versions therefore have a significant effect on the way subjects ranked utterances in terms of degrees of finality. The closest fit would be a straight line, and the results must therefore be interpreted as gradient rather than categorical.

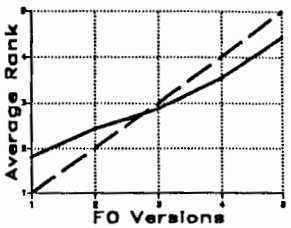


Figure 3. average ranks of finality: This graph shows the average ranks of finality (continuous line), against the predicted ranks (broken line).

4. DISCUSSION

4.1. The perception of finality

The first question posed by this experiment must be answered with yes: there is clear evidence that the starting point of a final fall influences the listener's perception of finality. There is no evidence to suggest that this perception is a categorical one. Whether we take the average scores of finality, or the average ranking of the different versions relative to one another, the result is a gradient. This suggests that finality is perceived in terms of **degrees** rather than in terms of categories, binary or otherwise.

4.2. Finality and f0

The answer to the second question - how does this relate to f0 - is not so clear. Is there such a distinction to be made as 'high' and 'low' fall, and if so, how are these to be defined?

It may be that the height of final falls is not perceived relative to any preceding syllables but to the speaker's overall range. All the versions of sentence 10, for example, were fairly low in the speaker's range. This might explain the tendency to judge this sentence as inherently more final than others.

Menn and Boyce [5] claim that the endpoint of a sentence-final falling contour can be regarded as constant in relation to the speaker's norm. This was assumed to be the case in the experiment described above. Nonetheless, a similar experiment in which the endpoint of each final contour was sytematically changed would complement the present study and perhaps throw light on results which are not accounted for here.

5. CONCLUSION

The tentative conclusions to be drawn from this study are as follows.

- The height of a fall given a constant low endpoint is perceived as a gradient. There is no evidence to support a categorisation of falls into high and low.
- The starting point of the final fall influences the degree of perceived finality. The lower the starting point of such a fall the more final the utterance is perceived to be. There are however other prosodic influences on the degree of perceived finality which cannot be accounted for here.
- Height seems to be perceived in relation to a speaker's norm and not in relation to the pitch of preceding syllables, accented or not.

REFERENCES

[1]BERKOVITS, R. (1984), "Duration and fundamental frequency in sentence-final intonation." *Journal of Phonetics* 12, 255-265

[2]FARNETANI, E. (1989), "Acoustic correlates of linguistic boundaries in Italian: A study on duration and fundamental frequency." *European Conference on Speech Communication and Technology. Proceedings Eurospeech '89 Paris* Vol 2, 332-335

[3]Lancaster/IBM Spoken English Corpus

[4]McCULLAGH. P. (1979), *PLUM An Interactive Computer Package for Analysing Ordinal Data* Working Paper, Department of Statistics, University of Chicago

[5]MENN, L., BOYCE, S. (1982), 'Fundamental frequency and discourse structure.' *Language and Speech* Vol 25 Part 4

STYLISED PROSODY IN TELEPHONE INFORMATION SERVICES: IMPLICATIONS FOR SYNTHESIS

J. House* and N. Youd**

*University College London, UK / Infovox AB, Solna, Sweden
**Logica Cambridge Ltd, Cambridge, UK

ABSTRACT

This paper studies the phonetic and phonological characteristics of stereotyped, often *stylised*, intonation patterns used by natural speakers to express routine procedural moves in telephone information services. It considers the appropriateness of such patterns in synthesised implementations within an automated information service.

1. INTRODUCTION

In an automated telephone information system, of the type under development in the SUNDIAL project, the role of the agent (A) in informing the caller (C) is taken over by a message planner + linguistic generator, with output speech provided by a synthesis-by-rule system (for British English, an adaptation of the INFOVOX text-to-speech system [1]).

Rules determining prosody are sensitive to a range of pragmatic and syntactic annotations, including labels for 'dialogue acts' (House & Youd [3]). Some such acts are primarily concerned with the *phatic management* of conversation, oriented less towards information transfer than to the interaction between (A) and (C). Observations of natural dialogue have verified that exchanges of this type are often associated with the use of *stylised* intonation.

Contexts for these acts recur on a routine basis, and will be equally applicable when (A)'s contribution to the dialogue is automatically generated. To maximise naturalness, the intonation used in synthesis should be modelled on the patterns found in natural speech. In practice, we must also ensure that the patterns used in synthesis are acceptable to the caller.

2. INFORMATION DIALOGUES

Over three hours of recorded dialogues between callers and airline agents were studied. Speakers in the (A) role were predominantly female, but no notable gender-based differences in prosodic form were observed. Patterns described below were based on an auditory analysis and transcription of (A)'s speech. Individual speakers sometimes favoured particular intonation patterns, but these were not speaker-dependent.

2.1 Dialogue Structure

In the Conversation Analysis (CA) tradition, summarised by Levinson [5], there are three major components of overall structure:

(i) opening section
(ii) topic-oriented slots
(iii) closing section

To this we add an optional *absence section*, particularly relevant during (ii), where the agent may need to ask the caller to hold the line, while information is being looked up.

2.2 Intonational Clichés

The opening, absence and closing sections -- those parts least concerned with information transfer -- represent the most routine contexts. The stereotyped responses to these routine situations regularly triggered the use of *intonational clichés* (Fonagy et al [2]), contours seemingly stored as holistic tunes. A subset of the clichés we observed

involved *stylisation*, conventionally regarded as a phonetic correlate of routine. Our definition of stylisation follows Johnson & Grice [4] in considering *monotone* to be a prerequisite. Although produced in conjunction with set phrases, the tunes themselves are considered to be independent of any specific text.

3. OBSERVATIONS

3.1 Phonetic notation

Our notation of the examples uses a three-tone system: pitch values are High ***H***, Mid ***M*** or Low ***L*** relative to our assessment of the speaker's current range. Symbols precede the syllable to which they apply in the text. The symbol -- denotes rightward spreading of the preceding tone, as distinct from a simple interpolation between values; a final ^ denotes an upglide at the end of the domain governed by the preceding tone; and an initial ^^ indicates high register. This simplified notation is adequate for the cliché tunes, where downstep and declination do not apply.

3.2 Openings

These are concerned with identifications, greetings and with eliciting the nature of (C)'s task. In our dialogues, identifications were always present, and usually began (A)'s opening move, which could also optionally include a greeting. Intonational clichés were found on all components; true stylisation occurred most readily on the identification component(s), less often on the greeting component.

The majority of openings in our study could be analysed as realisations of a very limited set of tunes. Two tunes, (i) */LHM/* ('calling contour') and (ii) */HLM/*, accounted for a high proportion, if one allows a gradient analysis of (i) which can accommodate both fully stylised and 'less' stylised variants (4.1). Examples:

Tune (i), */LHM/*, stylised:

(1) ***L***--Flight infor ***H***--ma ***M***--tion
(2) ***L***--British ***H***--Air ***M***--ways
(3) ***L***--Good after***H***--no***M***--on
(4) ***L***--Can I ***H***--help ***M***--you

Tone sustention was a regular feature of this variant; the ***H*** - ***M*** interval was typically around a minor 3rd.

Tune (i), */LHM/*, with upglide:

(5) ***L***--Flight infor***H***--ma***M***^tion
(6) ***L***-- ***H***^ ***M***^

The ***L*** tone was normally spread, but an upglide could occur on ***H*** and/or ***M***. Usually turn-final, this rising variant might be analysed as an overlay on the conventional stylised form, indicating a turn-giving cue.

Tune (ii), */HLM/*:

(7) ***H***Flight ***L***--informa***M***tion
(8) ***H***--British ***L***Air ***M***ways
(9) ***H***Good ***L***mor ***M***ning
(10) ***H***--Can I ***L***help ***M***you

Spreading is only shown for this cliché tune where ***H*** or ***L*** continues over more than one syllable. The tune, phonetically similar to a *fall-rise* nuclear tone, or to a *high (pre)head + low rise* (see 4.1), lacked any genuinely stylised, sustained monotone on the final ***M*** syllable.

Other recurrent tunes included:

(11) ***H***Flight ***M***--information (*/HM/*)
(12) ***L***--Flight infor ***M***--mation (*/LM/*)

and variants of these were also found with final upglide.

Each opening component could act as an independent domain for one of the tunes, while a succession of components typically, but not invariably, involved tune repetition. Components could also be combined into arguably composite versions of tunes (i) and (ii); in both cases this was achieved principally by extending the */L/* tone:

Tune (i), */LHM/*, composite:

(13) ***L***--British Airways flight infor ***H***--ma ***M***--tion

Tune (ii), */HLM/*, composite:

(14) ***H***Good ***L***--morning British Airways ***H***flight ***L***--informa ***M***tion

With rare exceptions, the final pitch in these moves was at a mid-level, or rising from mid to high. A wide range and relatively high register were common.

3.3 Closing sections

These may be divided into two components: *preclosings*, in which mutual intention to close is established; and *terminal exchanges*, which accomplish 'signing off'. Typically, (C) began the preclosing move, using downstep + low

fall (on e.g. 'Thank you very much') to convey task or dialogue completion. (A)'s response was frequently produced as a prosodic cliché involving a ***HL*** or ***ML*** drop to a very low pitch termination:

(15) ***H/M***You're ***L***--welcome

The final exchanges regularly used variants on the calling contour:

(16) ***H***--By ***M***--ye
(17) ***H***--Bye ***M***--bye
(18) ***L***Bye ***H***--By ***M***--ye

By contrast with the preclosings, final low pitch was apparently avoided.

3.3 Absence sections

Absences in our data were *proposed* by (A), and *accepted* by (C). A wide range of prosodic possibilities included variations on the calling contour:

(19) ***H***--Hold ***M***--on
(20) ***L***Hold ***H***--o***M***--on
(21) ^^***L***--Hold ***H***--on a ***M***--momemt
(22) ***L***--Would you ***H***^hold ***M***^please
(23) ^^***L***--Hold the ***H***--li ***M***--ine,
L--I'll just ***H***--che ***M***--eck

Another possible stylisation was:

(24) ***M***--Can you ***H***--^hold the line please

Some speakers used different idioms, such as a downstepped contour, in the proposal position; others preferred a non-stylised cliché, a version of ***/HLM/***:

(25) ***H***Hold ***L***--on a mo ***M***ment
(26) ***H***Let me just ***L***--check that for ***M***you

To indicate return from an absence, (A) often made use of a calling contour, with or without final upglide. Register tended to be high, especially if a second reconnection attempt was needed:

(27) ^^***L***He ***H***--llo ***M***--(^)o

(C)'s response often matched this. One major function of the calling contour seems to be *line checking*, ascertaining whether or not the interlocutor is present.

4. PHONOLOGICAL STATUS

In formal terms we have characterised the intonational clichés as sequences of the tones ***H***,***L*** and ***M***, extending over a whole intonational phrase. Functionally, the phatic role of the cliché utterances overrides any notions of information focus. A *holistic* abstract representation of these tunes would seem to be better motivated than, say, a nuclear tone-based analysis, in which intonation groups are made up of component parts such as *prehead*, *head*, *nucleus*, and *tail*. Such an analysis is weak on both formal and functional grounds. Examples of the two most popular clichés, ***/LHM/*** and ***/HLM/***, illustrate the difficulties.

4.1 Nuclear tone?

Since the ***H*** in the ***/LHM/*** tune is always aligned with a metrically prominent syllable, we would have to propose a stylised ***HM*** nuclear tone, with optional preceding low head (***L***). A problem arises with upglide ^ variants like (5) and (6): are these to be regarded as variants of the ***HM*** tone, or perhaps of the fall-rise? The latter analysis would maintain a categorical stylised/ non-stylised distinction, while the former acknowledges that versions with and without upglide may be used virtually interchangeably in comparable contexts.

In a nuclear tone framework, the ***/HLM/*** tune is ambiguous: in (7) it is consistent with a fall-rise on the first syllable, but in (8-10) we would have to posit a low rise on the ***L*** syllable, preceded by high head or high prehead. In so doing we would lose sight of the similarity between the tunes, involving the same pitch sequences but with different mappings over the text.

4.2 Accentual function?

Nuclear accent conventionally signals information focus and coincides with the metrically most prominent syllable. In stereotyped phatic utterances the location of this prominence may be variable (17-18; 19-20). The phrase 'flight information' appears to be ambivalent between a reading as a compound with early stress (7) and a phrasal reading with late stress (1). In practice, these variations in prominence appear to be tune-dependent; versions of ***/(L)HM/*** such as:

?(28) ***H***--Flight ***M***--information

or of ***/HLM/*** such as:

?(29) ***H***--Flight in for ***L***ma ***M***tion

were not favoured.

4.3 A holistic analysis?

The implication must be that there is a trade-off between preferred prominence relations and the requirements of the cliché tune. For instance, (1) may be preferred over (28) because of a strong pressure to include the ***L*** component in ***/LHM/*** where there is room to do so. Conversely, the ***/HLM/*** tune can accommodate both (7) and (29) equally, but (7) wins because it respects the compound stress pattern. Overtly contrastive possibilities like:

?(30) ***H***Bri ***L***--tish Air ***M***ways

are also avoided. Metrically prominent syllables will always be at a turning-point in the contour, but the precise mapping of tune to text may be flexible.

5 SYNTHESIS: IMPLICATIONS

On the assumption that any automated dialogue system will have a structure similar to that outlined above, we must decide on an intonation for the synthesised phatic utterances. In synthesising informative utterances, the desirability of exploiting prosody to clarify information structure and communicative function has been long recognised. In phatic utterances, where the inter-personal relationship is foregrounded, we must consider the most appropriate prosodic form. It may be right to question whether what is highly acceptable in natural speech will be equally appropriate in a context where (C) knows that (A) is not a real person; will (C) accept prosodic clichés, and particularly stylisations, as markers of stereotype and routine when they emanate from an inanimate source? As part of a programme of acceptability testing, stylised, 'less' stylised and non-stylised cliché variants of the phatic utterances are being synthesised (by hand, initially) and integrated into our automatically generated dialogues.

If any of the cliché tunes are indeed deemed suitable for dialogue synthesis, then they must be implemented by rule, and an abstract representation incorporated into the phonological model of prosody in the rule system. We propose a holistic representation, with realisation rules bypassing the 'normal' nuclear tone assignment, but sensitive to metrical prominence. Candidate utterances will be identified by the markers passed on at the interface between linguistic generator and synthesis system.

ACKNOWLEDGEMENT

This research was supported by ESPRIT and by Infovox AB, Sweden.

REFERENCES

[1] CARLSON, R & GRANSTROM, B (1986),"Linguistic processing in the KTH multi-lingual text-to-speech system", *Proc. ICASSP 86, Tokyo, 2403-2406*

[2] FONAGY, I, BERARD, E & FONAGY, J (1984), "Clichés mélodiques", *Folia Linguistica 17*, 153-185.

[3] HOUSE, J & YOUD, N (1990), "Contextually appropriate intonation in speech synthesis", *Proc. ESCA Workshop on Speech Synthesis, 185-188.*

[4] JOHNSON, M & GRICE, M (1990), "The phonological status of stylised intonation contours", *Speech, Hearing and Language 4, work in progress*, University College London.

[5] LEVINSON, S (1983), *Pragmatics*, Cambridge: CUP.

LINGUISTIC MECHANISMS OF WORD ACCENTUAL PROMINENCE IN THE TEXT

Tatiana Skorikova

Russian Language Department, MIIT
Moscow, USSR

ABSTRACT

A new approach in intonology - the introduction of lexics into intonological analysis is put forward in this paper. We try to show that the relationship between lexical meaning and accentuation is based on certain linguistic factors. The mechanism of this relationship can be described if we take into account the text-forming potential of word accentuation.

INTRODUCTION

Following T.M. Nikolayeva we distinguish two functionally different types of word prosodical prominence in the utterance:
1) the neutral sentence stress(SS) which refers to the plan of expression and serves as a means of syntagma phonetic organization and intonational segmentation of speech;
2) the sentence accent (SA) which is related to the semantic aspect of the utterance and is determined by the context and communicative intention of the speaker [4, p. 486-487].
Our research deals with SA or accentual prominence (AP) (the term was proposed by T.M. Nikolayeva [5]) as " a textual communicative phenomenon" [5, p. 9].
In the recent years scientists accepting the idea of functional difference between these types of accent are inclined to treat AP as a multi-aspect object of study. Nowadays the attention of Russian intonologists is concentrated on such problems as the description of linguistic factors as regards the SS and SA [11]; the accentual phrase structure [3]; the interaction between words semantics and AP [6, 10] on the one hand, and AP and text organisation [1, 8], on the other hand.
Taking into account the latest results in the investigation of AP's functions in Russian spontaneous speech we state that there exist certain semantic regularities in the accentuation of various lexical classes within a given type of the text. We try to establish these regularities by examining modern Russian Scientific Discourse (RSD) the linguistic properties of which have recently been described in the fundamental work by O.A. Lapteva and others [2].
We would like to put forward an idea that AP helping to reveal the speaker's intentions in the communicative act plays an essential role in text-formation process as its one of the main pragmatic components.
To give prove to the proposed point of view we consider two main questions:
1. What are the linguistic factors of AP realisation in RSD?
2. Can the degree of word accentuation potential in discourse be evaluated objectively?
The vollowing material served as the basis for our research:
- spontaneous uttered texts of a scientific character (lectures, reports, discussions);
- summarized phonetic transcription of text fragments;
- listener's reactions to different types of accentual patterns. The experiment on perception shows that AP in RSD is perceptually marked for the audience and can be considered as a relevant linguistic feature of the text prosodical structure.

RESULTS

1. Linguistic Factors of AP Realization in RSD

The context analysis of accentually marked elements in the utterance brings us to the conclussion that word accentuation in RSD is regulated by a number of closely interrelated factors such as: thematic and situational text parameters; lexical meaning of the word and its syntactic position; different contexual loads as well as pragmatic orientation of the utterance.
Let's consider the mentioned factors in detail:
1. Strong AP in RSD may be laid on the so-called keywords of the message:(terms proper names, titles, etc.) conveying information about the theme and communicative act.
2. Regularly accentually prominent become groups of words with appraisal, qualitative and attitudinal semes in their meanings.
3. As a rule, thematic and other lexical elements in the utterance are exphasized within given syntactic contexts (connections may be expressed in the following ways: x, y ...; i X, i Y; X ili y, Y i Y; ne tol'ko X, no i Y; ne X, a Y; kak X, tak i Y etc.).
4. With the help of AP the speaker often singles out and determines the boundaries of speech segments in RSD thus fascilitating auditive perception of a spontaneous monologue. In this connection, certain types of functional and auxiallary words at the beginning of syntagmas and utterances as well as initial components of nominative word-combinations and attributive constructions may acquire a strong AP.
5. Accentually marked in RSD are usually words connected by theme/rheme relations (AP marks either theme or rheme R — T; R — T).
6. Semantically interrelated lexemes may also be marked with strong phrase accent in the context. AP here performs its deictic function in RSD exposing more explicitly the semantic ties of the text components.
7. Text lexical signals facilitating the orientation in the discourse and conveying different pragmatic loads (adresation, motivations, qualifications, attitudes etc.) are usually accentually emphasized and serve as prosodi-

cal markers for the listeners in RSD.

2. Evaluation of Word and a Group of Words Accentuation Potential in RSD

Examing different parts of speech accentuation in RSD we came to a conclusion that word accentual potential in discourse can be evaluated objectively. For this purpose we introduce a special criterion - the relative accentuation index (i) showing the ratio between the number of cases when the word (group of words) is found in the accent position and the number of cases of words non-accent positions in the text.
Accentuation of the parts of speech in RSD may be presented as a following scale:

	i
Adjectives	0,335
Adverbs	0,320
Predicatives	0,289
Nouns	0,278
Verbs	0,227
Numerals	0,217
Parentheses	0,173
Pronouns	0,144
Particles	0,135
Conjunctions	0,122
Prepositions	0,053

As seen from the scale, adjectives, adverbs and predicatives - words with wide qualificative semantics - top the list as regards the relative accentuation index. Accentuation potential of auxilliary words is lower if compared with meaningful words.
In this way we qualified the AP indexes of particular meanings of the most frequently used in RSD words as well as accentuation indexes of some lexemes, lexico-semantic groups* and wide semantic zones of the text**.
It is necessary to note that AP potential of different lexical groups, seperate lexemes and their meanings varies greatly within one particular part of speech.
The analysis of accentual structure and semantics of attributive word-combination (adjective+substantive) in RSD /9/ testifies to this fact.
The latest results of our research can be presented in the following way:
1. If we consider nouns, we'll see that the greater accentual load is laid on the lexical units characterizing an object or a person from different points of view (i = 1,30) as well as on those denoting qualities, properties and speaker's attitudes (i = 1,27).
2. Among the adjectives high i-valued are lexemes with opposition/comparison semes (i = 2,7) as well as lexemes denoting the highest degree of some property (i=2,2) and attitudes (i=1,1).
3.We can observe different AP abilities of words within separate lexical classes. For example, among Russian pronouns declined like adjectives high accentually marked is lexeme drugoi (i=1,55), medium AP-index characterizes lexemes odin (i=0,83) and kazdyj (0,59). The majority of pronouns of this type are within the range of low accentuation indexes: nas (0,44), nikakoj (0,41), ves' (0,40), tot (0,39), etc.
4. As regards the total accentual loads in RSD qualificative semantic zone possesses the highest degree of AP (i=0,733) and nominative one - the lowest (i=0,427).

* Under the term of "lexico-semantic group" we mean "any semantic class of words (lexemes) characterized by at least one lexical paradigmatic seme in common" /12, p. 110/.

** Considering the text semantic zones we follow N.Yu.Schwedova stating that according to language functions we can single out in any text such semantic areas as nomination, communication proper, qualifications and attitudes, connections and correlations /7/.

CONCLUSIONS

1.The results obtained give sufficient grounds to state that such parameter as relative index of accentuation should be listed as one of linguistic characteristics of a word when it is regarded as a discourse unit.
2.As soon as we can measure AP potential of words we can take a new approach towards classification of lexics based on the relative accentual values of lexemes and their text loads. This research will contribute to composing "a dictionary of prosodical potentiality of lexemes (their prominence and phonation possibility)"/4, p.490/.
3.The main thesis of our work comes as follows: all the pragmatic sense components (denoting qualifications, speaker's attitudes, modality, various text orientations) are usually prosodically marked in the form of AP of some lexical elements in RSD.

REFERENCES

/1/Drozdova,T.J.(1988), "The Key-words of the text and their Prosodical Properties", Cand.Diss.,Leningrad (in Russ.).
/2/Modern Russian Scientific Discourse(1985).Vol.1, Krasnojarsk (in Russ.).
/3/Nadeina,T. (1985), "Accentual Structure of the Utterance in Russian", Cand. Diss., Moscow (in Russ.).
/4/Nikolayeva,T. (1987), "The Intonology of the 80-es", Proc. XIth in ICPhS, vol.2. Tallinn, 486-491.
/5/Nikolayeva,T.M. (1982), "Semantics of Accentual Prominence", Moscow (in Russ.).
/6/Pavlova,A.V. (1987), "Accentual Phrase Structure in its Correlation with Lexical Semantics". Cand.Diss., Leningrad (in Russ.).
/7/Schwedova,N.Yu. (1985), "One of the Possible Ways of Building up Russian Functional Grammar", The Problems of Functional Grammar, Moscow, 30-37 (in Russ.).
/8/Skorikova,T.P. (1987), "Functions of Accent Prominence in Speech", Proc.XIth ICPhS,vol.4, Tallinn,279-282.
/9/Skorikova,T.P.(1982), "Functional Capacity of the Intonation Pattern of Word-group in Speech-flow", Cand. Diss., Moscow (in Russ.).
/10/Skorikova,T.P. (1985), "The Accentual and Semantic Capacity of Adjectives in Russian Scientific Discourse", Scientific Literature, Moscow, 118-137 (in Russ.).
/11/Svetozarova,N.D. (1987), "Linguistic Factors in Sentence Stress", Proc. XIth ICPhS, vol.6,Tallinn,110-113.
/12/Vasiljew,L.M. (1971), "The Theory of Semantic Fields". Language Proc., N 5, 105-113 (in Russ.).

L'INTONATION DU POINT DE VUE DE LA PHONOLOGIE

Dr. Galina Ivanova-Loukianova

Université des langues étrangère M.Thorez
Moscou, URSS

System interdependent between information and syntax make it possible to determine three intonems, the sense-making of the language which serve to differentiate the meanings of phrases with similar lexical-grammar components, in particular, the intonem of the rising tone, the intonem of the falling tone and the intonem of the level tone.

L'analyse de l'intonation du point de vue phonologique demande qu'une distinction de principe des unités de langage soit prise au niveau segmentaire et supersegmentaire. L'intonation, à l'opposé de la phonème, est multifonctionnelle, c'est pourquoi, il faut y distinguer les fonctions liées à la différenciation du sens et celles qui ne le sont pas (fonctions expressives, stylistiques, modales, esthétiques et autres). Seule cette intonation "nue", selon l'expression de B.Tomachevski, peut être l'objet d'une étude phonologique. L'intonème, pareil à la phonème, se présentera comme une unité de langue d'importance fonctionnelle.

Puisque la phonologie ne s'occupe que des fonctions différentielles du sens, il est nécessaire de décider quels sens peuvent être différenciés par l'intonation et lesquels sont directement liés à la différenciation du sens.

L'opposition traditionnelle des intonations ascendantes et descendantes dans l'interrogation et l'affirmation nous permet de considérer que l'opposition phonologique du mouvement du ton est liée à la transmission des différentes significations syntaxiques. L'intonation, comme moyen d'expression des significations syntaxiques,

transmet aussi des significations qui sont propres à toutes les formations syntaxiques. C'est la signification de dépendance-indépendance(ou de subordination - autonomie) et d'achevé - d'inachevé. L'intonation du type ascendant nous donne le sens de la dépendance et de l'inachevé, et celle du type descendant, le sens d'indépendance et d'achevé. Le type d'intonation d'une phrase dépend de la concordance des significations, prises par paires, avec le sens de la phrase, en tant que formation syntaxique. Si la signification syntaxique de la phrase correspond avec les significations du ton ascendant ou descendant, alors l'intonation de cette phrase ne peut avoir de variantes: elle doit être présentée dans un mouvement tonal ascendant ou descendant. S'il n'y a pas de correspondance, alors il est possible de varier les intonations, c'est-à-dire, que la phrase peut être présentée avec une intonation ascendante, aussi bien que descendante et même dans un ton monocorde. La corrélation de la syntaxe et de l'intonation permet de dégager trois intonèmes, unités différenciatrices, qui distinguent le sens des phrases à composition lexico-grammaticale identique: intonème ascend. (IA), intonème descend.(ID) et intonème plate (IP).. L'intonème se réalise sur un segment de chaîne du langage par une phrase égale (=syntagme) avec tout un complexe de moyens supersegmentaires: modification du ton, frontières de la segmentation et du centre d'intonation. Le signe distinctif de l'intonème c'est le mouvement du ton dans la phrase. Donc trois phrases, ayant la même composition lexico-grammaticale, mais des intonations différentes, expriment trois sens différents. Par ex.: Il neige?-Il neige. Il neige,/on ne voit pas la fin de l'hiver/. Ici, l'intonation est le seul moyen de différencier le sens. Ceci nous donne le droit de parler de trois intonèmes.

Les intonèmes se différencient par la forme(haut, bas, plat), le sens(IA-sens de la dépendance et de l'inachevé; IP-sens de l'indépendance et de l'inachevé; ID-sens de l'indépendance et de l'achevé) et les fonctions(distinguent les ty-

pes de propositions communicatives et le caractère des relations syntaxiques.

Dans la langue parlée,les intonèmes sont réalisées par des modifications ascendantes, descendantes et plates(le système du CI de Bryjgounova est utilisé). Les nombreuses réalisations de l'intonème résultent de l'influence des particularités du lexique, de la morphologie et de la syntaxe sur le caractère de l'intonation,qui en plus du sens, rendent les nuances de l'expression,du style, de la modalité.La dépendance lexico-grammaticale sur le choix de l'intonème peut être plus ou moins forte. Phonologiquement cette influence peut être considérée comme une position : faible, si le choix dépend de la composition lexico-grammaticale(phrases avec différents moyens exprimant la question, la motivation, le recours); forte, si le choix de l'intonème n'est pas déterminé par sa position, c'est-à-dire, qu'une même phrase,grâce à l'intonation, peut exprimer l'interrogation, l'affirmation et l'inachevé d'une action. Les possibilités de différenciation du sens ne se manifestent que dans des conditions de position forte, lorsque dans une même phrase, il est possible de réaliser les 3 intonèmes avec trois sens différents. Pour que la position forte puisse se réaliser,il est nécessaire que les intonèmes répondent à deux positions:

I°- Toutes les trois intonèmes sont en opposition dans une phrase neutre par le style et à un seul composant lexico-grammatical. (Allons au bois?- Allons au bois.- Allons au bois....) et,

2°- L'intonation est le seul moyen de différencier le sens, autrement dit, quand le composant lexico-grammatical n'influence pas le choix de l'intonème ou lorsqu'il n'est pas motivé par la position.Dans la position forte l'intonème se présente ainsi: ID-CI-1; IA-CI-3; IP- ton monocorde. Cette intonation est de style neutre, mais elle peut être utilisée stylistiquement(intonation expressive). C'est le CI-4,6,2 pour l'IA; CI-2,5 pour l'ID. Ici,les fonctions non-grammaticales de l'intonation vont se manifester. Selon la position(position faible),la dépendance du

choix a deux aspects.

I.L'intonème est représentée par sa variation quand le choix de l'unité est marqué par la position,c'est-à-dire, quand les moyens lexico-grammaticaux expriment le sens principal de la phrase et l'intonation n'est que complémentaire. Par ex.:Pourquoi..? -Si... Ce que...-Bonjour,... Véra apportez-moi.(A comparer avec-Véra a apporté). Plus le sens est exprimé par les moyens lexico-grammaticaux, moins l'intonation a de rôle à jouer (Selon A.N.Grozdev) et ceci peut aller jusqu'à la neutralité totale. (Quelle heure est-il?)

L'intonation des propositions du type communicatif (question, demande...) se conforme à certaines règles dans chaque langue les siennes. Ici, le rôle différenciateur de l'intonème est affaibli,car des moyens dexpression du sens de la phrase,autres que ceux de l'intonation, y sont employés.

II. L'intonème est présentée par sa variante. Dans cette position, les intonèmes ne s'opposent pas.C'est le cas lorsqu'est employé le type "interdit" d'intonème, c'est-à-dire, quand l'intonation n'est pas employée dans sa signification première lorsque prévalent les relations syntaxiques. Par ex. l'intonation de la fin d'une phrase narrative peut être dite avec le CI-3,4,6. La neutralisation est aussi possible : Tu as oublié les gants?- Tu as oublié les gants.- Tu as oublié les gants (et le parapluie aussi).Ces intonations portent toujours une coloration stylistique expressive. Ainsi, les particularités de la réalisation des intonèmes dans le texte servent de base à la détermination du rôle stylistique de l'intonation.

La résolution de la question des intonèmes permet aussi de juger des caractéristiques du rythme et de l'intonation d'un auteur (ou d'un texte).

THE INTONATION OF INTERROGATION IN TWO VARIETIES OF SICILIAN ITALIAN

Martine Grice

Department of Phonetics and Linguistics, University College London

ABSTRACT

A study of intonation contours in Palermo and Catania Italian shows that surface dissimilarities between interrogative forms may be simply due to timing differences and the existence of a non-functional tone.

1 INTRODUCTION

In Italian, intonation plays a major role in the communication of interrogation. In the case of yes-no (polar) questions, there are no interacting morphological or syntactic cues; it is solely by virtue of their intonation contours that they are perceived as questions rather than as any other illocutionary act. Nonetheless, the tonal pattern which marks this function in different accents of Italian is not uniform. In the two Sicilian varieties examined here, those spoken in Palermo and Catania, it is marked with a rise-fall and a rise respectively.

Use is made here of recordings carried out as part of a more extensive study. These include spontaneous speech, and questions and statements which were read aloud. The latter had accompanying contexts clearly indicating the desired focus structure. An auditory and instrumental examination of these data for five speakers of each variety (all speaking regional Italian rather than a dialect) provides the basis for discussion.

2 THE TONAL FORM OF INTERROGATION

According to Crystal [3] (210-11) most intonologists view the final direction of pitch movement as paramount in the classification of tones; a rise-fall is therefore considered to be a variant of a fall rather than a rise. In Bolinger's pitch accent analysis, a rise-fall lies, along with falls, within the Accent A category, except where the fall has a shallow gradient.

If it is the terminal pitch direction which is crucial to the marking of interrogation, then the two varieties of Sicilian Italian, Palermo Italian (PI) and Catania Italian (CI), make use of entirely different intonation patterns: the terminal pitch movement is falling in the former and rising in the latter. However, the two are mutually comprehensible as far as interrogative function is concerned. It is therefore of interest to examine whether there is a common element in these two contours.

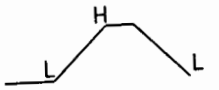
It could be argued that it is the *rise* which signals interrogation - intonation group-finally in CI and before the final fall in PI. Although this is tenable in CI, the situation is not clear-cut in PI, where non-final clauses are distinguished from polar questions by intonational means, even though both types of contour contain a pre-terminal rise. Alternatively, interrogation could be signalled by the presence of *high* pitch, manifesting itself as a high terminal in CI and as a boosted peak in PI; but there is no simple correlation here either, as boosted peaks are not confined to interrogatives. They occur on non-final clauses and exclamations, both of which are intonationally (whilst not necessarily syntactically) distinct from polar questions. A consideration of these other forms is important in clarifying the tonal form of yes-no questions.

3 TIMING IN PI CONTOURS

A closer look at the interrogative in PI suggests that it is the *timing* of the rise that distinguishes polar questions from other sentence types. In the former, the rise begins and ends on the accented syllable; in fact it is generally accomplished during the vocalic portion of the syllable (which is also the part with highest sonority (cf. Silverman and Pierrehumbert [5]).

Figure 1 illustrates the F0 contour of a PI yes-no question with narrow focus on the last lexical item: "Glielo porta domani?" (Will s/he bring it tomorrow?). As is most common in Italian, the penultimate syllable is stressed. The final part of the contour (in the region of the final accented syllable and beyond) may be described as the following sequence of tones: LHL. This LHL sequence occurs in the non-final clauses and exclamations mentioned above.

In the schemata presented below, an initial L tone is taken to occur at the point at which the F0 gradient becomes positive (the beginning of the rise), the H tone the point at which a zero gradient is reached; this could be a turning point (peak) or the beginning of a plateau. The final L is the low point reached at the end of the utterance. Schematically:



In all cases, the final L occurs at the end of the intonation unit. We shall therefore concentrate on the timing of the LH sequence.

In the polar question, L occurs early and H late in the accented vowel. The contour may be schematised thus:

a)

In non-final clauses, L occurs between one and two syllables before the accented syllable, H occurs late in the vowel, as follows:

b)

In certain types of exclamation, L occurs before the accented syllable and H occurs early in the accented vowel, as follows:

c)

The existence of these contours makes it difficult to account for the distinctive timing in terms of association rules, which allow for the association of one tone with a metrically strong (accented) syllable. Optionally, another tone may lead or trail. An association of the type LH* would have to be used for cases (b) and (c) above and L*H for case (a).

Where the accented syllable is word and utterance-final, the rising movement is timed in the same way as in penultimately stressed words but the fall is not completed, ie. does not terminate low; the rise(-plateau)-fall in Figure 1 has as its equivalent the rise(-plateau)-slump (Cruttenden's terminology [2]) in **Figure 2**."Gliel'hai detto tu?" (Did you say it?) The timing of the LH part of the contour is such that both L and H occur on the vocalic part of the syllable; it can be argued that there *is* a final L if it is considered to be undershot. It appears, then, that the final drop to low does not play an important role in signalling interrogation.

4 TIMING IN POLAR QUESTIONS IN PI AND CI

Figure 3 illustrates the CI contour of the yes-no question "Glielo porta domani?" - equivalent to Figure 1 in PI.

Two alternative hypotheses might account for the mutual comprehensibility between PI and CI interrogative contours, both assuming that LH is the crucial sequence.

The first relies on the concept of *alignment* which has been explored in various ways in a number of theoretical frameworks. The work of Bruce and Garding [1] accounts for dialectal variation in terms of whether a F0 peak is early or late in relation to the accented syllable. Ladd [4] formulates this in terms of the binary feature [±delayed peak]. The notion of precise alignment in the above mentioned work is adopted here. However, no constraints are placed on the

number of tones aligned with any given unit.

The second hypothesis makes use of the concept of *association* where one tone only is associated with a unit in the syllabic tier (cf Pierrehumbert [5]). In each case, the low end-point in PI is accounted for differently:

Hypothesis A: In PI, L is aligned with the beginning and H with the end of the accented syllable. Once the H target is reached, the pitch falls to a contextually-determined L tone (this low is accompanied by low amplitude and reduced spectral definition). The low F0 is realised when additional segmental material follows the accented syllable with which the LH is aligned, as is more often than not the case in Italian.

In CI, the L is aligned with the end of the accented syllable and the H with the end of the phrase; alignment is consistently later.

Hypothesis B: For PI, the contour is analysed as L*H L%. However, both the L* and the H fall on this syllable. This is due to tonal repulsion of the H (cf. Silverman and Pierrehumbert [6]), by an accent-specific, *contour independent obligatory L%* boundary tone.

For CI, the boundary tone is not obligatorily low; it may be high or low. The contour is analysed as L* H% or L*H H% although no independent evidence in favour of the latter tonal form with a bitonal pitch accent has been found.

In opposition to Hypothesis B is the existence of the contour in **Figure 4** of the PI question "Ma e' andato al cinema?" (But did he go to the cinema?) where narrow focus is underlined. Both L and H fall on the accented syllable "da". The rest of the contour consists of a high plateau followed by HL. Although the L% in the contour in Figure 1 could be seen as shifting the position of the H back onto the stressed syllable, there is no reason for this to be the case here.

Another problem with Hypothesis B is that, according to Silverman and Pierrehumbert, tonal repulsion occurs in order to allow the tones to be fully produced in the time available. In the example illustrated in Figure 2, where no segmental material follows the accented syllable, the final L is not fully produced. Furthermore, a comparison of a number of contours by a number of speakers shows that there is no noticeable difference between timing of the LH in penultimate stress contexts (as in Figure 1) and in final stress contexts (as in Figure 2). A theory of tonal repulsion would predict that the proximity of the tones in the latter context would shift the H tone even further back. This is not the case.

Independent evidence in support of Hypothesis A may be found by considering the tonal timing of the other LHL contours in PI. An account of this alignment requires four alignment points, all of which can be used distinctively:

```
>     V-    V+    ]
```

where, in relation to the segmental tier, > is prior to the accented syllable (equivalent to a leading tone); within the accented syllable, V- is early and V+ is late;] is at the end of the utterance.

The yes-no question (a) is aligned thus:

```
L     H     L
V-    V+    ]
```

the non-final group in (b):

```
L     H     L
>     V+    ]
```

and the exclamation in (c):

```
L     H     L
>     V-    ]
```

The status of] in PI is such that it does not align with tones which carry a functional load. There is no choice on the part of the speaker, as there is a L tone in this position in all utterance types. It is perhaps due to the linguistic insignificance of L in this position that it undergoes undershoot when there is insufficient segmental material for the movement to L to be achieved (ie. it is too close to V+). There is certainly no evidence of the H on V+ being lowered for this reason.

This is not the case in CI where either H or L may be aligned with], a linguistic choice depending on the illocutionary act.

It may therefore be concluded that in PI only a L tone may be aligned with a], whereas in CI a tone of the speaker's choice is aligned with it. This suggests that in the former case the boundary tone is phonetic (ie. contextually determined) and in the latter it is phonological (implying a linguistic choice). In this case the alignment of the CI polar question:

```
L     H
V+    ]
```


can be seen to be equivalent to the alignment of that in PI:

```
L     H     L
V-    V+    ]
```

where both have the LH aligned with the last two meaning-bearing alignment points of the utterance.

5 CONCLUSIONS

What appears to be a difference in the intonation contours of interrogatives in Palermo and Catania Italian can be analysed as a similarity of tonal form as follows: the rise or LH sequence which is the marker of interrogation is aligned in both cases with the last two meaning-bearing alignment points in an intonation group. The rise-fall in PI therefore consists of a LH sequence followed by a contextually determined L which has no functional load.

The existence of functional versus non-functional tones requires further corroboration, some of which may be gleaned from a study of other languages which use the rise-fall as marker of interrogation. This may provide an explanation for why these languages do not fit in with the universal tendency for a terminal rise to mark interrogation.

In addition, more detailed phonetic analysis needs to be performed on the PI interrogative contours and a more systematic comparison made with other non-interrogative contours. In particular, the alignment of each tone needs to be carefully controlled. To this end, a perceptual study is planned whereby the alignment of LHL contours is systematically varied.

6 REFERENCES

[1]Bruce, G and E Garding, 1978, A Prosodic Typology for Swedish Dialects, in Garding et al,*Nordic Prosody*, 219-28.

[2]Cruttenden, A, 1986,*Intonation*, CUP.

[3]Crystal, D, 1979, *Prosodic Systems and Intonation in English*, CUP.

[4]Ladd, DR,1983,Phonological Features of Intonational Peaks, *Language*, 59,4.

[5]Pierrehumbert, JB, 1980, The Phonology and Phonetics of English Intonation, MIT dissertation.

[6]Silverman,KE & JPierrehumbert, 1990 The timing of prenuclear high accents in English, in Kingston and Beckman, *Papers in Laboratory Phonology:*, CUP.

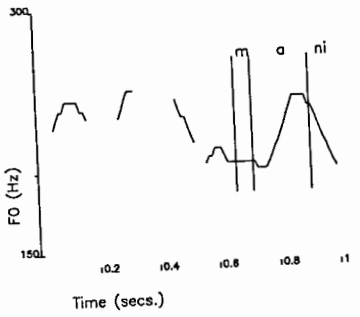


Fig.1 : "Glielo porta domani?" (PI)

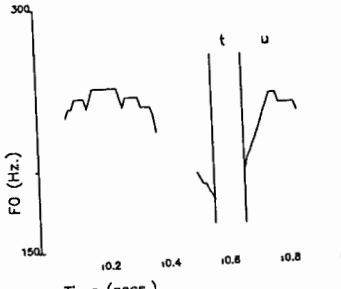


Fig.2 : "Gliel'hai detto tu?" (PI)

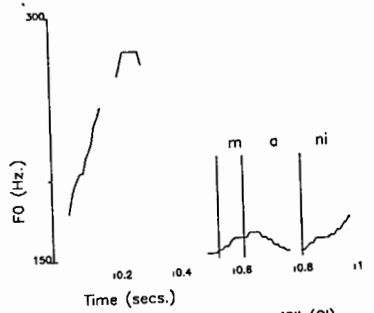


Fig. 3 : "Glielo porta domani?" (CI)

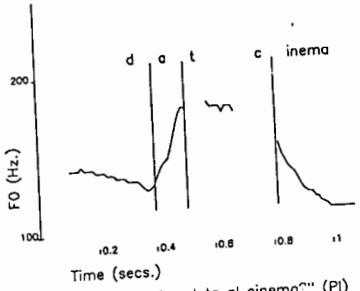


Fig. 4 : "Ma e' andato al cinema?" (PI)

Fo DECLINATION AS A CUE TO DISCRIMINATION OF TONAL CLASSES AND PHRASING IN FRENCH

Pascal Roméas

Institut de Phonétique d'Aix-en-Provence, France.

ABSTRACT

Fo keypoints follow an overall decay from the beginning to the end of French utterances. This is best accounted for when the keypoints are distributed into 3 tonal classes (L, H, S). We compare the significance of linear and 2nd-order polynomial regressions to account for the Fo declination of these 3 classes. This latter regression generally shows a negative second derivative, which leads to a discussion. We find that class S determines the occurrence of declination resettings. The regression significance may be better, under certain conditions, inside sections delimited by S-points than in the span of the whole utterance. We discuss whether regressions may be a cue to resettings.

1. BACKGROUND.

This study deals with the organisation of the melody keypoints in French utterances, from both frequential and temporal points of view.

The speech material is taken from a French simulated man-machine dialog in which only users' requests have been taken into account.

Our earlier works ([10], [11], [12]) have shown the existence of a two-mode organisation of tone in this type of utterances. We distinguish between :

1-suprasyllabic tonal patterns, whose domain and function refer to the lexical relative information load,

2-intrasyllabic contours, which, along with other redundant cues (pauses, etc.), assume a function at the syntactic level (marking of phrase ends).

These two tonal phenomena can be distinguished from three different points of view : acoustic features, phonological association with the syllable string, functions (at lexical, syntactic, and informative levels).

2. AIM OF THIS STUDY.

Considering that the approaches involving sequences of Fo targets (e.g. [1], [7]) have to be tested on our material, we now examine this tonal organisation in a new manner. The pitch maxima of suprasyllabic patterns have been labelled as H. Those of intrasyllabic contours have been labelled as S. All syllables located off these patterns and contours are considered as unstressed (which for French means : low tone). The center of their vocalic part has been labelled as L. Both time and Fo values of these keypoints have been saved in appropriate files.

Our aim is to show that satisfactory regression functions in the time x Fo space can be found to account for these (x,y) keypoints, under some conditions :

- the functions must be calculated independently for each class of keypoints (H, S, L),
- polynomial functions (generally second order) may often provide a better model than linear regressions,
- the model may often be improved when the utterance has been parsed into sections delimited by the S-points (these sections generally match phrasing, since S-points have a syntactic function).

This paper actually draws the first trends, but complete results and general conclusions will be available in our thesis dissertation by September 91.

3. METHODOLOGY.

We deal with 125 utterances, produced by 5 speakers. Fo calculation and representation, as well as tonal labelling, has been run on a Masscomp-5400 mini-computer, using the SIGNAIX speech signal processor [4]. Data have been transferred to a personal computer in order to run statistics.

The necessity for calculating independent regression functions for H, S, and L, is shown in an indirect way, since it relies on the analysis of variance of the three groups.

For each utterance, and for each tonal class, we compared the R-squared and the probability for linear and for polynomial regressions.

When the utterance had S-points, the same regressions have been tested on the sections delimited by these points. We could then see if the R-squared and p were better in the case of sections.

4. RESULTS.

4.1. Regressions must be applied separately to 3 tonal classes.

We said that acoustic features allow a distinction between tonal events involving H (the so-called suprasyllabic patterns) and tonal events involving S (intrasyllabic contours). The major two features are Fo glide threshold and vowel duration. The average vowel duration in the corpus for all syllables except those bearing an S-point is 85ms. As shown in figure 1, vowels bearing an S-point have much longer durations:

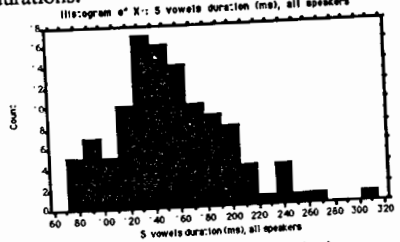


Fig 1: S-vowels duration, all speakers (ms)

mean:	153ms
standard deviation:	43
range:	70 to 304ms
(for 113 values)	

The Fo glide magnitude between S and the preceding L is bigger than the Fo glide magnitude between H and the preceding L:

Ratio Fo(H) / Fo (L), all speakers:	
mean Fo(S)/Fo(L):	1.23
standard deviation:	0.11
(for 113 values)	

Ratio Fo(S) / Fo(L), all speakers:	
mean Fo(H)/Fo(L):	1.16
standard deviation:	0.09
(for 466 values)	

Otherwise, the regression functions applied to the S group alone have a higher constant than the regression functions applied to the H group alone in 98% of cases.

The analysis of variance of the three groups (H,S,L) confirms that the Fo values organisation in the time dimension must be studied for each group separately. We shall call these groups tonal classes.

4.2. linear vs second order polynomial regressions.

Fo declination is generally described as a progressive Fo downdrift from the beginning to the end of the utterance. Declination models often make a distinction between top-line and base-line downdrift ([1], [8], [9], [13]). As seen in (4.1.), and as shown in figure 2, we find it useful to analyse this phenomenon for 3 separate classes, which provide an L-line, an H-line, and an S-line.

These lines are obtained by regression functions. S-lines have low significance since utterances have few S. However the S-lines slopes do not usually differ significantly from the slopes of other classes. The general shape of the downdrift shows a slight convergence between classes rather than a strict parallelism.

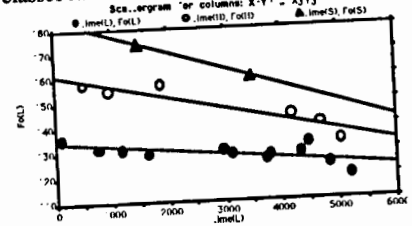


Fig 2: L-line, H-line, and S-line obtained by linear regressions. Time is in milliseconds, Fo in Hertz. Utterance: "j'aimerais connaître le temps prévu le douze juin mille neuf cent quatre vingt deux sur le versant alsacien des Vosges".

The number of H points in the H-lines may be lower than 4 (mean 4.3 per utt.), so that no significant regression can be calculated in these cases (41% of the occurrences). 32% of the H-lines are better accounted for by a linear regression although

approximately 2/3 of these do not reach the probability p=0.1. Actually, if we consider utterances with a greater number of H, the 2nd-order polynomial regression appears to be a better model. This is the case for 27% of H-lines, out of which more than 2/3 have p<0.05.

The same tendency can be noticed for L-lines, which gather more keypoints than H-lines (mean 9.6 per utt.). We found that 60% of the L-lines (generally the ones provided with a greater number of L-points) are better accounted for by 2nd-order polynomial functions. Most of them provide a satisfactory R-squared, and over 95% have p<0.05.

The x^2 coefficient was found to be negative for over 90% of L-lines. In most cases, the lines can be divided into two temporal phases : first, they increase, but they have a negative second derivative (shorter phase); second, Fo drifts down and the second derivative remains negative (which means that Fo steepens with regard to time). See figure 3.

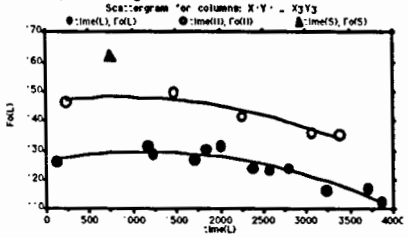


Fig 3: L-line and H-line obtained by a 2nd-order polynomial function. This utterance has only one S-point. R-squared for L is 0.887, p<0.0001. R-squared for H is 0.915, p<0.0855

A discussion of these results is presented below.

Yet 40% of L-lines are better accounted for by linear regressions. One explanation is that most of these L-lines are provided with few L-points. Moreover, only 1/3 out of those cases reach p<0.1 (which is partly due to the low number of points).

4.3. Utterance vs sections.

Trying to find one function to account for keypoints has less and less justification as utterances get longer. Many authors ([1], [3], [13]) have noticed that the course of declination may be reset at major boundaries. Our material provides many long utterances interrupted by silent pauses. The pauses generally occur immediately after S-tones.

We found that roughly half of the L-tones that immediately follow an S-tone (L2) have a higher Fo value than the L-tone that immediately precedes the S (L1). Moreover, if we now consider the Fo difference between L-tones and the y-values of the regression function provided with the same respective x-values, we find that most of the L1 have a negative difference while most of the L2 have a positive difference. This seems to indicate that the resettings must be interpreted with regard to an overall downdrift which covers the whole utterance, rather than to the rough Fo scale. This point deserves further investigation and will not be discussed in this paper.

Another criterion for resetting is the significance of the regression applied to sections delimited by the S-tones, as compared to the regression on the whole utterance.

This criterion is disappointing at first sight. Our hypothesis was that the utterances could be successfully parsed into sections delimited by S-tones. Some utterances do not have S-tones. Otherwise parsing has been attempted as long as S split the utterance in a way which provided the resulting sections with at least one L and one H (thus excluding final S). Finally both linear and polynomial regressions were run on the span of sections. The main problem that we encountered is the lack of points in the sections, especially for H-points.

In cases where the section slopes are obviously reset (relative Fo difference between L2 and L1, silent pause interruptions), the significance of section regressions often remains low. It is lower than the corresponding utterance regressions for 87% of sections, although 38% still provide a p below 0.05.

Yet if we now assume that there is no linguistic reason why the declination models obtained above by regressions on long sequences of points should not be implemented on shorter sequences, we may consider that the R-squared is a better cue than the probability (which is directly related to the degree of freedom). Actually, 63% of section R-squared are higher than the corresponding utterance R-squared.

See illustration in figure 4 & 5.

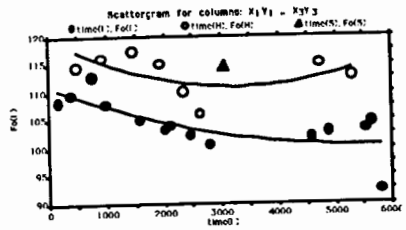


Fig 4: Utterance "Quelles sont les temperatures maxima et minima aux environs de Gérardmer à plus de huit cents mètres aujourd'hui". R-squared for L-line: 0.561, p=0.0108. R-squared for H-line: 0.302, p=0.4076. See next figure for section results.

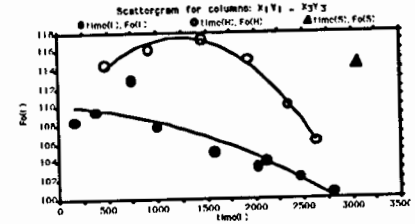


Fig 5: 2nd-order polynomial regressions for the first section of the same utterance as in figure 4. The section ends in the S-point represented by a black triangle. R-squared for L-line: 0.83, p=0.0049. R-squared for H-line: 0.99, p=0.0011.

Further discussions about resetting cues will take place in later publications. We now prefer to focus on the point of the negative second derivative that has been found for most of the declination slopes.

5. DISCUSSION

Negative x^2 coefficient does not confirm the previous observations on the general slope of declination ([1], [5]), which was found to follow an exponential decay (phrase component in Fujisaki's model). This shape of the overall decay has been claimed to be conditionned either by subglottal pressure [6] or by crico-thyroid activity [2]. Beyond this controverse, it may be assumed that declination is mostly determined by the linguistic struture of utterances, and therefore pre-planned independently from physiological constraints ([8], [9], [13]). Since the keypoint values are linguistically conditioned, the exponential decay model cannot be conceived as language- and context-independent. We infer that the steepening slope in our material is due to the specific structure of these French utterances. Since the slope remains a function of time, its shape cannot be conditioned by lexical or syntactic factors. On the contrary, the phatic function may be assumed to weaken smoothly as the S-tone linguistic information nears. The steepening slope may be a pre-indicative cue for the perception of S-tones (i.e. boundaries). As a consequence, the pitch of unstressed syllables in these French utterances should be considered the result of a controlled active process.

REFERENCES.

[1] BRUCE, G., (1984), "Aspects of declination in swedish", *Lund Working papers in Phonetics*, 27, 51-64.

[2]COLLIER, R., (1985), "Setting and resetting of the base-line", *R.I.L.P. Ann. Bull.*, 19, University of Tokyo.

[3] CRYSTAL, D., (1969), *Prosodic systems and intonation in english*, Cambridge University Press.

[4] ESPESSER, R., BALFOURIER, O., (1989), *SIGNAIX, mode d'emploi*, unpublished.

[5] FUJISAKI, H., (1981), "Dynamic characteristics of voice fundamental frequency in speech and singing", *4th symposium F.A.S.E.*, Venice, Apr. 21-24.

[6] GELFER, C., HARRIS, K., COLLIER, R., BAER, T., (1983), "Speculations on the control of fundamental frequency declination", *Haskins Status Report on Speech Research*, 76, 51-63.

[7] HIRST, D.J., (1980), "Un modèle de production de l'intonation", *Travaux de l'Institut de Phonétique d'Aix*, 7, 297-315.

[8] MAEDA, S., (1976), *A characterization of american english intonation*, Ph.D., M.I.T.

[9] PIERREHUMBERT, J., (1979), "The perception of fundamental frequency declination", *JASA*, 66, 2, 363-369.

[10] ROMEAS, P., (1990,1), "Prosodie et lexique : tendances majeures observées en dialogue homme-machine", *Proceedings of the 1st French Congress on Acoustics*, Lyon, April 9-13 1990, Editions de Physique, pp545-548.

[11] ROMEAS, P., (1990,2), "Apport d'information lexicale et marques prosodiques", *18èmes Journées d'Etudes sur la Parole*, Montréal, 28-31 Mai 1990, pp17-20, Pub. de l'Université de Montréal.

[12] ROMEAS, P., (1991), "Organisation prosodique et accès lexical en dialogue homme-machine", *2èmes Journées du PRC-GRECO Communication Homme-Machine*, Toulouse, January 29-30 1991.

[13] SORENSEN, J., COOPER, W., (1980), "Syntactic coding of fundamental frequency in speech production", in COLE, R.(ed.), *Perception and production of fluent speech*, 399-440, LEA, Hillsdale.

RUMANIAN INTONATION STEREOTYPES

Laurenţia Dascălu-Jinga

Institute of Phonetics and Dialectology,
Bucharest, Romania

ABSTRACT

Relying on the criterion of intonation-text relation, the author proposes a classification of intonation stereotypes (ISs) into three categories: (1) ISs depending on a grammatical structure, e.g. Verb+Indefinite Article+Noun etc; (2) ISs representing a higher degree of connection between intonation and text; e.g. intonations specific to some idiomatic phrases; (3) ISs occurring in the absence of any text, i.e. the intonations of "hummed messages" used to express "Yes", "No" etc.

0. PRESENTATION

The term "intonation stereotype" (IS) as used here is meant to be more comprehensive than "stylized tones" or "clichés mélodiques"; it refers to a more or less fixed pattern which is constantly associated with the same semantic and/or pragmatic content.

We propose the classification of ISs into three categories:

1. INTONATION STEREOTYPES DEPENDING ON A GRAMMATICAL STRUCTURE

1.1. The exclamative structures
ce "what" + noun
cît "how many" + noun
may have two intonational variants:
- a one-peaked pattern: an upskip to the accented syllable of the noun is followed by a downskip; e.g.:

Ce bogă$^{\check{t}i}$ *e!* "What abundance!";

- a two-peaked pattern, with the first peak on the exclamative word and the second one on the accented syllable of the noun; e.g.:

C$^{\hat{i}}$*te tab*$^{l^{\circ}}$*uri!* "How many pictures!"

1.2. The structure
ce + noun can be also used with a somewhat opposite attitude, i.e. that of rejecting the partner's statement. In most cases, this rejective exclamation is pronounced in the low part of the voice, with a slightly falling intonation and therefore a narrow pitch range. Since in Rumanian the main function of *ce* is the interrogative one, the rejecting IS may be contrasted to the "homonymous" interrogative pattern; compare:

Speaker A:	Speaker A:
Vrea să plece in Himalaia.	*Am găsit o idee pentru film.*
"He wants to leave for Himalayas."	"I've got an idea for the movie."
Speaker B:	Speaker B:
Ce idee!	*Ce idee?*
What an idea!	What's the idea?

1.3. A high degree in terms of quality (superlative) may be expressed in Rumanian, beside other devices, by using the structure
verb + indefinite article + noun
e.g.: *Era un frig!* "It was terribly cold!" (literally: "It was a cold!"). This "superlative" IS consists of a two-peaked pattern generally ending with a suspended high pitch: the two peaks correspond to the verb and the noun, respectively, the indefinite article being constantly pronounced on a low pitch. The last peak is usually followed by the lengthening of the last vowel on a high sustained pitch:

E ^{o} *mi z*^{e}*riee!* "What a mess!" (Lit.: It's a mess!).

1.4. Another "superlative" Rumanian structure, this time in terms of quantity, implies the same IS:
verb + preposition *la* "at" + noun
e.g.: *Au venit la oameni!* "There came hosts of people!" (Lit.: There came at people!), where the emphatic effect is the result of an ensemble of factors: the necessarily undefinite form of the noun, the unusual word order, the special suspensive intonation pattern, the extra duration of the last syllable vowel:

Au ve^{nit} *l*$_{a\ o}$ ^{a}*meeni!*

2. INTONATION STEREOTYPES REPRESENTING A HIGHER DEGREE OF CONNECTION BETWEEN INTONATION AND TEXT

This is the case of many phrases specific to any language, where a set sequence of concrete words implies a certain IS, or, as Bolinger says: "What we find is either a set intonation or a very restricted range of intonations as part of the set meaning" [2, 98].

Certainly, "there is no string of words that has one necessary intonation" [7, 57] and no intonation represents exclusively a certain text [6, 180]. In fact, to the enormous number of idioms of a given language there corresponds a rather limited number of intonations.

2.1. Most of the idioms have resulted from an ellipsis, accompanied by their semantic reduction. This often creates homonymous utterances with the original ones (with "full" meaning). Perhaps many languages have a number of such "expressions à deux lectures" of which one is idiomatic. In these cases, the intonation represents the only element, (beside the context), which determines the meaning, so that it has a distinctive function.

Actually, a great many of the Rumanian idioms may be regarded as "minimal pairs" of some utterances with the same wording and syntax, which are pronounced generally with a different intonation; compare:

Speaker A:	Speaker A:
Am nevoie de tot salariul tău.	*Care dintre aceste două cărţi îţi trebuie?*
"I need all your wages."	"Which of these two books is of use to you?"

Speaker B: Speaker B:

A^s ta-i bu nă! *As ta-i bu nă!*

"That'a good joke!" (Lit.: do". "This one will do). This is good).

Therefore the ISs attached to the phrases function as such only in their specific linguistic and situational context [5, 4]. We have dealt with some of Rumanian examples in other papers [3; 4]: *Nu mai spune!* "You don't say so!", *De unde!* "Not at all!", *Ce folos!* "What's the use of it!", *De ce nu!* "Why not!", *Nici vorbă!* "Nothig of the kind".

2.2. It seems necessary to make a distinction between the expressions with a meaning by themselves and the ones which resort to the intonation [1, 276-277].

Some idioms have become "frozen" in an odd, ungrammatical form, so that they have no homonymous free pair. Rumanian has many such examples: *Ce mai!, Nici vorbă!, Ce dacă!, Cum să nu!, Ce-are a face!, Ce dracu!, La ce bun!, Vezi să nu!* etc. Generally they have a specific intonation, but not a distinctive one; e.g.:

Speaker A:

O să-mi dai și cărțile tale. "You'll give me your books too."

Speaker B:

V^ezi să n_u! "By no means!"

(Lit.: "See to not").

3. INTONATION STEREOTYPES OCCURRING IN THE ABSENCE OF ANY TEXT

Some of these "intonation carriers" [2, 97] are uttered without opening the mouth; most of them function as different types of replies and probably it is this "sequential" position in the dialogue which makes possible their capacity of being wordless.

According to Fónagy [5, 104], these "tonal gestures" are not less conventionalized than other equivalent responses, like "Yes", "No" etc. Revealing their full and strict meaning, Karcevskij [8, 222] calls them "real algebraic symbols of sentences". Let us see one example:

The hummed message used in Rumanian as an affirmative answer consists of two syllables formed by two syllabic [m]s separated by a "pure nasal aspiration which is generally voiced" [9, 81] Its specific IS represents a rise on a pretty fixed interval of a major second (a slightly greater rising interval implies more interest or participation).

In other languages the rising interval is different; for instance, Fónagy describes the French equivalent as a labial nasal [m] accompanied by an abrupt rise of a seventh, whereas in Hungarian it is characterized by a slower and smaller rise, a "bisyllabic sixth", with two peaks of intensity, one at the beginning, the other at the end of hum [5, 104].

In other languages still, it seems that the same IS may be used with a different pragmatic value, e.g. in United States it is heard as a gentle and shorter answer to "Thank you".

4. COMMENTS

All types of ISs we have dealt with are used in the colloquial, informal language, most of them being affectively or attitudinally marked.

In the case of the first type, a somewhat set pattern expresses a peculiar sense (such as "superlative" or "rejective" etc). In the second case, a definite pattern is assigned to a concrete verbal formula, the ensemble having its own meaning in a given language; that is why the ISs associated with idioms are to be "learned as part of the whole", as pointed out by Bolinger [2, 101].

The functional efficiency of our ISs is obviously decreasing from type 1 to type 3.

Our first type of ISs are known in other languages too, similarly to the so-called "intonation morphemes", whereas the second type represents ISs specific to every language, the same as their corresponding idiomatic text. As for the third type, the use of hummed messages is probably universal; what differs from one language to another is their phonetic aspect and/or meaning.

5. REFERENCES

[1] BALLY, Ch. (1951), "*Traité de stylistique française*", III-ème éd., Genève, Paris, vol.I.

[2] BOLINGER, D. (1981), "Some intonation stereotypes in English", "*Problèmes de prosodie*". Vol.II. "*Expérimentations, modèles et fonctions*" (Studia Phonetica 18), 97-101.

[3] DASCĂLU, L. (1982), "Cîteva răspunsuri interogative și intonația lor în limba română" (Some interogative replies and their intonation in Rumanian), *Studii și cercetări lingvistice*, 33, 39-46.

[4] DASCĂLU-JINGA, L. (1988), "Rumanian idiomatic intonations", *Revue roumaine de linguistique*, 33, 229-236.

[5] FÓNAGY, I. (1982), "*Situation et signification*", Amsterdam/Philadelphia.

[6] FÓNAGY, I., BÉRARD, E., FÓNAGY, J. (1983), "Clichés mélodiques", *Folia linguistica*, 17/1-4, 153-185.

[7] GUNTER, R. (1974), "*Sentences in Dialog*", Columbia, South Carolina.

[8] KARCEVSKIJ, S. (1964), "Sur la phonologie de la phrase", "*A Prague school reader in linguistics*" (compiled by J. Vachek), Bloomington, 206-251.

[9] PETROVICI, E. (1930), "*De la nasalité en roumain. Recherches expérimentales*", Cluj.

PARAMETRIC DESCRIPTION OF GERMAN FUNDAMENTAL FREQUENCY CONTOURS

Bernd Möbius, Grazyna Demenko*, Matthias Pätzold

Inst. f. Kommunikationsf. und Phonetik, Bonn, FRG
*Inst. of Fund. Technol. Research, Poznan, Poland

ABSTRACT

This paper presents a parametric description of German fundamental frequency (F_0) contours as obtained by applying of Fujisaki's intonation model to German. The parameters of the model were extracted by an automatic approximation to naturally produced F_0 courses. The parameter values were standardized using statistical procedures. Finally, intonational prototypes that may be related to linguistic categories were developed by rule. Utterances resynthesized with prototypical F_0 contours were judged highly acceptable by phonetically trained listeners.

1. INTRODUCTION

Intelligibility and naturalness of artificially produced speech may be improved considerably if one allows for prosodic information. Therefore, the generation of prosodic features by rule is an important component of text-to-speech systems. This implies that an adequate description of the intonational variations of the language concerned is at hand. The most outstanding acoustic correlate of intonation is the temporal course of fundamental frequency (F_0).

The aim of this contribution is to separate analytically those factors that determine the F_0 contour of German utterances. This is achieved by applying the quantitative intonation model presented by Fujisaki [1] to German. The model has been elaborated by Fujisaki for the analysis and synthesis of Japanese intonation. It is based on the superposition of a basic value (F_{min}), a phrase and an accent component. The control mechanisms of these components are realized as critically damped second-order systems responding to impulse and rectangular functions, respectively. Thus, the model provides a parametric representation of intonation contours resulting in a considerable data reduction in analysis and synthesis applications.

The extraction of the parameters was attained by a close approximation of naturally produced F_0 curves with the contours generated by the quantitative intonation model. The fitting procedure was implemented in a computer program. Prototypical parameter configurations were derived by statistical analyses and related to linguistic categories, e.g., word accents.

The major issue of this paper is the classification of the parameters and the derivation of intonational prototypes. But first, the speech materials and the method of extracting the parameters will be presented.

2. PROCEDURE

2.1. Speech Materials

In this investigation, the speech material was limited to German declarative sentences containing only one prosodic phrase. 25 test sentences were realized by three male and two female speakers, respectively. For one speaker, the recording was repeated two months after the first session. The same speaker realized another corpus of 25 test sentences.

The stressed syllables of an utterance were determined in a listening test. By definition, each stressed syllable that is characterized tonally by F_0 movements, will henceforth be called accented. Following Thorsen's [2] description of "stress groups", we define an accent group as a prosodic unit that consists of a leading accent syllable optionally followed by unaccented syllables. This unit is independent of any word boundaries but sensitive to major syntactic boundaries.

2.2. Parameter Extraction

Extraction of the parameter values was done automatically. Using a procedure of analysis-by-resynthesis, the original F_0 course is decomposed into the components of the quantitative intonation model. The parameter values are determined by approximating the contour generated by the model to the original F_0 curve. Based on the principle of superposition, the parameter extraction may be carried out for each component of the model separately.

In our interpretation of the model, the phrase component is considered the baseline of the intonation contour with its maximal value at the very beginning of the utterance. In declarative sentences, a standardized negative phrase command was introduced to allow for the final fall. Each accent group is modelled by the contour resulting from one single accent command. The accent command parameters are determined by the method of least squares.

The parameter values extracted by this procedure were treated statistically with the aim of classification and standardization. The final goal was the derivation of intonational prototypes that are perceptually as acceptable as naturally produced contours. The results of the statistical analysis are presented and discussed in the next section.

3. RESULTS

3.1. Basic Value F_{min}

There is a relatively small dispersion of the basic value F_{min} with all five speakers. 50% of the observed values are found to fall into an interval of about 3.0 Hz around the arithmetic mean of each individual speaker. This small variation makes it reasonable to keep the value of F_{min} constant in experiments with resynthesized speech.

3.2. Damping Factors

The damping factors α and β of the phrase and accent components, respectively, are treated as constants. Fujisaki [1] has shown that the approximation of naturally produced F_0 contours by the model is not impaired if α is assumed to be constant. In our investigation, a fixed value of 3,1 s^{-1} was used. The β range was restricted successively in the present study. Finally, the value of 16,0 s^{-1} proved to be suitable for all speakers and all utterances.

3.3. Phrase Amplitude

Since the phrase component is interpreted as the baseline of the contour, only the phrase amplitude has to be considered here. In an analysis of variance (ANOVA), individual characteristics of the speakers, structure of test sentences and corpora, utterance length, and overall speech tempo were tested as potential sources of variation of the phrase amplitude. The most important factors are *speaker* (F=13.7, p<0.0001) and *sentence* (F=2.9, p<0.001). The significant influence of utterance length was reduced to a strong dependence on the factor *speaker*. Other factors were not found to be significant.

Further analyses revealed that the speakers may be classified into three groups. Taking into account the strong dependence of

phrase amplitude on the structure of the sentences, we looked for features common to those sentences with similar phrase characteristics. Since the phrase amplitude is directly related to the steepness of declination, the global downward trend, so our hypothesis, should be stronger especially in those utterances that begin with an accent syllable and end in an unaccented syllable. This hypothesis was confirmed by the result of the ANOVA showing the significant influence of the distribution of accents (F=36.3, p<0.0001).

3.4. Accent Parameters

For this section of the study, a further restriction of the speech materials proved to be necessary. The parameters of the accent component were found to be highly dependent on the position of the respective accent group within the utterance. In order to facilitate the comparison of different utterances, we chose only those sentences that required four accent commands. Thus, the materials consisted of 62 utterances containing 248 accent groups.

With respect to the amplitude of the accent commands, speakers may be grouped into two types. This classification is consistent for all positions of the accent group within the utterance. At the second and third positions, *type of speaker* is the only significant source of variation of the accent amplitude. The duration of the accent group plays a significant role in the initial and in the final position. Furthermore, in utterance initial position, the accent amplitude depends on the word classes: Adjectives require an amplitude value clearly lower than other content words. In the other accent positions, no similar effect of word classes was found.

There is a high positive correlation (r=0.815) between the duration of an accent command and the duration of the accent group to which it is applied. This is also shown by the result of the ANOVA (F=84.1, p<0.0001). No other significant factors determining the duration of accent commands were found.

The temporal distance between the beginning of the accent group and the onset of the command varies with the position of the accent group within the utterance (F=13.7, p<0.0001). While in the first, second and third positions the command is set, on the average, after 10% of the duration of the accent group, the command onset is found immediately after the beginning of the accent group in utterance final position.

Further analysis revealed that there is a direct link between the timing of the accent command and the direction of the accent lending F_0 movement; the effect is significant (F=52.5, p<0.0001). In the speech material under investigation, there is a preponderance of rising and rising-falling movements in the first (98%), second (89%), and third (85%) accent positions. Utterance final accent groups, however, are marked to a large extent (76%) by falling F_0 movements. Since in our interpretation of the model, these movements are approximated by the decreasing part of the contour generated by an accent command, the early command onset in utterance final position is not very surprising. The particular characteristics of final accent commands, differing in some respect from commands in the other positions, is in accordance with the observation that accent groups are sensitive to major syntactic boundaries (cf. section 2.1.).

4. RULE-GENERATION OF F_0 CONTOURS

In the preceding sections, we presented those factors that were found to be responsible for the variation of the parameter values. Standard values were derived subsequently on the basis of the statistically significant factors. A set of rules was formulated in order to generate prototypical intonation contours. By means of LPC analysis and resynthesis, the original F_0 data were replaced by the rule-generated contours.

Acceptability and naturalness of these artificial intonation patterns as well as the adequate realization of the word accents were examined in a listening experiment by six phonetically trained subjects. It turned out that the prototypical intonation contours are highly acceptable: None of the 36 stimuli in the test were rejected by the listeners as being not acceptable or unnatural. With regard to the word accents, more detailed judgements were obtained that led to the improvement of two specific rules, one concerning the slope of the final accent, the other predicting the duration of the accent command in an accent group containing the focus of the utterance. An illustration of a rule-generated intonation contour is given in Fig. 1.

5. CONCLUSION

Fujisaki (1983) has shown that an intonation model based on the superposition principle is a highly useful tool for the analysis and synthesis of the complex F_0 contours in various languages. The effects of different linguistic and speaker-specific features may easily be separated and controlled for, an advantage that facilitates quantitative investigations. The parameters remain constant for a defined stretch of time and may thus be related to linguistic units, in our interpretation to accent groups.

Our application of Fujisaki's work to German has now resulted in a set of rules predicting the parameter values for declarative sentences. Further investigations will have to comprise interrogative sentences, too. The modelling of questions will be particularly interesting, since sentence modality is supposed to be reflected mainly in the phrase component of the intonation model.

ACKNOWLEDGEMENT

This study was supported by the Deutsche Forschungsgemeinschaft and the Alexander von Humboldt-Stiftung.

LITERATUR

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

[2] THORSEN, N.G. (1989), "Stress group patterns, sentence accents and sentence intonation in Southern Jutland (Sønderborg and Tønder) - with a view to German". *ARIPUC (Copenhagen)*, *23*, 1-85.

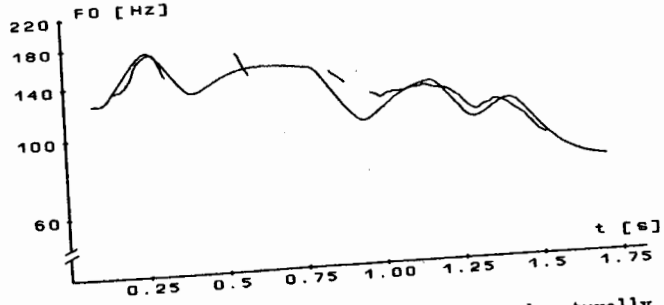


Fig. 1. Rule-generated (continuous line) and naturally produced F_0 contours of the utterance "Hans ißt so gerne Wurst".

PREDICTION PROSODIQUE DE LA MODALITE INTERROGATIVE EN ARABE MAROCAIN

T. BENKIRANE

Université des Lettres et des Sciences Humaines de Fès

ABSTRACT

In moroccan Arabic, the yes-no question melodic pattern is rising-falling as the assertive one.
The prosodic information concerning this modal opposition is located at the terminal part of the utterance. But, at the same time, a distinction is operated by a drastic increase of the frequency range in yes-no question.
Does this particular fact serve as a predictive cue below the nucleus ?
The results of this perceptive study confirm the melodic anticipation. The difference of frequency range assumes a predictive rôle during the activity of processing.

1.INTRODUCTION

Le point d'émergence de cette étude découle d'un ensemble d'observations |5|, |6| concernant les caractéristiques prosodiques de la question totale à réponse oui/non dans le parler arabe du Maroc. Cette dernière,qu'elle comporte un morphème interrogatif ou non, est réalisée avec une configuration fréquentielle montante-descendante qui évoque le patron de l'énoncé déclaratif. A première vue,cette similitude formelle entre les courbes mélodiques de la question et de la déclaration pourrait prêter à équivoque dans la mesure où la chute finale est généralement considérée comme l'apanage du mode assertif. Néanmoins, un examen plus attentif des données acoustiques nous a permis de relever des différences notables régies par l'opposition modale. Ainsi, contrairement à la déclarative, la partie descendante du patron intonatif de la question est rigoureusement synchronisée avec la syllabe finale. Cette chute mélodique qui semble être une propriété inhérente |language specific| à notre parler est matérialisée par un glissando de FØ négatif d'une ampleur supérieure à l'octave et qui se conjugue -interdépendance des paramètres oblige- avec un accroissement de la durée de la syllabe porteuse finale. En revanche,dans le cas de la déclarative, la dernière syllabe est marquée par un ton statique situé dans le grave et le point d'inflexion qui inverse la pente mélodique vers le niveau plancher de la gamme tonale ne concorde pas forcément avec l'ultime portion de l'énoncé. Quant à la partie ascendante du patron interrogatif, elle commence dès la syllabe initiale,culmine sur la pénultième et se caractérise globalement par une FØ plus élevée que celle de la partie assertive correspondante.Cette constatation implique que le fondamental usuel de la voix dans la question est plus haut que le niveau de voix propre à l'assertion.
Sur un autre plan, les résultats d'une investigation perceptive |5| indiquent que l'information ayant trait à l'opposition modale est massivement localisée sur les 2 syllabes finales et que la chute mélodique haute constitue la clef de voûte de la construction interrogative. Cela nous a autorisé,lors d'un premier bilan,à instaurer une ligne de partage entre côté amont, les deux dernières syllabes et côté aval, toutes les syllabes qui les précèdent.

2.PROBLEMATIQUE

Cependant, le fait que ces dernières soient systématiquement dotées d'une FØ plus élevée que celle de leurs homologues déclaratives nous suggère qu'il pourrait faire office d'indice avant-coureur susceptible d'orienter l'auditeur pendant la saisie du signal de parole, de lui fournir la possibilité dynamique d'ajuster ses hypothèses |7| et de l'aider à identifier avant terme, le statut modal de l'énoncé.
Cette problématique qui s'inscrit dans la perspective prédictive de la prosodie |1|, |9|,|10|,|11|,|14| constituera le fil directeur de la présente étude. Elle parfaitement résumée dans l'une des interrogations soulevées par Grosjean,F.|10|: "does the listener know that a yes-no question is being asked only on the last stressed word, or does some information exist before that point ? ".Cette information étant ici de nature prosodique, il est légitime de chercher à déterminer parmi les propriétés acoustiques celles qui par l'importance de leur contribution se haussent au rang d'indice à valeur prédictive. En outre, nous pourrons tenter de savoir à partir de quel moment, si information anticipée il y a,les indices disponibles permettent une identification modale efficace.Par exemple, est-ce que l'auditeur est en mesure de préjuger du caractère interrogatif de l'énoncé qu'il est en train de décoder dès la première syllabe? dès les 2 premières syllabes ou lui en faut-il bien davantage?
Afin d'apporter une réponse à ces préoccupations, nous avons conçu l'expérience perceptive qui suit.

3.EXPERIENCE 1

3.1.Procédure expérimentale

Une phrase a été enregistrée en chambre anéchoïque par 2 locuteurs (A.R.)et (M.B.) de sexe masculin dont la langue maternelle est strictement l'arabe marocain (pour éviter toute interférence avec le berbère). Cette phrase a été réalisée une fois avec une intonation déclarative sans pause ni emphase et une seconde fois avec une intonation de question.Cette phrase composée de 14 syllabes à voyelles pleines est la suivante :
|ʒabu|lmalika|lmagana|djal lmakina|ʒʒdida|
("Ils ont apporté|à Malika|le compteur|de la machine|neuve|").
Les barres verticales correspondent,grosso modo,à des limites entre mots. L'enregistrement obtenu a été numérisé (convertisseur 16 bits |17| et analysé grâce à un système de traitement du signal fonctionnant sous Unix |8|.

3.2.Méthodologie

Les 4 phrases ont ensuite fait l'objet d'une segmentation opérée sur une console graphique.
De chacune des 4 phrases-mères (dorénavant Pm) nous avons extrait,par troncations successives, 8 séquences ou sections|S1 à S8| qui serviront de support au test perceptif :

Pm	ʒabu lmalika lmagana djallmakina ʒʒdida
S1	ʒa
S2	ʒabu
S3	ʒabu lma
S4	ʒabu lmalika
S5	ʒabu lmalika lma
S6	ʒabu lmalika lmagana
S7	ʒabu lmalika lmagana djallma
S8	ʒabu lmalika lmagana djallmakina

La bande expérimentale est constituée des seules portions tronquées S1 à S8.la Pm n'apparaît jamais dans sa totalité.En effet la clausule|ʒʒdida| est supprimée parce qu'elle est porteuse de l'essentiel d'information concernant l'opposition modale. Les sections opérées répondent à des critères syntaxiques (cas des sections paires) ou phonologiques (cas des séquences impaires: S3,S5,S7).Ces dernières exploitent judicieusement le fait que certains mots débutent par la séquence phonique |lma| qui correspond,dans la langue,au signifié "l'eau".Cette coïncidence a rendu possible une coupe à l'intérieur des mots qui déborde la frontière syntaxique sans,par ailleurs, porter atteinte à la cohérence sémantique de la section ainsi obtenue.

3.4.Confection de la bande expérimentale

Une fois les 8 sections générées et répertoriées,nous avons procédé à leur appariement: les séquences extraites de la Pm assertive forment avec leurs correspondantes issues de le Pm interrogative des paires de stimuli. Suite à un tirage aléatoire le premier stimulus de la paire est tantôt d'origine affirmative,tantôt interrogative. Cette précaution permet d'obvier à l'éventualité d'un effet d'ordre sur les réponses Chaque paire est annoncée par un bip sonore et comporte un silence interstimuli d'une seconde.A la fin du second stimulus, l'auditeur dispose de 3 secondes pour donner sa réponse.Nous avons constitué par un tirage au sort automatisé 12 séries (6 par locuteur) de présentation et chaque série comporte 16 paires de stimuli.

3.5.Consigne

Au total,15 auditeurs marocains ont participé,par séance individuelle,au test qui se déroulait dans une pièce calme. les séries étaient présentées par écouteurs à un niveau normal de parole (65dB).La tâche du jury a consisté à identifier,sous choix binaire forcé, les stimuli originaires de la question totale.L'auditeur disposait sur la feuille de réponse de la transcription en arabe de toutes les sections rangées selon leur ordre de présentation sur la bande expérimentale et devait après l'écoute de chaque paire, désigner le stimulus reconnu comme interrogatif soit par le chiffre 1 quand le premier stimulus de la paire est interrogatif, soit par le chiffre 2 dans le cas contraire. Dans le but de familiariser le sujet avec la bande,une écoute non comptabilisée de la première série a été effectuée.

4.RESULTATS DE L'EXPERIENCE 1

Si l'on admet que le seuil significatif de détection se situe à 75 % ,alors comme cela ressort des résultats confinés sous forme de pourcentages (%) dans le tableau 2,les taux de réponses,pour toute section, sont largement supérieurs à cette valeur liminaire.

Tableau 2: Scores (en %)d'identification des stimuli interrogatifs en fonction des sections et des 2 locuteurs.(Variables en jeu : FØ et Durée)

	S1	S2	S3	S4	S5	S6	S7	S8
A.R.	84	100	95	100	95	100	98	100
M.B.	93	91	95	100	95	91	100	100

5.DISCUSSION

L'ensemble du jury a fait preuve d'une conduite cohérente et ne semble pas avoir été gêné par l'effet de troncation et par son corollaire la réduction de l'empan temporel des séquences stimuli.Cela indiquerait que l'information prosodique située en aval de la clausule suffit à l'identification précoce de la question.La variabilité acoustique interlocuteurs n'a exercé aucun effet notable (p=.1).Cependant, au vu des résultats,le score qui est relativement le plus faible (84 %)est le fait de la séquence monosyllabique S1 du locuteur A.R. Cette atténuation de la performance,toutes proportions gardées,est vraisemblablement imputable à l'insuffisance des indices prosodiques dans ce contexte. En effet, le seul indice opérant dans ce cas demeure l'écart mélodique (3.2 DEMI-TONS),la durée étant pratiquement similaire Pour s'en convaincre,il suffit de considérer,dans le cas de M.B.,le score de 93 % enregistré dans le même contexte et ce,à la lumière des facteurs en jeu : un écart mélodique de 5.6 demi-tons et une différence de durée supraliminaire |15|,ici de l'ordre de 40 %. Cette amélioration du score (+9 %), est-elle due à l'effet cumulé des 2 facteurs (FØ + Durée) ou bien à l'importance de l'écart mélodique seul ?
Cette interrogation qui mérite d'être étendue à l'ensemble de nos résultats nous a incité à déterminer de façon stricte la part de la FØ dans le processus de décodage anticipé de la question.L'expérience suivante sera menée dans cette optique.

6.EXPERIENCE 2

La neutralisation des différences de durée va nous permettre d'une part,d'être dans une situation expérimentale rigoureuse à une seule variable et d'autre part, d'évaluer l'impact d'une telle réduction de la donne prosodique sur les performances.

6.1.Procédure expérimentale

Pour cela, nous avons repris les mêmes séquences S1 à S8 assertives et interrogatives des 2 locuteurs de l'expérience 1 et procédé à l'égalisation de leurs longueurs réciproques en ramenant la durée de chaque section affirmative à celle de sa correspondante interrogative et vice versa.De la sorte,l'égalisation temporelle interstimuli a été effectuée dans chaque paire,une fois par référence à la durée plus longue du stimulus affirmatif (dilatation) et une autre fois en prenant pour cible la durée plus courte du stimulus interrogatif (compression).Ces procédures de dilatation et de compression ont été réalisées automatiquement grâce à un programme informatique implanté sur PDP 1123.
La bande expérimentale a été confectionnée selon le même canevas décrit dans le protocole précédent.Elle est composée de 12 séries (6 par locuteur)comportant 16 paires de stimuli chacune. Les sujets qui ont participé à l'expérience 1 n'ont pas été sollicité pour la seconde.Les 20 auditeurs qui ont bien voulu se prêter au nouveau test ont reçu la même consigne et passé l'expérience dans des conditions conformes à celles de la première.

7.RESULTATS DE L'EXPERIENCE 2

Toutes sections confondues,les stimuli du locuteur A.R. recueillent en moyenne 95.3 % des suffrages (écart-type=6.1) et ceux du locuteur M.B. 96.8 % (écart-type=2.4), soit pratiquement les valeurs moyennes obtenues dans le cadre de l'expérience 1. Le tableau 3 suivant fournit le détail des résultats enregistrés.

Tableau 3: Scores (en %)d'identification des stimuli interrogatifs en fonction des sections et des 2 locuteurs.(Variable en jeu : FØ)

	S1	S2	S3	S4	S5	S6	S7	S8
A.R.	82	90	99	98	98	98	98	99
M.B.	92	95	98	99	99	97	98	96

8.DISCUSSION ET CONCLUSIONS

Les très faibles fluctuations observées entre les résultats des 2 expériences ne sont dotées d'aucune signification sur le plan statistique. Est-ce à dire que le facteur durée, présent dans la première et neutralisé dans la seconde,doive être tenu pour quantité négligeable ? Nous pouvons répondre par l'affirmative dans la mesure où, sur le plan de l'encodage, les différences temporelles entre question et déclaration demeurent dans 14 sections sur 16 suffisamment en deçà du seuil de durée |15| pour susciter un effet perceptible. Cela justifie de faire l'impasse sur une possible troisième expérience où FØ et intensité seraient neutralisées au profit de la durée. En conséquence, c'est reconnaître que dans la palette des paramètres prosodiques qui contribuent à dresser un profil de la question distinct de celui de la déclaration, c'est la FØ qui exerce sans conteste, ce pouvoir séparateur.Il y a donc lieu de parler d'anticipation mélodique, puisque le rehaussement du registre fréquentiel de la question en regard de l'assertion est porteur,réellement,d'une valeur indicielle et préindicative. A ce propos, nous aimerions rappeler que l'écart mélodique moyen entre les 2 courbes considérées est supérieur à 3 demi-tons; valeur critique si l'on croit les résultats d'une étude de T'Hart rapportés dans Sorin |16|.Dans le même sens,nous avons établi dans |5| que lorsque l'écart mélodique tombe au dessous des 3 demi-tons, les réponses du jury s'accompagnent d'une grande incertitude.

En définitive et malgré les réserves et les limitations auxquelles notre étude n'échappe pas,l'orientation cohérente des réponses et le fait que les résultats de la seconde expérience étayent ceux de la première, tout cela abonde dans le sens de l'anticipation mélodique et confirme, dans les limites du paradigme expérimental adopté, la fonction prédictive assumée par la prosodie.

9.REFERENCES BIBLIOGRAPHIQUES

|1|AUTESSERRE,D.et DI CRISTO,A.(1972) "Recherches psychosémantiques sur l'intonation de la phrase française",Travaux de l'institut de phonétique d'Aix-en-Provence 1,61-98.

|2|BAGDASSARIAN,N.(1987)"Prédiction de la complexité syntaxique et prédiction de la modalité interrogative",Mémoire de D.E.A., Université de Provence.

|3|BENHALLAM,A.(1990)"Native speaker intuitions about moroccan arabic stress", La linguistique au Maghreb,collection dirigée par PLEINES,J. 91-109,Editions Okad, Rabat

|4|BENKIRANE,T.(1982)"Etude phonétique et fonctions de la syllabe en arabe marocain" Thèse de 3°cycle, Université de Provence.

|5|BENKIRANE,T.(1991)"Faut-il soulever la question ou la laisser tomber? Etude acoustique et perceptive de la question totale en arabe marocain",Linguistica communicatio,Vol.3,N°1,Casablanca.

|6|BENKIRANE,T.(1991)"Intonation systems : Western Arabic",Intonation Systems, édité par HIRST,D.et DICRISTO,A.,Cambridge University Press

|7|DARWIN,C.J.(1975)"On the dynamic use of prosody in speech perception",Structure and Process in Speech Perception,édité par COHEN et NOOTEBOOM,178-193.

|8|ESPESSER,R.(1985)"Signaix:un logiciel de traitement de signal sous UNIX",Travaux de l'institut de phonétique d'Aix-en-Provence,10,335-357.

|9|FONAGY,I.(1979)"Fonction prédictive de l'intonation",Phonetica,18,Vol.2,113-120.

|10|GROSJEAN,F.(1983)"How long is the sentence? Prediction and prosody in the on--line processing of language",Linguistics, 21,501-529.

|11|LHOTE,E.(1979)"Quelques problèmes posés par l'élaboration de règles prédictives de l'intonation",Current Issues in Linguistic Theory,Vol.9,310-319,édité par H.et P.HOLLIER,Amsterdam

PHONETIC CORRELATES OF THE 'NEW/GIVEN' PARAMETER

MERLE HORNE

Dept. of Linguistics, U. of Lund
Helgonabacken 12, S-223 62 Lund

ABSTRACT

Production data from American and British English speakers are examined to see whether the discourse parameter 'new/given' has phonetic correlates as regards accentual patterning in initial subject constituents. The results show no significant difference for the American speakers. For the British English speakers, however, it was observed that differences in Fo register width in the H* tone as well as the use of categorically different tonal patterns correlate with the discourse parameter 'new/given'.

1. BACKGROUND

In a previous study [3], we made a preliminary investigation to ascertain whether British and American speakers use intonation to distinguish between sentence-initial subjects which are contextually 'new' (brand new) versus those which are contextually 'given' (i.e., mentioned previously). In a related study, Eady et al. [2] measured Fo peak height and found no significant difference in this parameter for a group of American English speakers. In our study, we decided to measure in addition Fo register width on the subject, since it is known that differences in the size of an Fo obtrusion can lead to perceptually significant differences in prominence levels [4]. Results of our study indicated that, for both dialects, speakers do not make any distinction as regards Fo peak height on the stressed vowel (this result is in agreement with Eady et al. [2]). As regards register width on the tested word, however, it was found that the British, but not American speakers tested used this parameter to distinguish between new and given, with new information being assigned a wider register than given. That is to say, significant variations in the H*(igh) L(ow) tonal contour on the head word of the subject phrase were used to distinguish between contextually new vs given information. However, the data presented there were very limited (subject constituents containing one lexical word with one accentable syllable (*man, Mormon*) as well as the structurally ambiguous *young man* (compound or phrase?). Since there was for the most part only one accentable syllable present in the data, the speakers were very restricted in their choice of tonal contours for the subject constituent. This is because an accented syllable ('nucleus' [1] (which is normally H* in the dialects studied) is necessary somewhere in the intonational phrase ('tone unit') if it is to be well-formed. Thus, it is not possible to delete the accent (H* tone) on the subject if it is the only accent in the intonational phrase even if it is contextually given. Consequently, varying register width within a tonal category is a possible strategy for creating linguistic distinctions using prosodic parameters. For the present study, therefore, we decided to examine an additional number of cases with more than one lexical word and consequently more than one accentable syllable to ascertain if speakers use the same or different strategies in handling these more complex cases. With more than two accentable syllables, e.g. *new miller*, one could expect that in the 'given' cases, either the speaker could narrow the H*L tonal contour register as the British English speakers did in the previous study, or even use a different Fo contour (e.g. delete the accent on *miller*, provided an accent on *new* was realized in order to make the intonational phrase well-formed).

2. DATA AND SUBJECTS

The data in (1) were used in the investigation. Four speakers participated in the experiment (2 American English, one male (Kansas) and one female (Louisiana), and 2 British English, both female (one from N.E. England and one from N.W. England). All but the speaker from N.W. England had participated in the previous experiment and all but this subject have some degree of linguistic and/or phonetic background. The sentence pairs were typed on cards and were presented in random order along with 10 other filler sentences used in other experiments. The heads of the subject constituents in the final sentence of each sentence pair constituted the material to be investigated in detail, i.e. *miller, milliner, millionaire, Milan* and *Milwaukee*. The test words were also recorded in sentences where they functioned as subjects of embedded clauses, but, at the present time, these cases have not been analysed.

(1) (a) According to the farmers, there is a shortage of workers. *A new miller* will be very welcome.

(b) According to rumours, there will soon be a new miller. *The new miller* will be very welcome.

(2)(a) According to the merchants, there is a shortage of shops. *A new milliner* will be very welcome.

(b) According to rumours, there will soon be a new milliner. *The new milliner* will be very welcome.

(3)(a) According to the bankers, there is a shortage of investors. *A new millionaire* will be very welcome.

(b) According to rumours, there will soon be a new millionaire. *The new millionaire* will be very welcome.

(4)(a) According to reports, there is a need for a new tourist attraction. *A new Milan* will be very welcome.

(b) According to reports, a new Milan will be needed in the future. *The new Milan* will be very welcome.

(5)(a) According to the dope dealers, there is a shortage of marijuana in the East. *The marijuana in Milwaukee* is wanted in Washington.

(b) The gangsters in Milwaukee have just got a message from the East. *The marijuana in Milwaukee* is wanted in Washington.

Notice that in (5), it is just the phrase-final lexical item, and not the whole phrase, which is either given or new as is the case in the other test sentences.

3. ANALYSIS PROCEDURE

The sentence pairs in (1) were read four times and recorded in the sound studio at the Dept. of Linguistics, U. of Lund. This resulted in 5 test words x 2 parameters (new/given) x 4 speakers x 4 readings = 160 target sentences. Acoustic analysis of the final sentence in each of the pairs was performed using Lund University Prosodic Parser, a program developed by Lars Eriksson and implemented on a Macintosh II computer. The speech was first digitized at a sampling rate of 10 kHz. Examination of the Fo contours revealed that the speakers did not always use the same tonal pattern. In the majority of cases, the lexically stressed syllable of the subject head bore a H* tone as in our previous study. However, in a number of the 'given' cases, the British English speakers produced another pattern, with a falling or L(ow) tone on the stressed syllable of the phrasal head. These categorically different cases were not analyzed together with the H* tone data. The results, which are thus based on between 2 and 4 readings, are presented below. The following measurements were made: a) Fo peak (highest Fo value) in the lexically stressed syllable of the phrase-final lexical word, and b) the size of the Fo register on this word, i.e. the distance between the Fo peak and the bottom of the fall (L) after the final H* on the subject.

4. RESULTS

Results are presented below in Table 1.

Table 1. Means, standard deviations and ratios ('new/given') for four speakers. Test words are printed in bold type.

	Fo Peak (Hz)		Fo Register (Hz)	
	NEW	GIVEN	NEW	GIVEN
Am.Male				
Miller				
x̄	167	178	63	73
s	6.1	7.4	6.1	8.8
Ratio	0.94		0.86	
Milliner				
x̄	166	178	67	75
s	13.0	12.0	10.0	8.0
Ratio	0.93		0.89	

	Fo Peak (Hz)		Fo Register (Hz)	
	NEW	GIVEN	NEW	GIVEN
Milan				
x̄	154	154	58	52
s	4.4	11.1	2.8	6.7
Ratio	1.00		1.11	
Millionaire				
x̄	175	166	84	74
s	5.1	4.3	4.4	3.3
Ratio	1.05		1.13	
Milwaukee				
x̄	163	150	65	55
s	8.2	5.7	5.6	4.6
Ratio	1.09		1.18	
American Female				
Miller				
x̄	246	250	64	63
s	6.8	8.8	8.3	5.5
Ratio	0.99		1.02	
Milliner				
AE				
x̄	249	244	65	62
s	7.4	4.6	6.9	4.1
Ratio	1.02		1.04	
Milan				
x̄	245	242	64	55
s	4.3	14.8	5.0	14.8
Ratio	1.00		1.16	
Millionaire				
x̄	254	252	78	76
s	5.2	7.7	6.9	3.4
Ratio	1.00		1.02	
Milwaukee				
x̄	256	243	72	66
s	4.5	2.4	3.7	10.5
Ratio	1.05		1.09	
British (NE)				
Miller				
x̄	249	246	65	54
s	5.6	19.2	11.8	20.3
Ratio	1.01		1.20	
Milliner				
x̄	257	259	66	71
s	4.7	17.2	1.9	15.0
R	0.99		0.92	
Milan				
x̄	260	237	66	40
s	10.5	2.1	13.5	0.7
Ratio	1.09		1.65	
Millionaire				
x̄	251	243	64	48
s	2.6	11.3	8.3	12.0
Ratio	1.03		1.33	
Milwaukee				
x̄	259	258	84	72
s	14.9	17.6	17.8	12.7
Ratio	1.00		1.16	
British (NW)				
Miller				
x̄	284	253	121	97
s	9.5	1.6	9.4	5.5
Ratio	1.12		1.24	
Milliner				
x̄	322	257	165	108
s	29.1	27.0	40.0	22.3
Ratio	1.25		1.52	
Milan				
x̄	234	226	83	69
s	0	17.0	1.4	5.0
Ratio	1.03		1.20	
Millionaire				
x̄	234	No H*	81	No H*
s	6.3	data	8.4	data
Milwaukee				
x̄	259	212	103	58
s	18.7	12.5	22.6	6.1
Ratio	1.22		1.78	

In Table 2 are presented the average ratios (New/Given) for each speaker:

	Fo Peak	Fo Register
Am. Male	1.00	1.03
Am. Female	1.02	1.06
Br. N.E.	1.02	1.26
Br. N.W.	1.15	1.42

These results show that, as in the previous study, the American speakers do not differentiate between the categories given and new as far as peak height and register width are concerned. The biggest difference in register width, 1.18, corresponds to 1.1 semitones which is not perceptually distinctive (excursion size differences of 1.5 semitones have been found to cause a difference in the perception of prominence [4]). Even the British (NE) speaker does not in this study show any convincing variation of register width as was the case in the previous study, where a ratio of 1.54 (corresponding to about 6 ST) was obtained. The present mean ratio, 1.26,

corresponds to an actual difference of around 18 Hz, or 0.8 ST which is not sufficient to create any perceptual difference between new and given cases. However, in 25% of the given cases here, the speaker actually used a categorically different tonal pattern, 'deaccenting' the subject head (see Fig 2). This suggests that the speaker does have the option of distinguishing prosodically between the two discourse categories. The speaker from NW England, however, presents more convincing results; a mean 'new' vs 'given' ratio of 1.42 in register width corresponds to an actual difference of about 35 Hz or 2.44 ST, a difference which can be assumed to be perceptually distinct. This speaker, furthermore, used a categorically distinct tone in 35% of the 'given' cases, i.e. without a H* on the stressed syllable of the subject head.

5. CONCLUSION

The data presented here indicate that the discourse parameter 'new/given' can, but does not necessarily have prosodic correlates. The American speakers studied show no difference on this parameter. With respect to the difference in register width of the H* tone, it was seen, however, that one of the two British English speakers used perceptually significant differences between 'new' and 'given' as regards this correlate. Moreover, in 30% of the given cases, categorically different tonal patterns with respect to those produced in the 'new' cases were produced by the Br. English speakers.

6. REFERENCES

[1] CRYSTAL, D. (1969), "*Prosodic systems and intonation in English*", Cambridge : Cambridge UP.

[2] EADY, S., COOPER, W., KLOUDA, G., MUELLER, P. & LOTTS, D. (1986), "Acoustical characterization of sentential focus: narrow vs. broad and single vs dual focus environments", *Language and speech* 29, 233-250.

[3] HORNE, M. (1990), "Accentual patterning in 'new' vs 'given' subjects in English", *Working papers* (Dept. of Ling., U. of Lund) 36, 81-97.

[4] RIETVELD, A.C.M & GUSSENHOVEN, C. (1985), "On the relation between pitch excursion size and prominence", *Journal of phonetics* 13, 299-308.

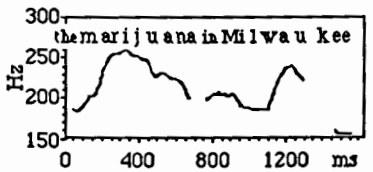


Fig. 1a. Fo contour produced by Br. Eng. (NW) speaker for *Milwaukee* 'new'.

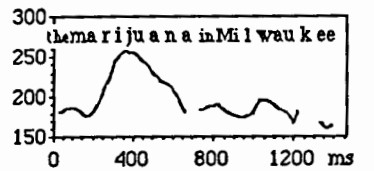


Fig. 1b. Fo contour produced by Br. Eng. (NW) speaker for *Milwaukee* 'given'.

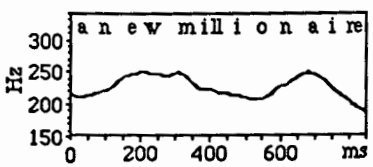


Fig. 2a. Fo contour produced by Br. (NE) speaker for *millionaire* 'new'.

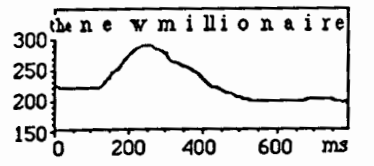


Fig. 2b. Fo contour produced by Br. (NE) speaker for *millionaire* 'given'.

CODING THE F0 OF A CONTINUOUS TEXT IN FRENCH : AN EXPERIMENTAL APPROACH.

Daniel Hirst, Pascale Nicolas and Robert Espesser

Institut de Phonétique d'Aix,
URA CNRS 261 Parole et langage.
Aix en Provence, France

ABSTRACT

An algorithm for the automatic modelling of fundamental frequency curves as a quadratic spline function was applied to a continuous text in French. The output of the model, a sequence of target pitches <ms; Hz>, was then coded using four successively more complex models which were subsequently used to generate synthetic versions of the recording, each version respecting the statistical distribution of the modelled target pitches. The resulting recordings were evaluated subjectively by native speakers. The two more complex codings obtained over 80% of the score obtained by the resynthesis of the text using the measured targets.

1 INTRODUCTION

A number of speech synthesis systems are available today for several different languages, capable of intelligibly synthesising isolated sentences. Results for continuous texts, however, are far less satisfactory. It is generally agreed that one of the principal weaknesses of such systems is the inadequate modelling of prosodic parameters : fundamental frequency, intensity and segmental duration.

In this paper we present the preliminary results of a project investigating the fundamental frequency structure of continuous texts in French.

2 METHOD

The research makes use of an automatic fundamental frequency modelling program MOMEL [6] combined with the PSOLA technique for time domain prosodic modification of speech [2].The F0 modelling program uses a dissymetric version of robust regression to provide an optimal fit for a sequence of parabolas,.factoring the F0 curve into two components, a microprosodic profile and a macroprosodic profile [1]. The output of the program is in the form of a sequence of target-points <ms; Hz>. These target-points can subsequently be used to generate a quadratic spline function [3] which is then directly usable as input for PSOLA resynthesis. For very high quality synthesis the microprosodic profile can be reintroduced, although owing to the high quality of the PSOLA synthesis, even without microprosodic correction, the resynthesis using this technique is practically indistinguishable from an original recording as far as the intonation is concerned.

3 CORPUS

The corpus used in the experiment was recorded from an introduction to science for French children. The text consists of 3 paragraphs, 8 sentences, 140 words and 232 syllables and develops a single topic "the atom".

The text, presented in normal orthography with its original punctuation, was recorded in an anechoic chamber by 6 subjects : 3 male and 3 female, all native speakers of French. The recording used for the experiment described here was that of a 30 year old female.subject whose reading was considered the most satisfactory of the six, presenting a harmonious rhythm and no hesitations. The complete recording including pauses lasted 55secs.

4 ANALYSIS

The fundamental frequency of the recording was analysed by means of spectral comb analysis [8], and the F0 was subsequently modelled by the automatic modelling program described above resulting in a set of 170 target points. Of these 170 values 3 were deleted and 3 others added manually after visual and auditory evaluation of the modelled curve. The resulting values, (the first 40 are illustrated in Figure 1) constitute the reference set **A0**.

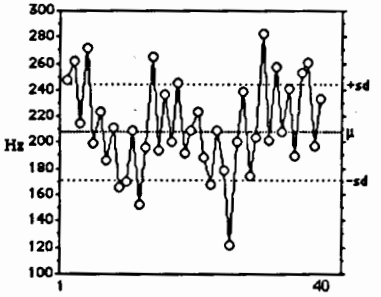


Figure 1 : first 40 target values from the reference set **A0**.

5 PROSODIC CODING

Four approximations to the original prosody of the text were obtained.by successively more complex prosodic coding of the different target points. In this analysis only the F0 values of the target-points were modelled : the time values being taken as given.

5.1. Model A1

The first approximation consisted of a sequence of values assigned randomly but with the same statistical distribution (i.e. the same mean and standard deviation) as the reference set **A0**. :

A1	N	mean F0	s.d.
	170	199	37.1

This approximation was introduced to test the possibility that the only relevant function of fundamental frequency variation is to avoid monotony.

5.2. Model A2

In a second approximation, each value was coded as either Higher (**H**) or Lower (**L**) than the preceding target point. This coding was intended to introduce a distinction between relatively high and low points of an F0 curve corresponding to the distinction used in a number of phonological models of intonation between High and Low tones constituting pitch accents [9],[3],[4].In a preliminary attempt, a set of target points was generated such that the intervals between successive points had the same statistical distribution as those corresponding to similarly coded points of the reference set **A0**. Informal listening showed however that such a model was very unsatisfactory since there was a tendancy for a sequence of values to drift into very high or very low regions, beyond the normal range of the subject's voice. To counteract this, a form of asymptotic declination needs to be introduced [7]. An extremely simple model of declination is given by the formula : $h_i = \sqrt{h_{i-1} * h_A}$ [4] where h_i is a pitch target and h_A is an asymptotic value. This formula was originally intended to model pitch lowering but the same formula can be used for both lowering and raising assuming simply distinct asymptotic values. Estimates of h_A can be then made simply. from each successive pair of values using the formula : $h_A = (h_i)^2 / h_{i-1}$ The successive values can then be generated using the declination formula and an asymptotic value with the same statistical distribution as that of similarly coded values of the reference set **a0**

A2 code	N	mean asymptote	s.d.
L	82	166	57.1
H	87	255	100.3

5.3. Model A3

In a third approximation, the text was segmented manually into Intonation Units on the basis both of textual content and of the F0 contour. The highest point in each Intonation Unit was then coded Top (**T**), the lowest point Bottom (**B**). The first point of each unit was coded Mid (**M**) if not already coded. Other points were first coded Higher (**H**) and Lower (**L**), as in **A2** and then recoded so that an **H** immediately followed by **H** or **T** was recoded as Down (**D**) while **L** immediately followed by **L** or **B** was

recoded as U. This coding corresponds to a transcription using INTSINT (an INternational Transcription System for INTonation) as proposed recently [5] in an attempt to set up a system capable of transcribing significant pitch patterns in any language. The coding incorporates two types of symbols : absolute symbols T, M and B, and relative symbols ; H, D, U and L. Target points coded with absolute symbols were assigned F0 values having the same statistical distribution as similarly coded points of the reference set A0.

A3	code	N	mean F0	s.d.
(absolute)	B	17	146	19.6
	M	13	204	23.7
	T	17	263	19.0

Targets coded with relative symbols were assigned values using the declination formula already used for A2 and an asymptotic value with the same statistical distribution as that of similarly coded values of the reference set A0:

A3	code	N	mean asymptote	s.d.
(relative)	L	44	152	24
	U	10	208	28
	D	24	147	20
	H	45	262	39

5.4. Model A4

The absolute values coded in A3 all show a fairly large standard deviation reflecting considerable variability. The final approximation A4 tested here was designed to limit the variability of T and B by restricting these symbols to target points respectively higher than 281 Hz and lower than 141 Hz.

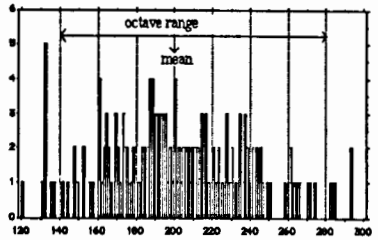


Figure 2 : distribution of target points from the reference set A0.

These thresholds were set to include only values which were beyond an octave range centred around the mean F0 of the recording. As can be seen from the frequency distribution in Figure 2 these values isolate fairly well the extreme values of the distribution which also correspond well to the beginning and end-points of paragraphs.

A4	code	N	mean F0	s.d.
(absolute)	B	10	133	5.5
	M	13	204	23.7
	T	4	288	5.8
	code	N	mean asymptote	s.d.
(relative)	L	51	149	24
	U	10	208	28
	D	24	147	20
	H	58	279	58

6 EVALUATION

In order to reduce the quantity of material to be evaluated, only the first two thirds of the text were used for the evaluation test. Three different series of recordings were obtained, each containing one version generated from the reference set A0 as well the four approximations A1, A2, A3 and A4 described above. No corrections were made for microprosodic effects.The five recordings in each series were presented in random order. The actual set of target values for each of the four models was different for each series but respected the statistical constraints described above. The first series was used as practice material. Subjects were asked to listen to the 5 recordings which, they were told, had undergone a number of different treatments which might affect the way the readings sounded. Subjects were then asked to listen to the second and third series of recordings and to underline portions of the text which they felt sounded least satisfactory. At the end of each recording subjects were asked to attribute a global score out of 20 reflecting their overall satisfaction with the recording.

7 RESULTS

12 subjects took part in the evaluation test, all students or personnel of the Université of Provence. None of the subjects had any particular training in phonetics. An analysis of variance of the subjects' scores showed a significant difference between the scores for the various models ($F(4,110) = 13.286$, $p = 0.0001$) but no significant difference between the two series ($F < 1$) and no significant interaction between model and series ($F < 1$). The reference set A0 received a mean score of 16.5. The four approximations received the following mean scores :

	A1	A2	A3	A4
score	8.5	11.1	13.4	13.6

8. DISCUSSION

Listeners showed a clear preference for the binary (H,L) coded targets (A2) compared to the random distribution (A1). They also showed a clear preference for the two versions of INTSINT tested as compared to the other approximations. The version of INTSINT incorporating a threshold (A4) received a slightly better score than the other version (A3) but the difference was not significant here. Both versions of INTSINT attained more than 80% of the score attributed to the reference set A0. This suggests that while improvement is still possible, these two models provide a very reasonable approximation to the estimate values. Further experimentation will be necessary, however, to decide whether the incorporation of a threshold provides a distinct improvement to the coding.

It is worth emphasising that the only manual intervention in the coding concerned the determination of the position of the boundaries of Intonation Units. Once these boundaries are determined, the coding used in both models A3 and A4 is entirely automatic. Inspection of the means of both the absolute and the relative values for the two versions of INTSINT suggests an interesting further generalisation. The mean values for T, M and B are quite close to the mean asymptote values for H, U and L/D respectively, with L and D seeming to have identical .asymptotic values. An extremely simple approximation to these values can be obtained from the mean F0 and the two extreme values covering an octave range centred on the mean, the same values which were used as threshold values in model A4. This means that the phonetic implementation rules described above could be implemented using a single individual parameter : the mean speaker F0 together with a standard deviation of say 10%. It remains to be seen, however, how far such a model can be generalised to other speakers and other languages.

9 REFERENCES

[1] Di Cristo, A. & Hirst, D.J. (1986) "Modelling French micromelody : : analysis and synthesis." *Phonetica* 43, 11-30

[2] Hamon, Moulines & Charpentier (1989) "A diphone system based on time domain prosodic modifications of speech". *Proc. Int. Conf. Assp.*, 239-241.

[3] Hirst, D.J. (1983) "Structures and categories in prosodic representations." in Cutler & Ladd (1983) *Prosody : Models & Measurements* (Springer, Berlin) , 93-109

[4] Hirst, D.J. (1987) *La description linguistique des systèmes prosodiques : une approche cognitive.* Thèse de Doctorat d'Etat, Université de Provence

[5] Hirst, D.J. & Espesser, R. (1991) "Automatic modelling of fundamental frequency." *Travaux de l'Institut de Phonétique* 15

[6] Hirst, D.J. & Di Cristo, A. (in press) "A survey of intonation systems." in Hirst & Di Cristo *Intonation Systems : a Survey of Twenty Languages.* (Cambridge University Press; Cambridge)

[7] Liberman, M & Pierrehumbert, J (1984) "Intonational invariance under changes in pitch range and length." in Aronoff & Oehrle (1984) *Language Sound Structure* 157-253

[8] Martin, P (1983) "Real time fundamental frequency analysis using the spectral comb method." *Proc. Phon. Sci. X*, Volume 2, 284-287.

[9] Pierrehumbert, J. (1980) *The Phonology and Phonetics of English Intonation.* PhD thesis; MIT.

INTONATION CURVES - NORMAL AND DEVIANT

Y. Frank and T. Most

Tel Aviv University
School of Education

ABSTRACT

This study was designed to observe individual differences in the production of intonation by normal-hearing (NH) and hearing-impaired (HI) children. Three types of intonation curves (declarative, Wh- and Yes/No questions) were produced in orally read sentences by 18 HI and 10 NH children (age group 9-12). F_0 measurements were obtained at the midpoint of vowels along the sentences. The results show that NH like HI may have some deviation in so-called normal patterns.

1. INTRODUCTION

It is known that hearing-impaired speech is characterized by high fundamental frequency, pitch breaks and difficulties in the production of the usual language-related intonation patterns. These characteristics, apart from articulation errors, may contribute to the general lack of intelligibility in the speech of the hearing-impaired. There have been a number of studies employing different approaches, which have demonstrated significant differences between normal-hearing (NH) and hearing-impaired (HI) children in both perception and production of intonation [4,6,7,8,9,11,12,13,14].

In many of the studies acoustic analysis concentrated on group values of NH children's intonation patterns which were compared with the values of HI children. In this study special interest was focused on the intonation curves produced by each of the normal-hearing children in order to observe possible individual differences.

2. METHOD

Eighteen HI children, 9-12 years old (mean age 10.4 years) and 10 NH children (mean age 9.9 years) were recorded in quiet rooms in their respective schools. The HI subjects who attended special classes for HI students in a regular school, were orally educated and were reported by their teachers to be good readers. Their average hearing loss in the better ear was 99.2 dB, SD 8.9 dB. The NH children were randomly selected. They had no hearing problem.

The recorded material consisted of 18 orally read sentences as follows: 3 short (5-7 syllables) and 3 long (11-14 syllables) sentences of each type: declarative, Wh- and Yes/No questions. The short sentences were taken from a previous list of one of the authors [2] and were extended by a modifier and a content word. Each sentence ended with appropriate punctuation and was written separately on a card. All of the children, NH and HI, were familiar with the function of punctuation. Each child read the sentences quietly to himself, and then he read the sentences aloud in random order. The recording was done by a SONY Casette Recorder TCM-848 with the microphone held at a steady distance of 12-15 centimeters from the mouth. Spontaneous speech was elicited by a picture of a simple family-scene. The children were instructed to describe some of the objects in the picture and then to ask the examiner questions about them. In this way we succeeded to get declarative sentences and also some questions from all the children.

The acoustical data consisted of 18 read and 4 spontaneous sentences. Measurement of F_0 was done at the midpoint of vowels, using a Kay Elemetrics Visipitch 6087 instrument. In this study only the data of the read sentences will be presented.

3. RESULTS

The mean fundamental frequency of the 18 sentences produced by the HI children was compared with the mean fundamental frequency of the NH children. The measured average F_0 of all the 18 sentences for the normal-hearing group was 247.73 Hz SD 26.10 Hz and for the hearing-impaired 316.73 Hz SD 38.03 Hz. The mean F_0 of the HI children was significantly higher (T=5.60, DF=24.70, p=.000).

In evaluating the intonation patterns, a grading system was used that presumed the possible configurations of the intonation curves (in relation to the 3 sentence types). It was found that there were no significant differences between the children of the two groups in the production of all the declarative sentences, neither the short nor the long ones. Also there was no significant difference in the intonation production of the short Wh-questions, while in the longer Wh-questions the difference was significant at the 5% level (T=2.46, Df=25.87, p=.021). Even so, statistical analysis of the measured values of all the Wh- questions together did not show a significant difference between the HI and the NH children. Only all the Yes-No questions, short and long, showed a significant difference (T=3.60, df=25.99, p=.000) between the NH and the HI children.

Until now group differences were presented. Examples of productions of individual children that may be informative are shown in the following two figures.

Fig. 1 shows two NH individual expressions of a short declarative sentence as compared to the mean F_0 pattern of the HI and the mean F_0 pattern of the NH. While the mean patterns of both groups show a declination, the individuals (A & B) are very different. B expresses the sentence in the usual way and A expresses it with a rising contour.

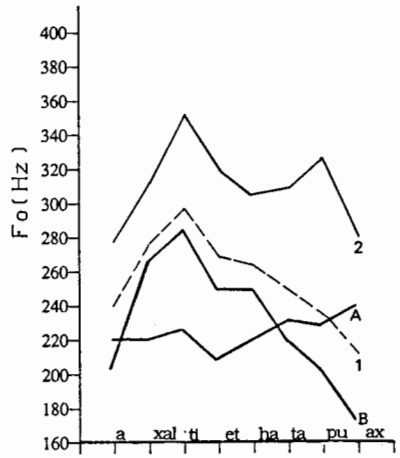


FIGURE 1: "axalti et hatapuax" (I ate the apple). 1=the average F_0 curve of NH. 2=the average F_0 curve of HI. A and B: different versions of 2 NH children

Fig. 2 shows 3 HI individual expressions of the same sentence as compared to both group means. In this example all the expressions have a falling contour.

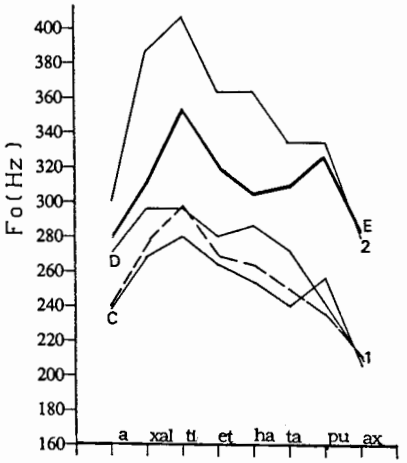


FIGURE 2: "axalti et hatapuax (I ate the apple). 1=the average F_0 curve of NH. 2=the average of HI. C,D,E different versions by 3 HI children.

4. DISCUSSION AND CONCLUSION

Higher average F_0 of HI children is a known phenomenon [9], although the measured values across the publications differ. It may be that different techniques for the measurement contribute to the varied values. In this study the mean F_0 of HI children is higher than that of the normal-hearing children of the same age group.

The evaluation of the patterns of the intonation curves showed that the Yes/No questions, short and long, were the most difficult to produce for the HI children. This result is in agreement with previous findings [13,2,91]. In the Hebrew language this type of sentence is identified by rising intonation only without any other syntactic cue. This type of production requires voice control and preplanning and therefore, may be very difficult for the HI child. Wh-questions allow for much more freedom of choice: after expression of the question word the rest of the sentence may show either a declination or a rise. Since HI children in this age group are able to produce the rise in a single word [2] they succeeded in executing this type of sentence when it was short. In the latter part of the longer sentences F_0 deviations may have occurred.

Declination of F_0 at the end of simple, neutral, sentences is considered to be a basic way of the intonation curve [1]. Observation of the mean F_0 curve of the declarative sentences, produced by the NH children as a group, confirmed this statement. However, when examining the individual curves of these NH children it was found that 2 out of the 10 children finished all their sentences with a rise of F_0 . One child exaggerated in pronouncing a final velar R, changing the declination to a rising curve, another one produced only flat and monotone curves. It is difficult to decide whether this phenomenon is "situation dependent" [10], a habit of expressing their wish to arouse attention, or it may be a language problem (they produced statements the same way in the spontaneous sentences).

Deviations in intonation patterns like these are generally attributed to the intonation production of HI children [7,5,9]. By grading the individual curves of the NH and those of the HI with the same expectations, according to the F_0 , the results showed no significant difference between the children in the production of simple read declarative sentences. The deviations in the intonation curves in the speech of HI children, as well as in the speech of the NH children, may be due to individual characteristics of expression and not necessarily to their hearing loss. This may be a way to explain the variations in intonation production by HI children with a similar hearing loss.

In applying intervention strategies for the intonation improvement of HI children, one should consider the possibility that deviant production of an intonation curve may be due to other causes beside their hearing loss.

ACKNOWLEDGEMENT

This study was supported by a grant from the Israel Academy of Science and Humanity.

4. REFERENCES

[1] COOPER, W.E. and SORENSON, J.M. (1981), *Fundamental frequency in sentence production*, Springer Verlag, New York, Heidelberg.

[2] FRANK, Y., BERGMAN, M., TOBIN, Y. (1987), Stress and intonation in the speech of hearing impaired Hebrew speaking children. *Language and Speech, 30* (4).

[3] FRANK, Y. (1989), Deviance in the intonation patterns: A contrasting study of the normal hearing vs. hearing impaired Hebrew speaking Israeli children in Tobin, Y. (ed.). *From Sign to Text: A semiotic view of communication*, John Benjamins, Amsterdam/ Philadelphia.

[4] LEVITT, H., SMITH, C.R., STROMBERG, H. (1976), Acoustical, articulatory and perceptual characteristics of the speech of deaf children, in Grant, G. (ed.) *Proceedings of the speech communication seminar*, Wiley, New York.

[5] LING, D. (1976), *Speech and the hearing impaired child: Theory and Praxis*, Alexander Graham Bell Ass., Washington, D.C.

[6] McGARR, N.S. and OSBERGER, M.J.(1978), Pitch deviancy and intelligibility, *Journal of Communication Disorders, 11.*

[7] MONSEN, R.B., (1979), Acoustic qualities of phonation in young hearing-impaired children, *Journal of Speech and Hearing Research, 22.*

[8] NICKERSON, R.S. (1975), Characteristics of the speech of deaf persons, *Volta Review.*

[9] OSBERGER, M.J., McGARR, N.S. (1982), Speech production characteristics of the hearing impaired. In Lass, N.J. (ED.), *Speech and language: Advances in basic research and practice*, New York: Academic Press.

[10] UMEDA, N. (1982), "F_0 " declination is situation dependent, *Journal of Phonetics, 10*, 279-290.

[11] PARKHURST, B. and LEVITT, H. (1978), The effect of selected prosodic errors on the intelligibility of deaf speech, *Journal of Communication Disorders, 11.*

[12] RUBIN-SPITZ, J. and McGARR, N.S. (1990), Perception of terminal fall contours in speech produced by deaf persons, *Journal of Speech and Hearing Research, 33.*

[13] STARK, R.E. and LEVITT, H. (1974), Prosodic feature reception and production in deaf children. Paper presented at the *87th meeting of the Acoustical Society of America.*

[14] SUSSMAN, H.M. and HERNANDEZ, M. (1979), A spectrographic analysis of suprasegmental aspects of the speech of hearing impaired adolescents, *Audiology, 5.*

ROLE OF BASAL GANGLIA FOR SPEECH RATE CONTROL : OBSERVATIONS FROM PATHOLOGY

C. Chevrie-Muller, M.T. Rigoard, C. Arabia and G. Chevaillier

INSERM, Laboratoire de Recherche sur le Langage
Paris, France

ABSTRACT - Speech rate was measured by having 29 patients with basal ganglia dysfunction (BGD) read a list of words. The patients showed, when compared with 10 controls : (i) a wide range in the figures for total duration, total word time (TWT), total pause time, mean, SD and variation coefficient of pause (VC), (ii) TWT significantly shorter, (iii) Pause VC higher. Intra-subject pause variability was a common symptoms in patients suffering from BGD.

The role of structures localized in basal ganglia for the control of speech rate has been clearly attested. Grewel (4) describing speech impairment associated with parkinsonism indicated that such patients employ extra long pauses and that the duration of each syllable is usually greater than normal. A particular speech behavior, i.e. uncontrolled rapidity, has also been noted and referred as "propulsive rate" (7), as "short rushes of speech" (3) or as "accélération paroxystique de la parole" (2). It seems, nevertheless, that not all patients with a Parkinson's disease diagnosis demonstrate a single "typical" speech impairment. According to authors like Canter (1) and Sarno (7) the rate of speech may be either fast or slow. In a previous study one of the authors of this paper found, in a sample of 81 patients with Parkinson's disease, 38 % with a rapid speech rate, 5 % with a slow one and the others with a normal rate (8). When, in place of clinical data, the effect of a stimulation of the thalamus ventro-lateral nucleus (in the course of stereotaxic operation for parkinsonism) is considered ; this was not always identical, both speech arrests and speech acceleration might be observed (5). The purpose of the present study was to examine, using measurement of phonation and pause time, whether or not patients with speech disorders related to Parkinson's disease were homogeneous. In addition, the same method was used to characterize speech rate in another group of patients with a symptomatology close to that of Parkinson patients, i.e. Progressive Supranuclear Palsy.

1. MATERIAL AND METHODS

1.1. Subjects

3 groups of French speaking male subjects entered the study : 1/ 22 patients with idiopathic Parkinson's disease (PD) who had never received L-Dopa therapy (or other specific treatment) with an age range of 50 to 79 (mean = 63 SD = 8) - 18 of the 22 patients belonged to the sample from which a perceptive description of speech was reported in a previous paper (8) ; 2/ 7 patients with typical features of Progressive Supranuclea Palsy (PSP), and especially no effect of L-Dopa, aged between 50 and 72 (mean = 61 years 4 months, SD = 5 years 1 month), and 3/ 10 controls, ranging age from 54 to 61 (mean = 59, SD = 4).
All subjects were at a cognitive and educational level which allowed reading without problem, except of a motor origin.

1.2. Material

The subjects were asked to read a list of words printed in a column on a sheet. This material was part of a more extended protocol including sentence reading, words repetition, automatic speech (numbers, months) and self-formulated speech. Speech was recorded in a sound-proof room, at the same time as an electroglottogram, using a two-channel tape recorder (REVOX A77).

1.3. Analysis of temporal patterns

Digital conversion of speech signal was performed at a sampling rate of 2 KHz Measurements were made from the integrated acoustic signal by a single operator (for all measures) using a mouse to determine the word limits on the screen. The time data were stored and further statistics obtained from the file. Statistical comparisons between groups (Student t test) were obtained for the following measures : total duration of the reading of the words list (TD), total word time (TWT), total pause time (TPT), ratio TWT/TD, mean of pause duration (MPT), standard-deviation of pause duration (SDPT), variation coefficient (VC = SDPT/MPT).

2. RESULTS

2.1. Comparison between Parkinson's patients and controls

The comparison of the group means comparison showed significantly shorter mean for total phonation time (TWT-t = 2.53 ; fd = 30 ; $p<.05$). The only other significant difference was for VC, i.e., on an average, a higher VC in the PD group than in controls (VC-t = -2.46 ; fd = 30 ; $p<.05$). There was actually a great heterogeneity in the pause duration of a given patient which was independent of the pauses duration as a whole (CV corresponded to the ratio SD/mean). It was striking that for all other measures the mean was close to PD and controls, but in the first group the range was very wide showing that there was an important variability in patients speech behavior.

2.2. Comparison between PSP patients and controls

The heterogeneity in patients pause duration for a single subject was confirmed in this group. When compared with controls SDPT and VC were, on the average, greater (SDPT-t = -3.38 ; fd = 16 ; $p<.01$. VC-t = -3.84 ; fd = 16 ; $p<.01$). Total word duration mean was shorter than in controls, but no significant difference was shown (note the PSP group small size). As noted for PD patients, speech behavior was different among patients with a higher variance than in controls.

2.3. Comparison between PD and PSP patients

No significant differences could be shown except for a higher SPDT in PSP than in PD patients (SDPT-t = -2.60 ; fd = 27 ; $p<.05$).

3. DISCUSSION

3.1. The data obtained in speech rate analysis in 2 groups of patients with basal ganglia dysfunction demonstrated a wide range in all parameters describing speech rate from very slow to very fast. Such a high variance in a group of patients is in agreement with data obtained in studies where self-formulated speech was juged by listeners (4, 7, 8). As far as the total pause duration (TPT) and mean pause duration (MPT) were concerned, the range was almost equally distributed on both sides of a mean not very different from that of controls. But for word duration (TWT) the duration was on the average shorter than in controls. One explanation for this might be that, conversely to pauses, words could not be lengthened beyond a certain limit. The words shortening which is obvious in at least part of the patients is not in agreement with Grewel's description (4) ; the shortening of phonation time (associated with an opposite pause lengthening) in PD patients who benefited of L-Dopa therapy (6) had seemed also to indicate that

the neurotransmitter defect led to slowing of word articulation.
In any case, the differences between patients for speech rate needs to be explained. Further research should test the possible relationship between the type of speech impairment and the clinical, biological and neuroanatomical features. Rough significant correlations has been described between the severity of speech impairment and that of other neurological symptoms in PD patients (8).
3.2. The only specific disturbance that differentiated patients from controls when means comparisons were computed was the intrasubject heterogeneity of pause duration (higher VC). Recalculating variation coefficient from Mawdsley and Gamsu's data on pauses between digits, in a counting task, it appeared that after L-Dopa therapy the pause variation coefficient dramatically decreased ($t = 3.28$; $df = 19$; $p<.01$). The defect in neurotransmitter seemed to have a reversible effect on the pause duration variance (in a same patient).
In tasks involving a periodicity it seems that nigrostriatal structures are necessary for the regularity of the rhythm.
3.3. Patients with PSP demonstrated the same intersubject variability as PD patients, and the same, even at a higher level, intrasubject pause variability.

4. CONCLUSION

There is a need for further research taking into consideration speech rate and rhythm characteristics in other modalities such as reading of sentences or paragraphs, repetition of words or sentences, spontaneous speech.
A first practical conclusion may be that any research on control of speech movements, of articulation or of prosody must be performed using either a sufficient number of subjects, or groups of patients defined on precise criteria (especially concerning speech rate and rhythm). A given medical diagnosis does not imply a single speech modification.

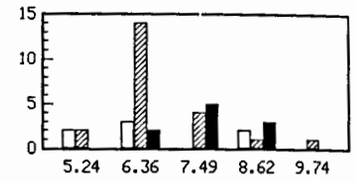


Fig. 1
Total phonation time (TWT) - Filled bars represent controls, stripped bars PD patients and clear bars PSP patients. Figures on X-axis correspond to the lower limit of each five classes of the histogram, on Y-axis they correspond to number of subjects.

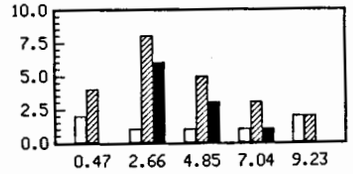


Fig. 2
Total pause time (TPT) - Same definition as in Figure 1

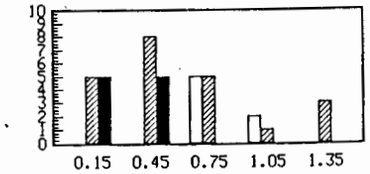


Fig. 3
Pauses variation coefficient (VC) Same definition as in Figure 1

Tableau I

Duration measurement (in second) - see definitions and comments in the text

	PD n=22	PSP n=7	CONT n=10
TD m	12,11	12,98	12,74
σ	4,18	4,54	2,00
TWT	7,27	7,13	8,32
	1,21	1,49	0,76
TPT	4,84	5,85	4,43
	3,32	3,95	1,66
TWT/TD	0,64	0,61	0,66
	0,15	0,20	0,08
MPT	0,44	0,53	0,40
	0,30	0,36	0,15
SDPT	0,25	0,54	0,15
	0,19	0,38	0,06
VC	0,79	0,98	0,42
	0,45	0,21	0,18

REFERENCES
(1) CANTER, G.J. (1963) "Speech characteristics of patients with Parkinson's disease, I. Intensity, pitch and duration", Journal of Speech and Hearing Disorders, 28, 221-229.
(2) CLAUDE, H., DUPUY-DUTEMP, S. (1921) "Forme céphalique du syndrome de Parkinson avec tachyphémie troubles oculaires et sympathiques", Revue Neurologique, 28, 716-720.
(3) DARLEY, F.L., ARONSON, A.E., BROWN, J.R. (1975) "Motor Speech Disorders", Philadelphia : Saunders.
(4) GREWEL, F. (1956) "Dysarthria in Post-Encephalitic Parkinsonism" Acta Psychiatrica Scandinavia, 32, 440-449.
(5) GUIOT, G., HERTOOG, E., RONDOT, P., MOLINA, P. (1961) "Arrest or acceleration of speech evoked by thalamic stimulation in the course of stereotaxic procedures for Parkinsonism", Brain, 84, 366-379.
(6) MAWDSLEY, C., GAMSU, C.V. (1971) "Periodicity of speech in Parkinsonism", Nature, 231, 315-316.
(7) SARNO, M.T. (1968) "Speech impairment in Parkinson's disease" Archives of Physical Medicine and Rehabilitation, 49, 269-275.
(8) SEGUIER, N., SPIRA, A., DORDAIN, M., LAZAR, P., CHEVRIE-MULLER, C. (1974) "Etude des relations entre les troubles de la parole et les autres manifestations cliniques dans la maladie de Parkinson", Folia Phoniatrica, 26, 108-126.

ACKNOWLEDGMENTS
The authors wish to thank their neurologist colleagues (Neurology departments of the Hôpital Pitié-Salpêtrière) for information about the neurological symptoms of patients as well as S. Orsoni and C. Arragon for editing the text.

Main author adress
Dr Claude CHEVRIE-MULLER
INSERM - Laboratoire de Recherche sur le Langage
Hôpital de La Salpêtrière
47, bd de l'Hôpital
75651 PARIS Cédex 13
(FRANCE)

DURATIONAL PATTERNS IN THE SPEECH OF FINNISH APHASICS

Pirkko Kukkonen

Department of General Linguistics
University of Helsinki, Finland

ABSTRACT

In this study, consonant and vowel durations were compared in normal and aphasic speech. Some aphasic speakers produced longer sounds than normal. The durations in aphasic speech were more variable than in normal speech, but on the average, the aphasics were able to produce the length opposition. For vowels, the increase in variation seemed to depend on intrinsic factors, whereas for consonants also the effects of surrounding sounds had to be considered. The implications of these findings to speech production models were discussed.

1. THE RESEARCH QUESTIONS

The present study addresses questions of timing and the neurophysiological programming of speech. There should not be any differences in the durational patterns due to aerodynamic factors between normal and aphasic speech, whereas different types of deviations should appear in the patterns due to neuromuscular constraints in aphasic speech.

The questions addressed in the present study were whether or not the aphasics were able to produce the length opposition, whether or not they produced longer sounds than normal, and whether or not the durations in aphasic speech were more variable than in normal speech. Factors contributing to the variation for durations were discussed.

2. MATERIAL AND METHODS

The acoustically analyzed words were elicited in a repetition test presented to eleven aphasic speakers and four age-matched control subjects. The aphasic speakers were accepted on a first come, first served basis. All the subjects were male, right-handed native speakers of Finnish. The etiology of aphasia and the lesion localization varied, and the time post onset of aphasia was between 1,5 months and 12 years. Detailed backgound information is presented in Kukkonen [1]. The present study is based on case descriptions, and symptom dissociations are searched for.

The phonetic composition of the test items was systematically alternated. The duration of the eight Finnish vowel phonemes in the first syllable was determined, as well as the duration of word-initial and word-medial consonants /ptksln/. For word-initial stops, the voice-onset-time was measured. In Finnish there is a phonological length opposition for word-medial vowels and consonants. The items were isolated words, and the analysis was based only on correct repetitions.

3. RESULTS

3.1 First-Syllable Vowels

3.1.1 Length Opposition

The means for short and long vowels were clearly different, and a figure obtained by dividing the duration of a long vowel phoneme by the duration

of the corresponding short vowel phoneme was approximately the same for both the control subjects and the aphasic speakers. The comparisons revealed two subjects whose vowels were on the average longer than normal (Subjects 4 and 14). This lengthening came into surface in a similar manner in all the vowels, and no differences were observed between, for example, closed and open vowels, or labial and illabial vowels.

The realization of the length opposition was further characterized by comparing the duration of the shortest long vowel with the longest short vowel for each speaker and each vowel phoneme. In the control data, there was always a "margin" between the short and long vowels. On the average, the duration of the margin was 61 - 85 ms for the control subjects. One of the aphasic speakers did not differ from the comparison group. For the other aphasic subjects, there was no margin between the short and the long vowels. The deviations were most notable for Subjects 6 and 11.

It was the easiest to produce the length opposition for vowel /æ/, and it was the most difficult to produce the opposition for vowels /y/ and /u/. These closed labial vowels are usually produced with lip protrusion in Finnish. Some differences were noted between the aphasic speakers, but there were not enough data to establish which of these differences were significant.

The deviations observed in the realization of length opposition were correlated with the increase in the amount of variation for vowel durations. All the subjects were obviously aiming at the correct phonological target phonemes.

3.1.2 Variation for Vowel Duration

In order to compare the amount of variation in the subjects' speech, the coefficient of variation was determined for each vowel. Furthermore, the means for these coefficients was calculated. As the variance for durations depends on the length of the segment--the longer the segment, the higher the variance--the present analysis was based on the logarithm of the vowel duration. For most of the aphasics, the c.v. was higher than for the control subjects. There were two aphasics (Subjects 9 and 11) with a remarkably high c.v.

3.1.3 Factors Behind the Increased Variation

It was expected that the vowels should be shorter between two stops than between two sibilants [3]. Different word structures were treated separately. First, the duration of the vowels occurring between two stops was compared with the duration of the vowels located between two sibilants (or between a stop and a sibilant). The effects of the surrounding consonants came mostly out as predicted, and the effects were similar for both the controls and the aphasic speakers. A comparison of word pairs (e.g. teettææ and seestyy) also supported the conclusion: there are no apparent differences between the controls and the aphasics in how the surrounding consonants affected the vowel duration.

The effects of word structure and word length were difficult to tell apart because the shorter words were of a different word structure than the longer words. This analysis did not reveal differences between the aphasics and the controls: the differences between the word structures were equally clear for all the speakers.

When discussing intrinsic duration, the different articulatory components should be compared. Ladefoged & al. [2] propose that vowel production is accounted for by three components: a posterior constriction, an anterior constriction, and a labial constriction. For all the subjects, and especially for the aphasics, the difference between short and long /æ/ (/æ/ requirs the least constriction of the Finnish vowels) was usually rather long. There were no clearcut differences such that

some patients would fail with posterior tongue movements and others with anterior tongue movements. However, one aphasic subject seemed to find it more difficult to control vowel length for labial vowels than for illabial vowels. We could assume that both the execution of the constrictive movements and the coordination of these movements may lie behind the observed deviations. The vowel centralization observed for one of the aphasics supports the conclusion that it may be the reduction in muscular movements that lies behind the observed deviations [1, 5].

3.2 Word-Initial Consonants

The voice-onset-time of the word-initial stops and the segment duration of other consonants were measured.

The duration of VOT depended on the stop's place of articulation. The VOT was the shortest for /p/, and the longest for /k/. In this respect, there was some variation between the subjects, but no differences between the controls and the aphasics were observed. The effects of the following vowel and of the word structure were analyzed but no effects were found.

In aphasic speech, the VOT was never lengthened, whereas the word-initial consonants tended to be longer than normal in the speech of certain (nonfluent) aphasics. Especially for the nonfluent aphasics, the VOT was shorter than normal. Thus, there may be a reciprocal relationship between stop closure and VOT [4].

3.3 Word-Medial Consonants

3.3.1 Length Opposition

In the comparison data, there was a "margin" between the longest single consonants and the shortest geminate consonants. The exact duration of this margin varied depending on the speaker and the consonant in question. On the average, the margin was the shortest for /n/ (45 ms) and for other resonant consonants, and the longest for /s/ (118 ms). Three out of four control subjects produced long margins, whereas one control subject produced margins that were often remarkably shorter than the average margins in the control data.

Most of the aphasic speakers also produced a margin between the longest single consonants and the shortest geminate consonants, but this margin was on the average shorter than in the control data. For two of the aphasic speakers, there was overlap between the duration of single and geminate consonants. For these speakers the coefficient of variation for consonant duration was remarkably higher than normal.

3.3.2 Increase in Variation for Consonant Durations

The variation for durations of the word-medial consonants was characterized by the coefficient of variation. In order to be able to reliably compare consonants of different length, the analysis was based on the logarithm of the duration. As compared to the control subjects, most aphasic speakers showed some increase in the amount of variation for consonant duration, and for two speakers the c.v. was remarkably high. These were the same subjects for whom there was overlap between the duration of single and geminate consonants.

3.3.3 Factors Behind the Increase in Variation

According to Lehtonen [3], the consonants are longer after labial vowels than after illabial vowels. The effect of labiality was very weak in the present data, and a statistical analysis did not support Lehtonen's conclusion.

The word-medial consonant preceeded by a short vowel is longer than a consonant preceeded by a long vowel [3]. For the control subjects, the above rule was true in general. There was some tendency for stops /pk/ to obey the rule more often than for the other consonants. One of the aphasics (Subject 8) produced longer consonants

after long vowels than after short vowels.

Voiceless plosives are the longest consonants, followed by fricatives and resonant consonants [3]. When different places of articulation are compared, labial consonants are on the average longer than dental and velar consonants [3].

In the present data, the manner of articulation had a stronger effect on consonant duration than place of articulation. For most of the speakers, resonants were on the average shorter than obstruents. For the control subjects, all the short consonants were shorter than the long consonants. For two of the nonfluent aphasic speakers, the geminate resonants were shorter than single obstruents.

The sibilant requires more sophisticated motor control than other consonants. In the present data, single stops were on the average longer than /s/, but the geminate /ss/ was often longer than geminate stops.

The comparison of different places of articulation (whether or not the labial sounds are longer than dental and velar sounds) did not give systematic results. One speaker had very long OFTs in the word medial position. His speech was not, however, distorted.

4. DISCUSSION

The length opposition was preserved in aphasic speech albeit some aphasics experienced difficulties with controlling the duration and therefore occasionally violated the length opposition.

Voice-onset-time is among the variables affected by aerodynamic factors. The duration of VOT seemed to be conditioned by the duration of the occlusion of the stop--the subjects with considerable consonant lengthening produced short voice-onset-times.

The present data gave only some hints to the processes between the selection of the phonological target and the aerodynamic processes. Lengthening was similar for both consonants and vowels, and the word structure did not have an effect on it. For vowels, the increase in variation for durations was not explained by the surrounding sounds or by word structure. Rather, there was some evidence for differing effects of the articulatory components of vowel production. For consonants, not only intrinsic factors but also for example the duration of the preceeding vowel should be considered.

Further evidence for different components of the articulation process was reported by Kukkonen [1] in connection with the different error types (some patients deleted word-initial consonants, some distorted them, and still others commited substitution errors that were not distortions).

The results point out that a comparison of acoustic properties of normal and deviant speech is a promising testing ground for theories of normal speech production. The findings will also have implications for the clinical classification of aphasics.

5. REFERENCES

[1] Kukkonen, P. 1990: *Patterns of Phonological Disturbances in Adult Aphasia.* Suomalaisen Kirjallisuuden Seuran toimituksia 529. Helsinki: Finnish Literature Society.

[2] Ladefoged, P. & R. Harsman & L. Goldstein & L. Rice 1978: Generating vocal tract shapes from formant frequencies. *Journal of the Acoustical Society of America* 64: 1027-1035.

[3] Lehtonen, J. 1970: *Aspects of Quantity in Standard Finnish.* Studia Philologica Jyväskyläensia VI. Jyväskylä: University of Jyväskylä.

[4] Suomi, K. 1980: *Voicing in English and Finnish Stops.* Publications of the Department of Finnish and General Linguistics of the University of Turku, No. 10.

[5] Ziegler, W. & D. von Cramon 1983: Vowel Distortion in Traumatic Dysarthria: A Formant Study. *Phonetica* 40: 63-78.

HEARING-IMPAIRED AND NORMAL-HEARING ADULTS' USE OF LOW-FREQUENCY CUES TO INITIAL FRICATIVE VOICING

L. Holden-Pitt, S. Revoile, and J. Pickett

Gallaudet University, Washington, D.C.

1. ABSTRACT

The contribution of various acoustic components to the perception of voicing in syllable-initial /f,s,v,z/ was investigated for hearing-impaired and normal-hearing listeners. Syllable-segment deletion and high-pass filtering were employed to eliminate potential voicing cues. Relative to the normal-hearing, the hearing-impaired group's voicing perception was more dependent upon low-frequency cues in the frication. For /vʌd/ and /zʌd/, with the frication segments deleted, above-chance fricative voicing perception, particularly by the normal-hearing, signified the existence of cues in the vowel stem.

2. INTRODUCTION

Perception of consonant voicing is often troublesome for persons with severe to profound hearing impairments, especially since this distinction is not easily conveyed through lip reading. During the past decade, some of our research efforts have involved the employment of various acoustic-signal enhancements to improve hearing-impaired persons' perception of spoken consonants. However, preliminary to the development of such enhancements, we must discover which acoustic elements in the speech signal can elicit particular consonant-feature distinctions.

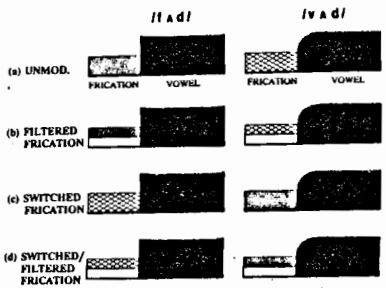
This study examined the contribution of various acoustic elements to the perception of voicing for the fricatives /f,s,v,z/ in the syllable-initial position of naturally-spoken /Cʌd/. We made controlled alterations to syllable acoustic-segments suspected to characterize voiced, that is /v/ and /z/, versus voiceless, /f/ and /s/, fricatives. To determine the perceptual utility of syllable segments for cuing fricative voicing, systematic syllable-modifications were performed to either eliminate or place in competition various components of the acoustic signal.

3. METHOD

3.1. Stimuli

The core set of stimuli consisted of 10 utterances each of /fʌd/, /sʌd/, /vʌd/, and /zʌd/, spoken citation-style by an adult male. Consonant-vowel boundaries for the utterances were established from the digitized waveform displays, guided by syllable segmentation criteria described in Revoile et al. [2]. Temporal and spectral measurements were made for the utterances' fricative and vowel segments, to enable examination of acoustic characteristics differing between voicing cognate syllables. Inspection of these attributes targeted the presence of a low-frequency component in the frications, and the vowel onset transitions as the most apparent indicators of acoustic difference between the voicing cognates. Thus, our syllable modifications were designed to examine the relative importance of these low-frequency components (i.e., potential voicing cues) by isolating or removing these elements, and by placing them in direct competition.

Fig. 1. Schematics of /fʌd/ and /vʌd/ stimuli per test condition.



From the 40 unmodified utterances, conditions of syllable-alteration (Figure 1) were prepared by computer manipulation of the waveform segments. The unmodified (panel a) utterances formed the basis for development of the other conditions. The rounded upper-left edge of the vowel symbol following the voiced frication in /vʌd/ signifies the presence of characteristic vowel-onset transitions. In the condition of filtered frication (panel b), the frication segments were high-pass filtered (1 kHz cutoff), as indicated in Figure 1 by the clear lower region in the frications. The 1 kHz filter cutoff was selected to eliminate the low-frequency spectral information present predominantly in the voiced frications. The next condition, switched frication (panel c), involves the exchange of frication segments between the voiceless and the voiced initial-fricative syllables. Between panels (a) and (c), note that the frication from the unmodified /vʌd/ -- represented by the cross-hatching -- has been appended to the vowel stem of the original /fʌd/, and vice versa. This alteration was intended to produce competition between voicing cues residing in the frication and those in the vowel stem. In panel (d), switched/filtered frication is a combination of the high-pass filtering and the switching of frications, carried out to examine whether the cue-competition effect expected with the switched frication stimuli would be nullified. Stimuli for a final condition, frication deleted (not shown in Figure 1), were developed by omitting the frications from the unmodified stimuli. This condition was intended to gauge the sufficiency of fricative-voicing cues remaining in the vowels.

3.2. Procedure

The 40 utterances in each condition were randomly presented in single-interval identification trials. Listeners' responses were limited to "FUD", "SUD", "VUD", or "ZUD". No feedback of correct response was provided. Stimuli were presented to each hearing-impaired subject's better ear at listener-determined most comfortable listening levels (MCL), using the procedure described in Revoile et al. [1]. Normal-hearing listeners were presented the stimuli at 73 dB SPL. At least five 40-syllable blocks per condition were tested per listener. Tests were administered in random order throughout 24 one-hour listening sessions.

3.3. Subjects

Twenty-two hearing-impaired and 10 normal-hearing young adults from Gallaudet University participated as paid listeners. The hearing-impaired listeners had tone-threshold averages ranging from 34 to 82 dB HL, with a median of 54 dB. Pure-tone threshold contours for these subjects were classified as either flat (n = 8) or sloping (n = 14). All subjects attained at least 70% correct fricative-voicing recognition for the unmodified test utterances.

4. RESULTS AND DISCUSSION

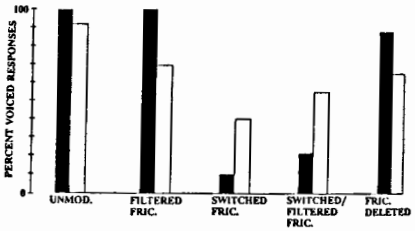
Percent correct voicing scores were calculated separately for the voiced, and the voiceless, initial-fricative syllables -- place of articulation errors disregarded. Mean voicing scores per listener were calculated for each test condition, arcsin transformed, then submitted to a repeated-measures ANOVA. Tukey's *hsd* was used for pairwise comparisons of condition means. A criterion alpha level of .05 was used for all tests of statistical significance. Interactions of listener group, fricative voicing, and test condition dictated that analyses be conducted separately within each listener group for the voiced and for the voiceless initial-fricative syllables.

4.1. Voiced Fricatives

Figure 2 shows percent "voiced fricative" responses for the /vʌd/ and /zʌd/ syllables. Fricative voicing recognition for the unmodified utterances approached 100% for the normal-hearing and the hearing-impaired listeners. While the pattern of response across test conditions was generally similar for the two listener-groups, magnitudes of the modification effects did differ between groups. Also, the hearing-impaired group's average standard deviation per condition (7%) was greater than the 2% seen for the normal-hearing -- evidence of greater intra-group performance variability among the hearing-impaired listeners.

The effect of filtering the frications' low frequencies can be observed by comparing voicing perception in the filtered frication condition with that for the unmodified utterances. While the normal-hearing group showed virtually no reduction in fricative voicing perception for the filtered stimuli, the hearing-impaired group's mean fricative voicing score of 69% represents a significant effect from the filtering. The importance of the frications' low frequencies to voicing perception is also exemplified by the similarity in the hearing-impaired group's performance for the conditions of filtered frication versus frication deleted. The 4% differential indicates that acoustic elements remaining in the frications after filtering did not contribute significantly to fricative voicing perception.

Fig. 2. Perception of VOICED fricatives by the normal- (black bars) and impaired- (clear bars) hearers.



Results from the switched frication condition show for both listener groups that appending the voiceless frications to the vowel stems from /vʌd/ and /zʌd/ greatly reduced the perception of voicedness. The normal-hearing group's identification of these hybrid stimuli as voiced in only 9% of the cases represents a near-complete domination of the voicing cues in the voiceless frication over the cues to voicedness in the vowel stem. For the hearing-impaired group, the perceptual decline of voicedness from the 92% in the unmodified condition to the 39% in the switched frication condition again shows that these concatenations of voiceless frications to vowel stems from voiced fricative syllables are identified /fʌd/ and /sʌd/ -- the appended frications dictating the perceived voicing value. Filtering the switched frications produced no

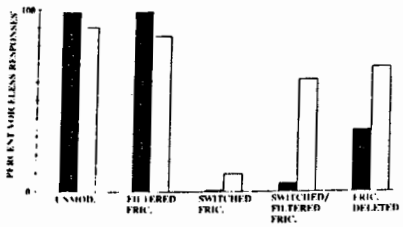
significant recovery from the perceptual domination of the frication segments for either group, as apparent from the increases of no more than 13% in fricative voicing scores for switched/filtered frication over the switched frication.

Finally, relative to perception for the unmodified stimuli, deletion of the frications significantly reduced voicing scores, though fricative voicing identification remained above-chance for both listener groups. This good voicing recognition in the absence of the /v/ or /z/ frications supports the utility of vowel stem characteristics for cuing fricative voicing.

4.2. Voiceless Fricatives

Figure 3 displays results for perception of the /fʌd/ and /sʌd/ syllables. Mean scores for the unmodified condition are 90% or better for both listener groups. The effect from high-pass filtering the voiceless frications was negligible, as relatively little spectral content existed in the lower spectral region of these voiceless fricatives.

Fig. 3. Perception of VOICELESS fricatives by the normal- (black bars) and impaired- (clear bars) hearers.



The switched frication condition again produced a dramatic reversal in fricative voicing perception, now with voiced frications appended to the vowel stems of /fʌd/ and /sʌd/. Neither normal-hearing nor hearing-impaired listeners provided many "fud" or "sud" responses to these stimuli. Thus, fricative voicing cues in the vowel stem again proved no match for those in the frication. For the hearing-impaired group, filtering low frequencies from the switched voiced frications did produce a significant release from this voicing reversal effect. For the normal-hearing group, higher-frequency components in the appended /v/ and /z/ frications were sufficient to sway listener responses toward voiced fricatives. Deletion of the /f/ and /s/ frications produced a significant decline in fricative voicing perception, when compared to performance for the unmodified utterances, though this reduction was more pronounced for the normal-hearing than the hearing-impaired group.

5. SUMMARY

The relative low-frequency energy in the frications of /vʌd/, /zʌd/ versus /fʌd/, /sʌd/ considerably influenced fricative voicing perception for the hearing-impaired listeners. When placed in competition, cues in the frication dominated those in the vowel stem, for prompting fricative voicing. However, in the absence of the frication segments, cues in the vowel stem were capable of eliciting voiced fricative percepts, particularly for normal-hearing listeners.

REFERENCES

[1] Revoile, S., Holden-Pitt, L., Pickett, J., and Brandt, F. (1986), "Speech cue enhancement for the hearing impaired: I. Altered vowel durations for perception of final fricative voicing", Journal of Speech and Hearing Research, 29, 245-255.

[2] Revoile, S., Pickett, J., Holden-Pitt, Talkin, D., and Brandt, F. (1986), "Burst and transition cues to voicing perception for spoken initial stops by impaired- and normal-hearing listeners", Journal of Speech and Hearing Research, 30, 3-12.

THE RELATIONSHIP BETWEEN MALOCCLUSIONS AND SPEECH DISORDERS : AN ACOUSTIC STUDY

M. Pettorino*, P. Diaco, A. Giannini*, A. Ferro****

*** Ist. Univ. Orientale, Fonetica Sperimentale, Napoli, Italia**
**** Ist. Clin. Odont. e Stomat., I Fac. di Med. Univ. Napoli, Italia**

ABSTRACT

A corpus of about 100 meaningful Italian words uttered by 36 normoccluded and maloccluded subjects has been analysed spectrographically. The results show that a direct relationship between different classes of malocclusions and speech errors does not exist.

1. INTRODUCTION

One of the questions which has always been and still is outstanding for orthodontists and speech pathologists is whether there is a relationship between dental malocclusions and speech disorders.

By far the greatest difficulty in this kind of research is to find a cause-effect relationship between a single dental anomaly and a particular speech impairment. In fact if on one hand "articulatory defects of speech may exist even though the dental occlusion is normal and, conversely, dental malocclusions may exist in person with normal speech" [3] (p. 921), on the other the greater the seriousness of disgnathic defects co-occurring in the same subject is, the greater his phonetic handicap will be [5] [8]. This is the reason why notwithstanding the great number of studies of the morphological aspects of the different malocclusions, as for instance vertical, anteroposterior and transversal relation ship of the jawbones, interincisal occlusion, overjet, openbite, spacing and crowding of the incisors, the lack of dental elements, the results have often been conflicting [1] [7] [2] [9] [6] [10] [4].

The aim of this research is to verify whether there is a close relationship between dental malocclusions and speech disorders.

Before facing this problem, we should draw some considerations. If dental malocclusions can be easily tested and then classified, the defect of speech is more difficult to identify. It can be recognized only perceptively. In fact the listener is the only one who can say whether a sound is similar to or different from a *normal* sound. His judgement, however, cannot go beyond a personal opinion which can be neither quantified nor experimentally verified, thus originating vague and often inaccurate classifications.

First of all we should say that two different auditory impressions must be the result of two different acoustic signals. However, the opposite is not always true, as two different acoustic signals, generated by different articulatory mechanisms, do not necessarily generate two different auditory impressions. This happens because each speech sound is a complex acoustic signal of which some components are vital, whereas some others, being redundant, can have various characteristics or even lack completely. Suffice it to say that compensatory articulatory movements are able to produce a sound which is perceptively accepted as *normal*.

At the light of what has up to now been said, it is possible to identify a speech sound as faulty only on the basis of the acoustic analysis of the signal. If we skip this step, observing directly the articulatory movements, at the most we will be able, thanks to sophisticate technologies, to reconstruct point by point the mechanism of the individual parts of the speech apparatus, but we will not be able to establish for certain whether such an articulatory mechanism affects the distinctive or the redundant components of the signal. Furthermore it must be borne in mind that the techniques employed nowadays can cause an emotional stress to the speaker that affects negatively the spontaneity of the utterance. This happens because they may be either tissue invasive owing to the attachment of lead pellets, of artificial palates, of electrodes and so on, or biologically unsafe because of radiation exposure.

2. METHOD

From the foregoing, it seems to us that we should start from the acoustic analysis of the signal, which allows to identify faulty sounds as well as to infer the incorrect articulatory movements that produced them. Many are the possibilities given by this method of analysis. In fact on a broad band spectrogram it is possible to deduce the behaviour of the vocal folds from the number and periodicity of the vertical striations and, consequently, to notice the presence of possible anomalies of the glottal pulse. Shiftings on the y-axis of the formants reflect the movements of the articulators and the shapes assumed by the supralaryngeal cavities. Formant frequencies are broad bands of energy represented on the spectrogram by clearly marked darkness areas. According to the different contextual situations, every speech sound has a particular formant pattern : any modification reflects an anomalous posture of articulators involving a change in place and manner of articulation. Nasality is represented on the spectrogram by a loss of energy expecially at the level of the second formant as well as in appearance of one or two extraformants in the low region of frequencies. So the spectrographic analysis allows us to say whether an oral articulation has been realized with an incomplete closure of the velopharyngeal port.

3. MATERIAL

A list of about one hundred meaningful Italian words has been prepared, where dental articulations [t d s z r l n ʦ ʣ] occurred in all phonological contexts. Also bilabials [p b m], labiodentals [f v], palatals [ʃ ʧ ʤ ʎ ɲ] and velars [k g] have been considered.

The list has been read in a silent room by thirty six speakers differently aged (7-9, 12-14, 17-19 years) selected by a clinical test from a total of 228 students. Nine of them were normoccluded subjects and twenty seven represented of the different classes of malocclusion (Class I, Class II, Class III). A structured questionnaire was used to obtain information about age, history of previous speech therapy and orthodontic therapy. All selected subjects had not received any treatment and all of them had normal hearing.

For this research a Nagra IV S recorder, a DSP Sona-Graph 5500 Kay and a computer HP Vectra have been employed. Of each word the broad band spectrogram (from 0 to 8 KHz) and the tracings of Fo, intensity and waveform have been obtained.

4. RESULTS

Subjects without any anomaly have been found in all categories. Table I summarizes the speech anomalies of

TABLE I. Types of articulatory disorders in speech acoustically diagnosed among normoccluded and maloccluded subjects.

SPEECH SOUNDS	DISORDERS			
	NORMOCCLUDED SUBJECTS	MALOCCLUDED SUBJECTS		
		CLASS I	CLASS II	CLASS III
dental stop [t] [d]	retroflex fricative		fricative	fricative affricate
dental fricative [s] [z]	palatal		palatal labiodental interdental whistled	
dental trill [r]	lateral trill		fricative retroflex	fricative
dental lateral [l]	lateral trill		lateral fricative	
labiodental fricative [v]			bilabial stop	
palatal affricate [tʃ] [dʒ]	dental			
palatal lateral [ʎ]				palatal stop
velar stop [k] [g]	fricative		fricative	fricative affricate
oral sounds		nasalized speech	nasalized speech	
voiced sounds		irregular glottal pulses		

normoccluded and maloccluded subjects. As we can see, most anomalies occur with dental articulations, but anomalies in velar stops, nasality and glottal pulses have been also noticed.

Figs. 1 - 4 show the spectrograms relative to the voices of normoccluded and maloccluded subjects.

The spectrogram of the word *sodo* (Fig. 1) uttered by a nornoccluded subject points out two different anomalies, regarding the fricative [s] and the stop [d]. The first one has a dental place of articulation because of a very strong signal starting from 6 kHz. In fact the acoustical signal of a dental fricative shows the highest frequencies and in this case we can see that the signal is cut at the upper edge of the spectrogram. In the meantime we can notice the presence of a strong signal between 3.5 KHz and 4.5 KHz due to a narrowing of the channel at the hard palate. So we can conclude that it is a palatalized dental fricative. As far as the stop is concerned we can say that it is a retroflex, because of an abrupt falling down of F3 and F4 of the adjacent vowels.

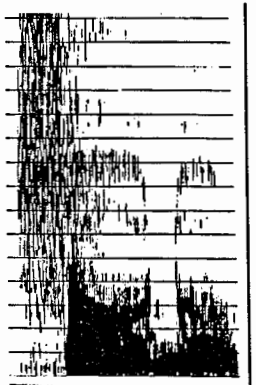


Fig. 1. Spectrogram of the word *sodo*

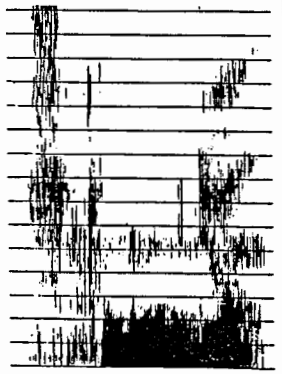


Fig. 2. Spectrogram of the word *studio*

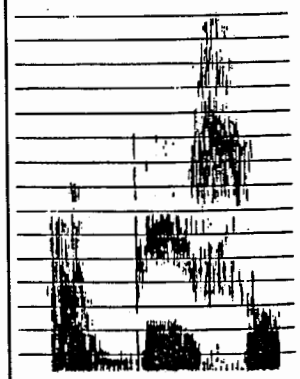


Fig. 3. Spectrogram of the word *avviso*

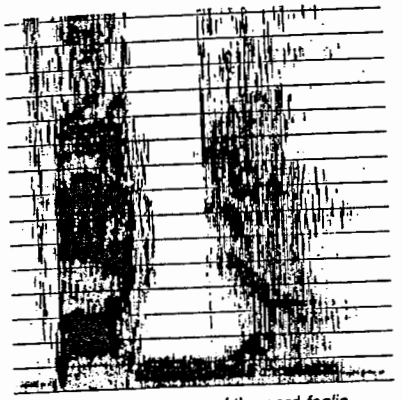


Fig. 4. Spectrogram of the word *foglia*

The spectrogram of the word *studio* (Fig. 2) uttered by a Class I malocclu-ded subject shows two anomalies, one relative to [s] having the same characteristics already seen in Fig. 1 and the other concerning the whole word, which is completely nasalized, as the presence of an extra formant at 2.5 KHz shows.

The spectrogram of the word *avviso* (Fig. 3) uttered by a Class II malocclu-ded subject, shows that [v] has been uttered as a stop (absence of signal followed by burst of noise), bilabial because of the F2 deviations of the adjacent vowels and voiced because of the periodical striations. Furthermore [s] shows in addition of the fricative signal, a whistled pure tone at about 4.5 KHz.

In the spectrogram of the word *foglia* (Fig. 4) uttered by a Class III malocclu-ded subject, both the absence of signal and F2 deviations show that [ʎ] is uttered as a voiced palatal stop [ɟ].

5. CONCLUSIONS

The data gathered in this experimental research point out that a direct relationship between different classes of malocclusions and speech errors does not exist. In fact the same speech sounds can give rise to different kinds of errors, aside from the kind of dental occlusion of the subject. [s] for instance is realized as palatal by a normoccluded subject, as either labiodental or whistled fricative by two Class II maloccluded subjects. Dental trill [r] is realized as a lateral trill by a normoccluded subject, as either fricative or retroflex trill, by two Class II maloccluded subjects.

As regards the different kinds of speech errors our results seem to suggest that fricatives tend to be realized at anomalous places of articulation; stops tend to be realized as affricates or fricatives; laterals and trills tend to be realized as fricatives. Moreover we have to say that the speaker's sex and age do not have any infuence on the occurrence of the different kinds of speech errors.

6. REFERENCES

[1] BERNESTEIN, M. (1954), "The relation of speech defects and malocclusion", *Amer. J. Orthod., 40,* 149-150.

[2] BLOOMER, H. H., (1958), "Speech as related to dentistry", *J. Mich. St. Dent. Soc., 40,* 258-266.

[3] BLOOMER, H.H. (1963), "Speech defects in relation to orthodontics", *Amer. J. Orthod., 49,* 920-929.

[4] DE SANTIS, M. (1986), *"Voce e linguaggio"*, Padova: Piccin.

[5] FAIRBANKS, G. and LINTNER M. A. (1951), "A study of minor organic deviations in functional disorders of articulation", *J. Speech Hear Disord., 16,* 273.

[6] MARX, R., (1965), "The circum-oral muscles and the incisor relationship. An electromyographic study", *Trans. Eur. Orthod. Soc.,* 187-201.

[7] MOORE, G. E., (1956), "The influence of the oral cavity on speech", *Br. Dent. J., 101,* 304-309.

[8] RENOCHE, G. de, *et al.,* (1983) "Il problema ortodonzia, deglutizione, fonesi nei soggetti neurolesi", *Saggi, IX,* 21.

[9] SNOW, K., (1961), "Articulation proficiency in relation to certain dental abnormalities", *J. Speech Hear. Dis., 26,* 209.

[10] SUBTELNY, J. D. *et al.,* (1964), "Comparative study of normal and defective articulation of /s/ as related to malocclusion and deglutition", *J. Speech Hear. Dis., 29,* 269-285.

PHONETIC AND PHONOLOGICAL LEVELS IN THE SPEECH OF THE DEAF

A-M. Öster

Dept of Speech Communication and Music Acoustics, Royal Institute of Technology, KTH, Box 700 14, S-100 44 Stockholm, Sweden.
Phone 46 8 7907557, Fax 46 8 7907854

ABSTRACT

The speech of eleven prelingually and profoundly deaf children, educated by sign-language, was videorecorded and was given a narrow phonetic transcription. The phonetic inventory of consonants used by the children in initial, medial and final word-positions was established. Analyses were also made to see what systematic deviations occurred for the speech sounds that the children could articulate correctly. The intention was also to get an opinion of average phonetic and phonological competence of this group of prelingually profoundly deaf children, with pure tone averages between 90-108 dB at .5, 1 and 2 kHz.

1. INTRODUCTION

Prelingually deaf children do not acquire speech spontaneously. They have to learn oral speech through visual information mainly and to rely on orosensorymotor control in maintaining speech movements. As the deaf child does not have any acoustic speech target to compare his own production with, his speech will be characterised by specific deviations and substitutions due to input limitations in speech perception such as auditory limitations and limited visibility of phonetic features and impacts of orthography and insufficient physiological control.

Despite the fact that prelingually deaf children have difficulties in producing normally articulated and auditorily acceptable speech some studies have reported that they can develop a phonological system through the limited information available, [2], [4], [5], [6], [8]. However, these systems will differ in some respects to those of normally hearing children.

Through a phonological assessment it can be determined to which extent an inadequate phonological system is obscured by phonetic deviations and the systematic deviant patterns can be identified. A detailed phonetic transcription, that describes the phonetic inventory and its application in different word positions, should form the basis of the phonological assessment.

In a study by Öster [6], it was shown that a deviant pronunciation in fact was an attempt to express a speech sound contrast. A child made a contrast between voiced and unvoiced bilabial stops but not through voicing. Instead the contrast was expressed by lip-protrusion in initial position and by the insertion of a neutral vowel in final positions. The training was then directed towards changing this inadequate way of expressing voicing contrast to improve the intelligibility of the child's speech.

A traditionally phonetic analysis describes the quality of a child's articulation of various speech sounds with no reference to their distinctive function in spoken language. Often distortions, substitutions and omissions are listed that show what the child is not capable of articulating. The sounds that the child articulates correctly are disregarded. Even if a child knows how to articulate a speech sound correctly, this does not imply that the usage is correct in his spoken language. Through a phonological assessment, on the other hand, it is possible to study systematic deviations in spoken language, of those speech sounds, which a child has shown to be capable of articulating. If these systematic deviations can be explained by limited phoneme perception in lip-reading, impact of orthography or insufficient physiological control valuable pedagogical information is obtained.

2. AIM OF THE STUDY

The intention was to investigate how phonetic deviations affect the phonological systems of deaf children. Assessments were made to establish which speech sounds most of the children could articulate, which of these sounds were applied correctly in their speech and what the substitutions and other deviations looked like.

3. SUBJECTS, PROCEDURES AND SPEECH MATERIAL

Eleven prelingually deaf children, educated by sign-language, participated in the study. One child was eleven years of age, while the others ranged from fourteen to seventeen years. Their pure tone averages were between 90-108 dB at .5, 1 and 2 kHz. The intelligibility of their speech varied from very poor to very good.

The children read a list of familiar words provided with stimulus pictures. The word list contained all Swedish consonants, which occurred at least twice in initial, medial and final position, if phonotactically possible. The videorecorded speech was given a narrow phonetic transcription, using the symbols of IPA and some additional diacritics. Some of those which Bush, Edwards, Luckau, Stoel, Macken and Petersen [1], Grunwell [3] and Roug, Landberg and Lundberg [7] have developed for the transcription of babbling and phonetic development in early infancy were used to transcribe those sounds in the speech of the deaf, which are not part of the IPA inventory. The phonetic inventory and phone distribution in the different word positions was established.

4. RESULT AND DISCUSSION

Figure 1 shows the number of children who, at least once in the material, controlled the articulation of each Swedish consonant correctly. The figure also shows the number of children who made correct use of their articulation in initial, medial and final word positions. In other words, the difference in heights between the two bars, representing each consonant, shows the number of children, who made phonological substitutions or deviations in some position. A big difference indicates that this speech sound is difficult for a deaf child to control in this position, for example /ʃ/. Five children could articulate that sound but no one controlled it in initial position, only three in medial position and four in final position. The types of deviations or substitutions observed in various positions are shown in figure 2, where it can be seen that a stop or fricatives produced at incorrect places of articulation were substituted for /ʃ/ in initial position.

The children controlled 70% of the articulation of Swedish consonants on the average but they could only make use of 43% of them in initial position, 50% in medial position and 50% in final position, which is shown in figure 1. This indicates a discrepancy between the children's phonetic and phonological competence. Some speech sounds, however, are in correspondence like /t/ in initial position, /p/ and /m/ in medial position and /ŋ/ in final position.

The influence of the visibility of articulation on speech acquisition of deaf children is obvious since those speech sounds that most of the children control are the unvoiced bilabial and dental stop, the unvoiced labiodental fricative and the lateral, which are all visually contrastive and easy to lip-read.

Figure 2 shows all deviant phonetypes in various positions of systematic deviations occurred for those speech sounds that the subjects could articulate correctly at least once in the speech material. Some of them probably represent different phonemes despite the phonetic similarity . For example, it can be assumed, despite the phonetic similarity to [b], that a child could make contrasts between /p/, /b/ and /m/ in initial position through lip-protrusion for /p/, devoicing for /b/ and voicing for /m/. Another child makes perhaps contrasts between /t/, /d/ and /n/ in final position, despite the phonetic similarity to [d], through a non-audible release, devoicing and through nasal air emission.

In Swedish /ɕ/ can be spelled as *tj, kj, k* or *ch* and /ʃ/ can be spelled as *sk, sch, sj, skj* and *ch*. The fact that two, and sometimes three, graphemes are pronounced as one sound is probably not obvious to some children. The impact of

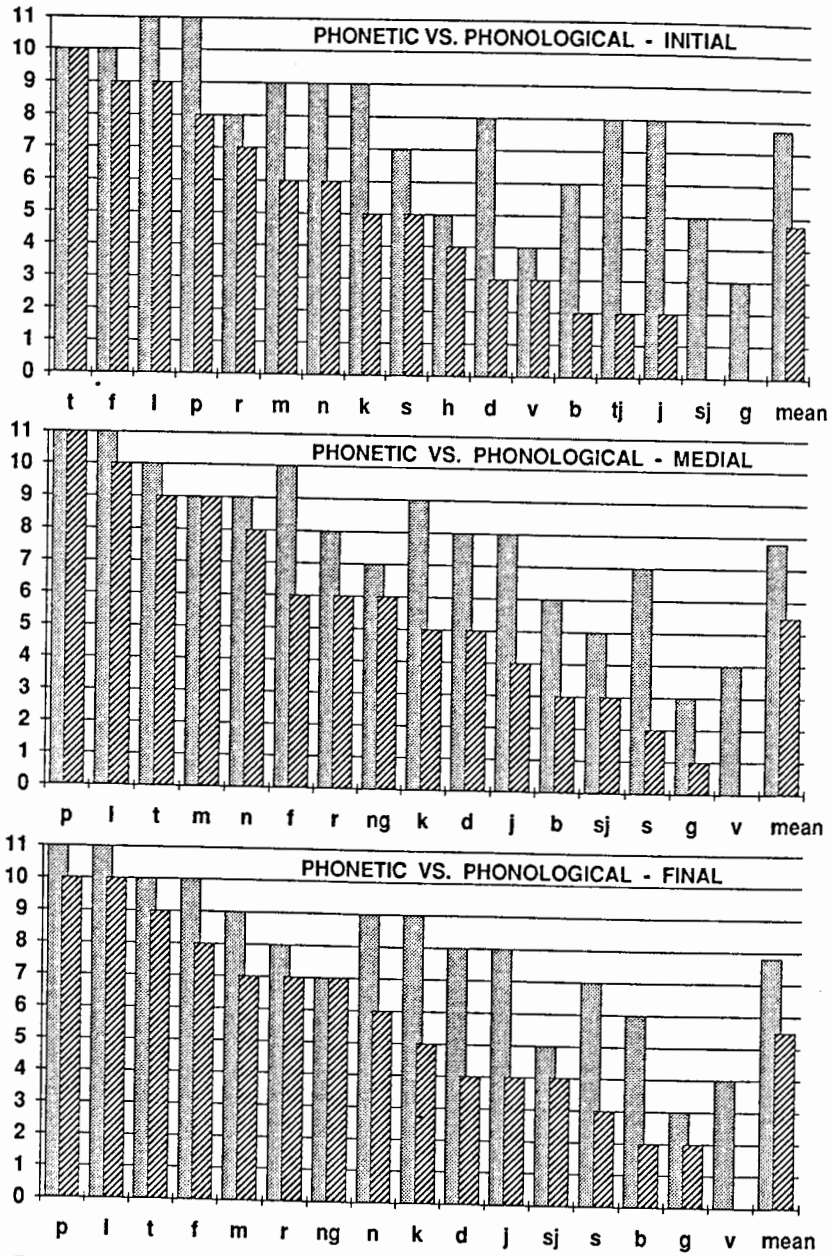


Figure 1. The number of children, who could articulate each consonant correctly at least once (▩ phonetic competence) and, who made correct use of their articulation in spoken language (▨ phonological competence) is shown in initial, medial and final word position. sj denotes /ʃ/, ng /ŋ/ and tj /ç/.

orthography can to some extent explain the deviations found for /ʃ/ and /ç/.

The fact that /s/, /ʃ/ and /ç/ are confused can also be due to incorrect physiological control. The cause can be that the air-flow is too weak and that the constriction is formed in an inaccurate way or produced at an inaccurate place.

	INITIAL	MEDIAL	FINAL
p	b̲,b,m		p̆
b	b̥,p	b̥	b̥,p
m	b,b̥		ᵐb,b
f	v	v,vᴷ	vᴷ
v	v̥	v̥,v̥ᴷ	v̥,v̥ᴷ
t		d	dˀ
d	d̥,t,l	t,n	d̥,g,n
n	d,d̥	d	d,d̥,d̃
l	r,d̥,ɳ	ɳ	m
r	l	l,ɬ	l,ɳ
s	ç,t,d	ç,ʃ,t	ç,n,ʃ
ç	k,t,tj,j,ts	N/A	N/A
j	ç,s,ɸ,i:	ŋ,n,ç	ŋ,n,ç,s
ʃ	k,s,ç	k,ç	k
k	g	g,x	g,k̥,ɸ
g	g̊,k,ɸ	g̊	g̊
ŋ	N/A	nj	
h	ɸ	N/A	N/A

Figure 2. Systematic deviations in different word-positions for the speech sounds that the children could articulate at least once in the speech material.
N/A = not applicable in this position. For an explanation of the diacritics, see reference [1], [3] and [7].

5. FINAL REMARKS

To assess the speech of deaf children phonologically is extremely important since valuable pedagogical information about systematical phonological deviations of speech sounds, which a child controls the articulation of, can be derived. By this means further development of deviant processes can be avoided during the speech acquisition of deaf children.

5. ACKNOWLEDGMENTS

The work has been supported by grants from the Bank of Sweden Tercentenary Foundation.

6. REFERENCES

[1] BUSH, C.N., EDWARDS, M.L., LUCKAU, J.M., STOEL, C.M., MACKEN, M.A., PETERSEN, J.D. (1973), "On specifying a system for transcribing consonants in child language: A working paper with examples from American English and Mexican Spanish", *Report, Dept. of Linguistics, Stanford University*:

[2] DODD, B. (1976), "The phonological systems of deaf children", *JSHD, 41, 2*, 185-197.

[3] GRUNWELL, P. (1987) "Clinical Phonology", *Williams & Wilkins, Baltimore.*

[4] OLLER, D.K., KELLY, C.A. (1974), "Phonological substitution processes of a hard-of-hearing child", *JSHD, 39*, 65-74.

[5] OLLER, D.K., EILERS, R.E. (1981), "A pragmatic approach to phonological systems of deaf speakers", *Speech and Language, Advances in basic research and practice*, 103-141, Academic press.

[6] ÖSTER, A-M. (1989), "Studies on phonological rules in the speech of the deaf", *STL/QPSR 1/89*, 59-162.

[7] ROUG, L., LANDBERG, I., LUNDBERG, L-J. (1987), "Phonetic development in early infancy. A study of four Swedish children during the first 18 months of life", *Report from Dep. of Linguistics, Stockholm University.*

[8] WEST, J.J. & WEBER, J.L. (1973), "A phonological analysis of the spontaneous language of a four-year-old, hard-of-hearing child", *JSHD, 38*, 25-35.

SPEECH TIMING IN ATAXIC DYSARTHRIA

Fredericka Bell-Berti,†,‡ Carole Gelfer,†† Mary Boyle,‡ and Claude Chevrie-Muller*

†Haskins Laboratories, New Haven,CT; ††Wm. Paterson College,Wayne,NJ; ‡St. John's University,Jamaica,NY; *INSERM, L'Hôpital de la Salpetrière, Paris

ABSTRACT

The production of articulate speech involves the complex spatiotemporal organization of articulatory gestures, resulting in characteristic timing patterns across speakers. Acoustic studies of ataxic dysarthria report lengthening of segments and increased mean syllable and utterance durations, but utterance-level durational effects have been described infrequently. In this study, French-speaking ataxic dysarthric speakers failed to show final lengthening effects.

1. INTRODUCTION

This study comparing dysarthric and normal speech was undertaken in an initial attampt to explore the roles of different parts of the CNS in two utterance-level timing patterns: compensatory shortening and final lengthening. The term dysarthria refers generally to a group of speech disorders resulting from neurological damage or disease, and may be characterized by slow, weak, imprecise and/or uncoordinated movements of the speech musculature [9]. We chose to study dysarthric speech because traditional descriptions of the speech characteristics of most dysarthrias include disturbances of speaking rate [3]. We hope that by describing the speech timing patterns of individuals with known neurological pathologies we may improve our understanding of the contributions of specific neural mechanisms to speech timing patterns. This should allow us, eventually, to locate the origins of specific timing differences at motor-execution or higher levels [e.g., 2, 5]. Furthermore, since the existing clinical descriptions of these disorders are sufficiently vague that the same descriptions are often used for patterns that are clearly distinguishable, an added benefit of these studies may be descriptions of timing patterns of the various dysarthrias that have diagnostic value. Since the cerebellum has been implicated in movement timing, and because there are few studies that attempt to relate empirical measures of dysarthric speech timing with perceptual judgments, we have begun with a study of ataxic dysarthric speakers--that is, speakers whose speech pathology arises from cerebellar pathology.

2. METHODS

The data were recorded at L'hôpital de la Salpetrière, in Paris. The subjects were eight native speakers of French, four ataxic dysarthric speakers, and four age-matched control speakers.

Two nonsense "target words," [pɑt] and [splɑt], were embedded in eight sentence frames. The target words occurred medially in four sentences and finally in four sentences. The sentences were of four lengths: 4-, 5-, 6-, and 7-syllables. A native French speaker produced exemplars of each sentence. These were presented in two random orders, with two tokens of each sentence produced on each of two presentations, resulting in eight repetitions of each sentence by each subject. The sentences were produced without pauses, and all were intelligible. All stop closures were achieved, but some ataxic speakers distorted the /s/ frication. All target words were produced with an aspirated /t/ release by all subjects on all repetitions in both target positions. We measured the durations of the /ɛ/ of "C'est une (1-4 syllables)_____" and "C'est une_____ (1-4 syllables)," the /s/ (of 'splat'), the /p/ closure, the vocalic nucleus (/lɑ/ in 'splat' and /ɑ/ in 'pat'), /t/ closure, and /t/ aspiration from the acoustic waveforms.

3. RESULTS

3.1 Overall Speaking Rate

For both subject groups, as the number of syllables per sentence increased, sentence duration increased; this is shown for "pat" target-word utterances in Figure 1, and was also found for "splat" target-word utterances. In this figure, it is also obvious that ataxics' sentences were longer than those of the control subjects. In one sense, at least, then, the ataxic subjects' speech was "slower" than that of the control subjects, as reported in the literature [3]. To see if this "slowness" was distributed uniformly across a sentence, we examined durations of medially and finally positioned target words and their constituent segments.

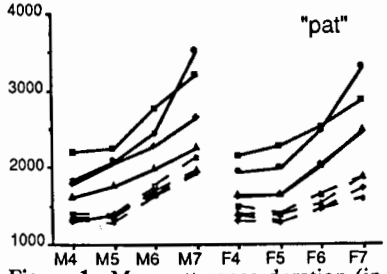


Figure 1. Mean utterance duration (in ms) for each subject and sentence. 'M' indicates medial position target words, 'F' final target words in sentences of 4-, 5-, 6-, and 7-syllables. Ataxic subjects' data are connected with solid lines; control subjects' data are connected with dashed lines.

3.2 Final Lengthening

In accord with previously published data on Swedish and English [e.g., 4, 8], our control subjects produced significantly longer target words when they occurred in final than in medial position (see Fig. 2) for "pat" target-word data; ("splat" data are equivalent for both speaker groups). The final lengthening effect, however, was not evenly distributed across the segments of the target words: the target-word vowels were about 25 ms longer for final-position targets (see Fig. 3 for "pat" data; "splat" data are equivalent for both speaker groups), while the corresponding /t/ closures were 75-100 ms longer (see Fig. 4 for "pat" data; "splat" data are equivalent for both speaker groups).

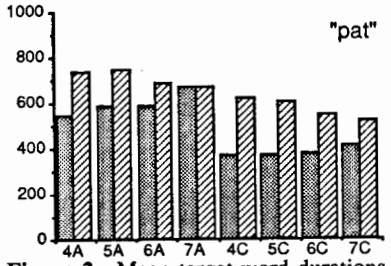


Figure 2. Mean target-word durations for ataxic ('A') and control ('C') subject groups, in 4-7 syllable-sentences. Filled bars represent durations of target words in medial position; striped bars, in final position.

The ataxic subjects, on the other hand, failed to differentiate the target-word durations as a function of sentence position. Indeed, in some comparisons the medial-position vowel was longer than the final-position vowel (Fig. 3), and the final lengthening so evident in the final /t/ closures of the control subjects was absent in the speech of these dysarthric speakers (Fig. 4).

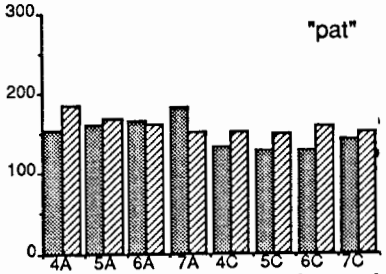


Figure 3. Mean target-word vowel nuclei durations for ataxic ("A") and control ("C") groups for 4-7 syllable-sentences. Filled bars represent durations in medial target words; striped bars, in final target words.

In order to present a more general picture of position effects on segment duration, we have collapsed the data for each group across the sentences of different lengths (Fig. 5). Not surprisingly, neither group showed a difference in initial /ɛ/ duration as a function of target-word position. The control subjects, however, did show differences in all target-word segment durations as a function of target-word position, while the ataxic speakers showed differences only for the initial consonant (/p/) and final aspiration. In the parallel data for "splat," the controls subjects again showed positional effects for all target-word segments ($p<.001$), while the ataxic subjects produced only the initial /s/ and final aspiration with greater length in final-position target words ($p<.05$ and $p<.001$, respectively). The /p/, no longer an initial segment, did not show durational differences as a function of target-word position.

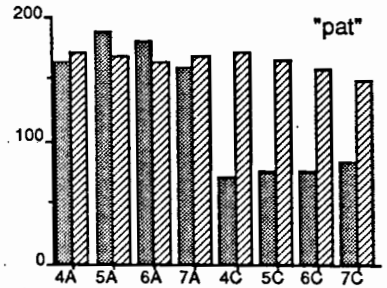


Figure 4. Mean /t/-closure durations for ataxic ("A") and control ("C") groups in 4-7 syllable-sentences..

3.3 Compensatory Shortening

There was no correlation between the duration of /ɛ/ and total utterance duration for the control or for the ataxic speakers. This is in contrast to an earlier study with a different group of French-speaking subjects [1], in which such a correlation was found. It is possible that the syllable-timed structure of French is only minimally compatible with compensatory shortening.

4. DISCUSSION

Our results show that the ataxic speakers differed from the controls in at least two ways: the total durations of their utterances were longer, and they failed to show final-lengthening effects. These differences can be attributed, whether directly or indirectly, to the cerebellar dysfunction. While the role of the cerebellum is incompletely understood, it is thought to be involved in coordinating the motor system. One hypothesis is that the cerebellum is involved in setting the initial parameters of the movement [7], perhaps by biasing muscle spindles; these parameters would be responsible for the effect of syllable position on segment duration observed

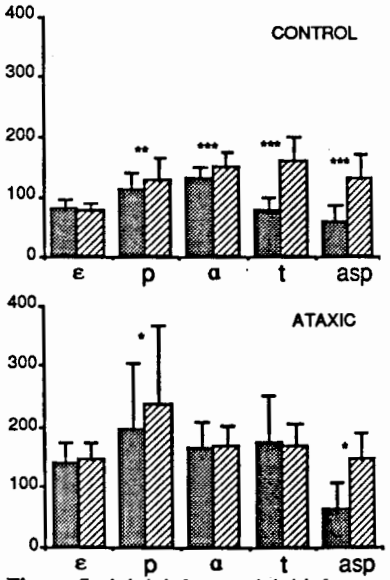


Figure 5. /ɛ/, /p/ closure, /a/, /t/ closure, and final aspiration durations. Segments occurring in medial position are shown with filled bars; final, with striped bars (*p<.05, **p<.01, ***p<.001.)

in normal speakers. Thus, it has been suggested that the hypotonia of cerebellar ataxia results from reduced spindle biasing and leads to the associated slowing of movement [7]. In addition, one of the hallmarks of cerebellar disorders is difficulty terminating movements, a difficulty that may contribute to the longer segment durations of our ataxic speakers. It is unclear whether the increased segment durations and absence of final lengthening are independent aspects of the timing disorder or reflections of disruption of a single underlying mechanism. Finally, the question remains as to whether the timing patterns observed in normal speakers result from speakers' intentions to produce durational differences that may be perceptually important to listeners [6], or whether the durational differences are due to a physiological unwinding as the motor act unfolds.

Our failure to find compensatory shortening effects may result from the syllable-timed structure of French, which may be only minimally compatible with compensatory shortening. That is, the effect may be so small that a much larger sample may be necessary to reveal any effect, and this may explain the difference between these results and those of Bell-Berti and Chevrie-Muller [1].

5. SUMMARY

The utterances of our ataxic subjects were of substantially greater duration than those of our control subjects. In addition, final-lengthening effects were present in the speech of the control subjects (although they were not uniform across the segments of the target words), but were not seen in the speech of our ataxic subjects. Finally, we found no evidence of utterance-length effects in either the control or the ataxic subjects.

6. ACKNOWLEDGMENTS

This work was supported by NIH grant DC-00121 to Haskins Laboratories, INSERM, le Ministère de Recherche et Technologie, and St. John's University.

7. REFERENCES

[1]BELL-BERTI, F. & CHEVRIE-MULLER, C. (to appear). Motor levels of speech timing: Evidence from studies of ataxia. In H. F. M. Peters & C. W. Starkweather (Eds.), *Speech Motor Control and Stuttering*. Amsterdam: Elsevier.

[2]BONNOT, J.-F. (1989). Timing intrinsèque et timing extrinsèque: le temps est-il une variable controlée? *Journal d'Acoustique, 2*, 287-296.

[3]DARLEY, F.L., ARONSON, A.E. & BROWN, J.R. (1975). *"Motor speech disorders"*, Philadelphia: W. B. Saunders Company.

[4]EDWARDS, J., BECKMAN, M. E., & FLETCHER, J. (1991). The articulatory kinematics of phrase-final lengthening. *Journal of the Acoustical Society of America, 89*, 369-382.

[5]KENT, R. D. (1983). The segmental organization of speech. In P. F. MacNeilage (Ed.), *The Production of Speech*, New York: Springer-Verlag, pp. 57-89.

[6]KLATT, D., & COOPER, W. E. (1975). Perception of segment durations in sentence contexts. In A. Cohen and S. Nooteboom (Eds.), *Structure and Process in Speech Production*. Heidelberg: Springer Verlag.

[7]LARSON, C. R., & SUTTON, D. (1978). Effects of cerebellar lesions on monkey jaw-force control: Implications for understanding ataxic dysarthria. *Journal of Speech and Hearing Research, 21*, 309-323.

[8]LINDBLOM, B. E. F., & RAPP, K. (1973). Some temporal regularities in spoken Swedish. *Papers in Linguistics, University of Stockholm, 21*, Stockholm.

[9]YORKSTON, K.M., BEUKELMAN, D.R. & BELL, K.R., (1988), *"Clinical management of dysarthric speakers"*, Boston: College-Hill.

MODIFICATIONS TO STUTTERERS' RESPIRATORY, LARYNGEAL, AND SUPRALARYNGEAL KINEMATICS FOLLOWING SUCCESSFUL FLUENCY THERAPY

P.J. Alfonso, R.S. Story, and J.S. Kalinowski

The University of Connecticut, Storrs and
Haskins Laboratories, New Haven, Ct. USA

ABSTRACT

Within-subject comparisons of respiratory, laryngeal, and supralaryngeal kinematics of severe stutterers immediately before and after successful completion of intensive fluency therapy reveal that an increase in post-therapy fluency co-occurs with a number of spatial and temporal modifications within and among each of the three monitored speech systems. Some of the post-therapy modifications can be distinguished from therapy-directed clinical targets and are presumed to be natural requisites to perceptually fluent speech.

1. INTRODUCTION

It is well known that stutterers' respiratory, laryngeal, and supralaryngeal movements during moments of overt stuttering are radically different from those observed in normally fluent speakers. However, it is not clear whether stutterers' control of the speech mechanism is generally abnormal; that is, abnormal even during production of speech that is perceived as fluent. Clarification of this issue would have important clinical ramifications. Thus, we have undertaken a research program that seeks to resolve the following three questions: 1) are certain kinematic profiles associated with stutterers' perceptually fluent speech distinct from those of normally fluent speakers, and if they are different, 2) which aberrant kinematic profiles, if any, can be modified by speech therapy to become more like those of normally fluent speakers? In addition, 3) we seek to determine which post-therapy kinematic modifications are requisite to perceptual fluency. We report here the results of experiments that focus on the third aim of this research program. The results demonstrate that not all of the kinematic modifications that are observed post-therapy are reflective of clinical instruction but rather are reflective of certain speech motor control strategies that are observed in normally fluent subjects.

2. PROCEDURES

Within-subject pre- and post-therapy kinematic comparisons of the respiratory system (using Respitrace inductive plethysmography), laryngeal system (using photoglottography), and supralaryngeal system (using optoelectric tracking to monitor the movements of the lips and jaw) were made from eight stutterers immediately before and after completion of either one of two intensive fluency programs, and from four control subjects. Program 1, the Summer Residential Stuttering Clinic of Geneseo, New York, represents a Van Riperian type of program that primarily emphases speech rate control, and Program 2, the Communication Reconstruction Center's (CRC) of New York City version of the Precision Shaping Fluency Program (PFSP), represents a highly structured physiologically oriented program. The comparisons across different therapy programs are primarily motivated by our attempts to differentiate therapy induced kinematic modifications from kinematic modifications that are requisite to stutterers' increase in fluency. For example, kinematic modifications that occur in Program 2 subjects who show post-therapy increased fluency but not in equally successful Program 1 subjects could be considered requisite to the achievement of Program 2 clinical targets but not necessarily requisite to increased fluency. On the other hand, a kinematic modification that occurs in all successful subjects, including those who completed Program 1, which does not emphasize physiological clinical targets, could be considered a physiological requisite to increased fluency. All stutterers who took part in these experiments were diagnosed as severe or moderate pre-therapy and as mild post-therapy. Kinematic measurements included traditional motor control indices, e.g. sequential ordering of articulator movements, and those that more directly address the achievement of the various clinical targets. Two paradigms were used: a variable-foreperiod simple reaction-time (RT) task and a paradigm that assesses relatively natural speech, the production of the phrase "he see CVC again" where "C" represents various stops and fricatives and "V" represents /i,e/. Only fluent utterances, defined perceptually and physiologically, are discussed here.

3. RESULTS

3.1. Respiratory-Laryngeal Kinematics in Reaction-Time Tasks

In a related reaction-time study, we showed that quantitatively different respiratory and laryngeal behaviors underlie stuttering severity and variable-foreperiod (response preparatory interval) effects on acoustic RT. Severe stutterers showed both delayed initiation and inappropriate organization of respiratory and laryngeal events leading to phonation at all foreperiods, while mild stutterers differed from severe stutterers primarily at short but not long foreperiods [6,7]. In our first experiment that compares within-subject pre- and post-therapy acoustic RT performance, we found that post-therapy increase in fluency covaries with acoustic RT improvement for both Program 1 and 2 subjects, that the magnitude of acoustic RT improvement depends on therapy program type and for some subjects approaches normal values, and suggests that acoustic RT represents a dynamic measure of respiratory-phonatory function rather than a fixed and presumably neurologically based delayed latency [1]. We continue to use the same variable-foreperiod RT protocols to evaluate the effects of therapy on respiratory-laryngeal kinematics in stutterers because we have found that an isolated vowel response in the RT paradigm represents a relatively easy stimulus for stutterers to produce fluently, presumably because an isolated vowel is less physiologically complex compared to reiterate speech, and because the protocol provides a large number of perceptually fluent responses in a relatively short amount of time. Equally important, the wide variety of respiratory-laryngeal kinematic patterns exhibited by stutterers in reiterate contexts makes analysis of respiratory-laryngeal control strategies much more straightforward in RT tasks. However, important subject differences in post-therapy respiratory-laryngeal pre-phonatory strategies are still evident, some of which cannot be explained in terms of achievement of therapy-directed clinical target behaviors. For example, the RT data shown in Figure 1 demonstrate that this Program 1 stutterer reduces post-therapy phonation response latency primarily by reducing the time, relative to pre-therapy performance, required to complete respiratory and laryngeal pre-phonatory maneuvers, e.g., appropriate levels of respiratory inflation and preparatory vocal fold adjustment for phonation. On the other hand, Figure 2 shows that a different Program 1 stutterer reduces post-therapy phonation latency primarily by improvement in respiratory-laryngeal temporal coordination, e.g., the moment of onset of respiratory compression relative to laryngeal adduction for phonation.Taken together, the results indicate that stutterers who increase fluency following therapy generally demonstrate RT improvement at acoustic, respiratory, and laryngeal levels of measurement. However, the differences in response strategies among stutterers indicate that the the physiological bases for the covariation between acoustic RT improvement and perceptual fluency improvement are complex and are not entirely related to clinical target behaviors. For example, respiratory-

laryngeal temporal coordination may make a greater contribution to improved acoustic RT than either respiratory RT or laryngeal RT in those stutterers who demonstrate either: 1) relatively short pre-therapy response latencies, and/or 2) appropriate levels of lung volume inflation for speech, and/or 3) appropriate laryngeal abductory/adductory gestures for normal phonation.

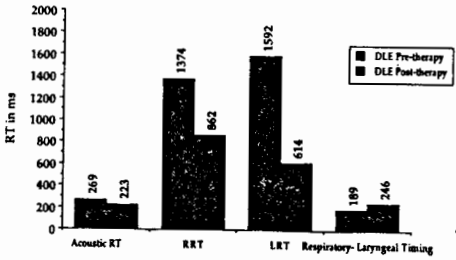


Fig. 1. Pre- (left bar) and post-therapy RT values (right bar), subject DLE.

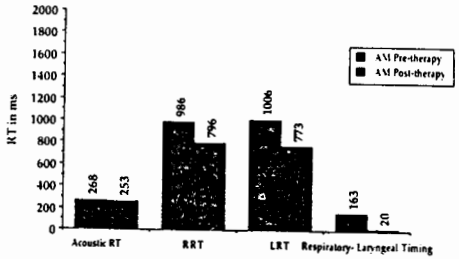


Fig. 2. Pre- (left bar) and post-therapy RT values (right bar), subject AM.

3.2. Supralaryngeal Kinematics in Phrase Length Utterances

In more natural speech tasks, post-therapy increase in fluency co-occurs with kinematic modifications at all measured levels of speech production. Some of these modifications appear to be related to specific therapy-directed clinical targets while others do not and appear to be related to motor control strategies observed in normally fluent subjects. For example, Figure 3 shows an example of stuttering severity and therapy influences on lip and jaw relative timing and sequence patterns during /p/ closure for perceptually fluent productions of /pit/ in "he see pete again" [5]. The data for two controls, shown on the left, represent two different sessions about six weeks apart and are consistent with the results obtained from a larger group of control subjects [3], with respect to both inter-articulator relative-timing and sequence patterns. The stutterers' data, shown on the right, are quite different. Recall that these stutterers were classified as severe pre-therapy and mild post-therapy. Considering inter-articulator latencies first, note that pre-therapy latencies for stutterer AB (specifically lower lip lag of the upper lip) and for stutterer PC (specifically jaw and upper lip lag of lower lip) are much greater than the corresponding control subject latencies. For both of these subjects, post-therapy latencies are significantly reduced relative to their pre-therapy latencies, even though their post-therapy speech rate was significantly reduced compared to their pre-therapy rate. Turning next to sequential order, note that two of the stutterers, KH and PC, do not show the expected upper lip, lower lip, and jaw sequence in either the pre- or post-treatment condition. Also note that for stutterer KH, the pre- and post-treatment comparison shows a complete sequence reversal. Similar results were obtained for /pet, fit, and fet/ and indicate that post-therapy increased fluent speech can be marked by improved inter-articulator relative-timing and, less frequently, by alteration of the sequence patterns, although the altered sequence may not be like that of the controls. The lip and jaw sequence pattern observed in normally fluent speakers most likely is related to neural and biomechanical interactions [2] and thus reflects differences in both neural control and biomechanical processes between stutterers and controls.

4. DISCUSSION

In conclusion, the results we have obtained thus far suggest that post-therapy increase in fluency co-occurs with spatial and temporal adjustments of the respiratory, laryngeal, and supralaryngeal systems. For example, we have observed 1) an increase in inspiratory and expiratory lung volume exchange,

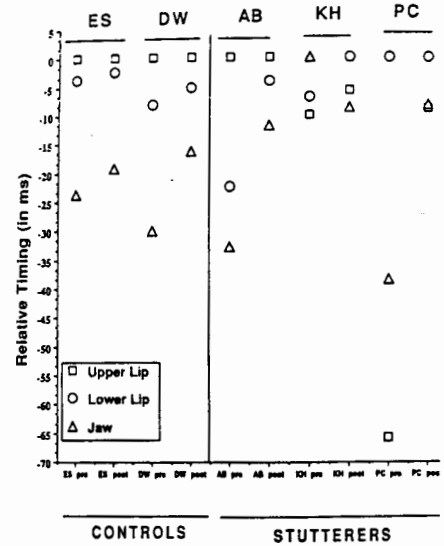


Fig. 3. Temporal organization of upper lip, lower lip, and jaw. Controls left, stutterers right.

duration, and flow, all of which approach values exhibited by normally fluent subjects during phrase length utterances, 2) an increase in the duration of laryngeal abduction and adduction gestures although speech rate decreases post-therapy, 3) a reduction in the frequency of inaudible and phase-locked respiratory-laryngeal kinematic abnormalities, and 4) a reduction in the displacement, peak velocity, and duration of lip and jaw movements in target obstruent-vowel sequences. In addition, certain intra- and inter-system spatial and temporal coordinative adjustments co-occur with post-therapy increase in fluency. Some of the kinematic modifications we observe appear related to the clinical strategies associated with specific therapy programs while others do not. The latter modifications may be manifestations of post-therapy adoptions of certain normal motor control strategies that are requisite to fluent speech production. Our plan is to compare the kinematic modifications of stutterers who successfully complete a variety of different therapy programs, the notion being that the most important modifications leading to fluency will be shared by all successful stutterers even though the clinical instructions to the different groups can differ. In this way, we hope to identify those kinematic strategies that are requisite to the production of perceptually fluent speech.

5. ACKNOWLEDGEMENT

This research is supported by NIH Grant DC-00121. The manuscript was prepared during the first author's tenure as Fulbright Research Scholar, The University of Nijmegen.

6. REFERENCES

[1] Kalinowski, J. S. and Alfonso, P. J. (1987), "Improvement in laryngeal reaction time with improvement in fluency", *Asha, 29-10,* 124 (A).

[2] Gracco, V. (1988), "Timing factors in the coordination of speech movements", *J. of Neuroscience, 8,* 4628-4639.

[3] Gracco, V.L. and Abbs, J.H. (1986), "Variant and invariant characteristics of speech movements", *Experimental Brain Research, 65,* 156-166.

[4] Story, R.S. (1990), *"Articulatory, laryngeal, and respiratory movements in stutterers' fluent speech before and after therapy",* Unpublished Ph. D. dissertation, University of Connecticut.

[5] Story, R.S. & Alfonso, P.J. (1989), "Temporal reorganization of lip and jaw gestures following an intensive therapy program", *ASHA, 31-10,* 65 (A).

[6] Watson, B. C. and Alfonso, P. J. (1983)," Foreperiod and stuttering severity effects on acoustic laryngeal reaction time", *Journal of Fluency Disorders, 8,* 183-205.

[7] Watson, B. C. and Alfonso, P. J. (1987), "Physiological bases of acoustic LRT in nonstutterers, mild stutterers, and severe stutterers", *Journal of Speech and Hearing Research, 30,* 434-447.

ETUDE D'UNE AIDE TACTILE POUR SOURDS PROFONDS: IMPORTANCE DU CODAGE.

Rémi BRUN

CCA Alésia , 26 rue Jean Moulin, 75 014 Paris, France

ABSTRACT:
To study the best way to present speech information to the skin, we have focused on the coding part of the process. The code needs to be optimised according to two fields of constraints; those linked to the speech,the information to be transmitted, and those relying on the abilities of the receptive processor, the tactile sense. A way to assess a code has been implemented, emphasising the temporal resolution versus the size of the repertoire.

1. INTRODUCTION

Se proposer de transmettre des informations contenues dans la parole par le sens du toucher, c'est s'attaquer à un problème de transformation d'un signal défini, l'onde sonore de la parole, en un signal devant non seulement être senti tactilement, mais surtout compris par le sujet recepteur.

Or il est clair que la simple transposition physique de la vibration de l'air, recueillie par un microphone, et transmise à un vibrateur après amplification n'est pas suffisante (4). Il est donc necessaire de réfléchir sur le moyen de coder l'information contenue dans le signal de parole en fonction des aptitudes spécifiques du recepteur tactile. La question du codage n'est plus dépendante des supports physiques de la transmission, mais doit satisfaire au mieux à ces deux ensembles de contraintes:

- d'une part, l'ensemble des propriétés de l'information à transmettre. Chacun sait, pour la parole, qu'il existe un grand nombre de connaissances sur la manière dont elle est construite: structures, hiérarchies, éléments, prosodie, redondance, variabilité, vitesse...

- d'autre part, l'ensemble des propriétés caractérisant l'acquisition et le traitement d'information par le recepteur stimulé. Les études sur la perception tactile sont moins nombreuses que celles sur la vue ou l'ouïe, mais il existe tout de même un savoir éparpillé: perception en fréquence, amplitude, localisation, orientation, forme, en fonction du temp, phénomènes de déformations liés au masquage, à la sommation, au mouvement apparent,... Par ailleurs, l'étude d'autres systèmes de traitement de l'information comme la vision ou l'audition peut fournir des renseignements précieux sur les moyens utilisés pour simplifier la tâche de comprehension.

Concevoir une prothèse tactile pour transmettre des informations de la parole, c'est essayer de trouver un codage intégrant au mieux ces deux champs de contraintes qui sont parfois contradictoires. Avant de chercher à trouver une bonne solution, il est necessaire de croire que certaines solutions sont meilleures que d'autres. L'exemple de la lecture tactile pour les aveugles est encourageant. Avant la création du code Braille, on utilisait pour transmettre l'écriture alphabétique sur une feuille cartonnée standard un principe de lettres normales imprimées en relief. Louis Braille a proposé un autre code, présentant la même information avec la même méthode de stimulation: le Braille à 6 points qui s'est révélé bien plus performant.

Dans notre cas, penser que des codes sont meilleurs que d'autres ne suffit pas si l'on ne dispose pas de moyens simples de réaliser cette évaluation. S'il faut que des échantillons représentatifs de sujets aient suivi suffisament d'entraînement, pour avoir une idée de la qualité de la comprehension de la parole permise par différents codes, l'étude risque d'être trés lente à donner des résultats exploitables. Par contre, des comparaisons trés simples et trés rapides peuvent ne pas refléter les performances possibles à long terme.

2. ETUDE

Nous avons commencé une étude avec un code donné (fig 1), défini en essayant de respecter les contraintes liées à la structure de l'information à transmettre ainsi que celles dictéees par les propriétés du recepteur tactile. Nous avons essayé d'utiliser des tests d'évaluation simples et pouvant être passés rapidement après le début de l'entrainement pour se faire une idée des qualités du code par rapport à celles visées lors de sa conception.

3. TESTS

Deux tests principaux:

- Test 1: phonème isolé.
Identification d'un stimulus parmi 16 ou 20 consonnes, les 12 voyelles ou tous les phonèmes réunis; ce test donne une bonne indication sur les distances entre les stimuli ainsi qu'une appréciation générale du pourcentage de confusion.

- Test 2: triphonème
Test d'identification d'un phonème parmi 3 présentés successivement. Le premier phonème est sujet au masquage arrière, c'est à dire à la que son identification devient plus difficile à cause de la présence de stimuli qui suivent proche dans le temps.

Le troisième est sujet uniquement au masquage avant, comme le masquage arrière mais avec le masqueur présenté avant le stimulus.

Le second est sujet aux deux masquages combinés et est donc celui dont la reconnaissance subit la plus sévère dégradation.

4. MATERIEL

Les expériences ont été réalisées sur une cellule expérimentale (3) construite autour d'un vidéolecteur piloté par ordinateur . On a ainsi la possibilité d'associer à chaque séquence de parole enregistrée sur le vidéodisque (image + son) les stimuli tactiles définis par le code. Des exercices de reconnaissances en tactile accompagné ou non de lecture labiale sont alors gérés par l'ordinateur.

Il est important de noter que le code est édité de façon manuelle, le traitement de la parole ainsi efféctué n'étant qu'une simulation, laissant pour le moment de coté le problème de la réalisation technique en temps réel.

Le système tactile utilisé pour stimuler le doigt était un OptaconII relié à l'ordinateur par une liaison RS232. Cet appareil permet de déssiner les dessins voulus sur une matrice de 100 picots (20*5) vibrant à la fréquence constante de 230 Hz, et dont l'amplitude ne peut être réglée que par un potentiomètre extérieur.(1)

Six sujets ont suivi ces expériences:

- A) 2 sujets adultes entendants ont suivi environ 20 heures d'entrâinement.

- B) 4 sujets enfants sourds (13-15 ans) ont suivi environ 40 heures d'entraînement.

4. RESULTATS

La première observation importante concerne les sujets et la rapidité de leurs progrès. Alors qu'il fallait 40 heures au sujets B pour parvenir à des résultats n'excédant pas 80% au test1, les sujets A sont arrivés à 90% en moins de 3 heures (2,3). Il semble cependant que la nature des erreurs est la même pour les 2 tests; on peut donc penser que des améliorations du code sensibles pour les sujets A le seront aussi pour les sujets B, et ainsi se concentrer sur des études avec les sujets A qui seront beaucoup plus rapides.

Les résultats du test1 nous renseignent de façon trés précise sur le taux de confusion relatif à un code, ainsi que sur les confusions les plus fréquentes, nous permettant de localiser les sources d'erreurs et d'essayer d'y remédier.
Mais des résultats proches de 100% ne sont pas une preuve suffisante de la qualité du code, et c'est là que le test 2 intervient. La dégradation générée par le masquage est trés nette .

sans masquage: **90%** (test1)

1er **50%** ; masquage arrière
2ème **35**; masquage avant et arrière
3ème **65%** ;masquage avant

Il nous semble donc que c'est à ce niveau d'integration temporelle qu'il faut travailler. Le test 2 peut être réalisé après une dizaine d'heures de pratique, et il permet d'enregistrer l'influence de l'entraînement sur les performances. Après 20 heures les sujetsA obtenaient les résultats suivants: 1er (70%); 2ème (50%) ; 3ème (85%); mettant en évidence les possibilités de progrès.

Nous comptons créer un nouveau test du même type, mais ou les stimuli pourraient être plus signifiant, pour évaluer la possibilité d'apprendre à reconnaitre des groupements de stimuli, formant des mouvements complexes. Un test syllabique, (inspiré d'un exercice sur l'identification de nombres entre 1 et 100 présentés phonétiquement qui avait fait apparaître l'importance de ces super-unités), est en cours de conception.

6. CONCLUSION

Dans cette étude se concentrant sur la question du codage, nous comptons persévérer dans la mise au point de tests qui permettront une évaluation rapide des qualités d'un code. La formule qui donne le débit d'informations en fonction de la taille du répertoire (R) et du nombre de stimuli perceptibles par seconde (N):

$$\text{Débits (bauds)} = N \times \ln_2(R)$$

reflète bien l'importance de N par rapport à R.

Cette démarche n'est d'ailleurs pas réservée à la prothèse tactile pour sourds, et peut s'appliquer à d'autres études, soit avec un récepteur différent (oeil, cochlée,...) soit avec des informations différentes (feedback sensoriel pour les orthèses orthopédiques...)

7. BIBLIOGRAPHIE

[1] BLISS (66)"Tactile perception of sequentially presented patterns" *Percept. & Psycho.* 1 125-130

[2] BRUN,R.P.,CHALLIER, G., IMBERT, G.(90) "Etude d'une aide tactile pour sourd profond" *XV Congrès de Biomécanique, ENSAM Cluny 90*

[3] BRUN R.P.,CHALLIER G., IMBERT G.(91) "Study of a tactile aid for the deaf:simulatio on an experimental set (computer, videoplayer, OptaconII)" *1st Eur. Conf. on Biomédical Engieering, Nice 91,*

[4] GAULT, R.H. (1924) "Progress in experiments in tactile interpretation of oral speech" *J. Abnorm.Soc.Psych* 14, 155-159

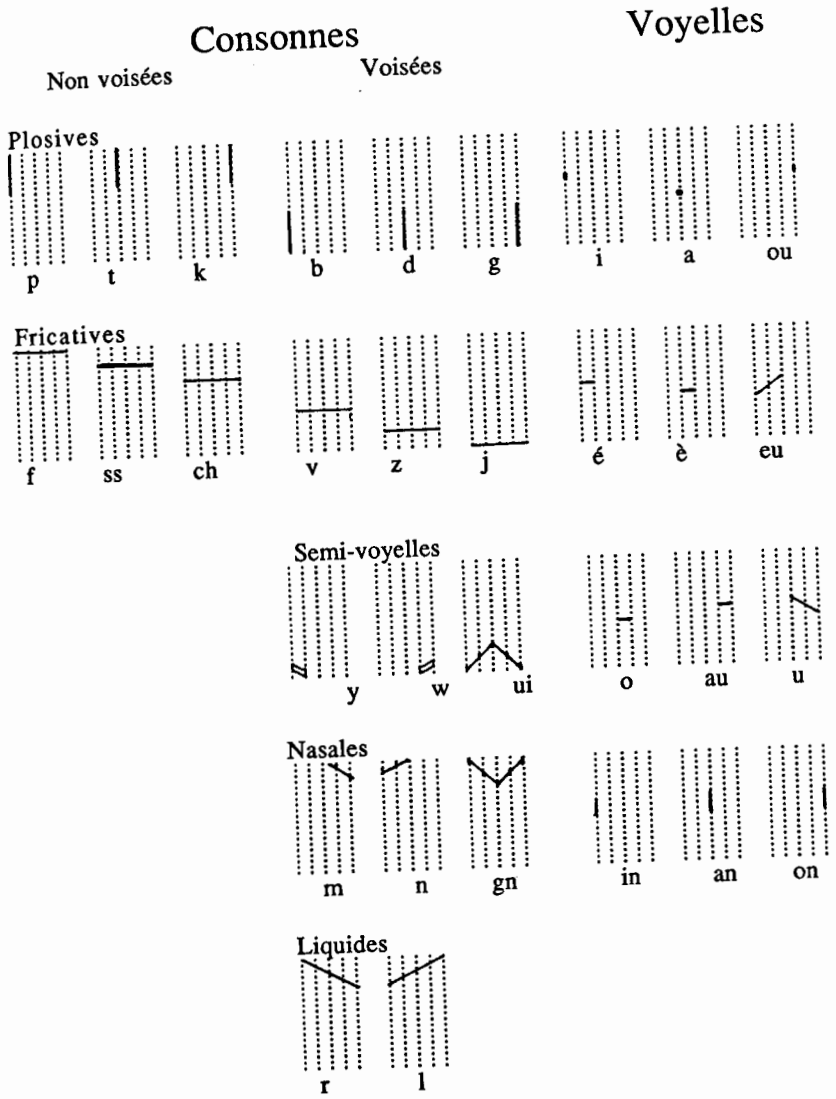


figure 1: Code utilisé dans cette étude. Il a été construit en essayant de respecter les contraintes liées aux propriétés de l'information à transmettre, la structure phonétique de la parole, et celles relevant des caractéristiques perceptives du recepteur tactile, allant du capteur aux capacités de traitement de l'information évolué.

L' EVALUATION OBJECTIVE DE LA DYSPHONIE: UNE METHODE MULTIPARAMETRIQUE.

Antoine GIOVANNI [1], Valérie MOLINES [1], Noel NGUYEN [2] et Bernard TESTON [2].

[1] Centre Hospitalier Universitaire de la Timone, Marseille.
[2] URA 261 du CNRS, Institut de Phonétique, Aix-en-provence.
FRANCE.

ABSTRACT

The aim of this study is to validate an aid for the evaluation of dysphonia with objective measurements.

We recorded exhaled airflow, fondamental frequency and sound level pressure, for a sustained vowel "a", with 51 dysphonic subjects and 15 normal subjects. The following measurements are made on these three parameters: mean value, standard deviation and coefficient of variation. The exhaled airflow volume was also computed for a duration of 2 seconds. A principal components analysis of the measurements indicated that it is possible to recognise the classes of vocal evaluation and vocal pathology. These findings reenforce on objective aid for the vocal evaluation of dysphonia.

1. INTRODUCTION

Dans un exposé fort célèbre, M. HIRANO [1] détaille remarquablement les différentes méthodes d'évaluation objectives de la production vocale : "*1- to diagnose the etiologic disease, 2- to determine the degree and the extent of the etiologic diseases, 3- to evaluate the degree and the nature of dysphonia, 4- to determine the prognosis and, 5- to monitor changes*". Son étude, basée sur plusieurs centaines de cas,décrit statistiquement les méthodes d'évaluation objectives au moyen généralement d'histogrammes, en fonction de différents paramètres tels que: le débit d'air buccal et le volume d'air associés au temps maximal de phonation, la fréquence fondamentale usuelle (Fo), celle de l'intensité acoustique de l'émission vocale (SPL), de la pression sous glottique, et d'autre paramètres moins communs. Cependant, HIRANO n'est pas allé plus loin dans ses investigations statistiques,en essayant par exemple de faire émerger des classes pathologiques ou différents degrés de dysphonie en fonction de plusieurs paramètres. C'est ce que nous cherchons à faire dans la ligne du travail de DEJONCKERE [2] dont nous apprécions la méthode. Notre démarche n'est pas innocente. En effet, le but de cette étude est de valider ou invalider une aide objective à l'évaluation vocale pour des applications cliniques et de rééducation dans le domaine de la Phoniâtrie, sans être pour autant aussi optimiste que HIRANO dans son introduction..

2. PROTOCOLE EXPERIMENTAL

2.1. Le dispositif

Il est schématisé dans la figure 1; L'expérience consiste à mesurer 3 paramètres de l'émission vocale. - Le débit d'air buccal en cc/s. - L'intensité acoustique (spl) en dB. - La fréquence fondamentale (Fo) en Hertz. Le débit d'air buccal est mesuré au moyen d'un aérophonomètre [3] développé dans le cadre d'un contrat de faisabilité de l'INSERM. Il se caractérise par une grande dynamique associée à une bonne linéarité. Le détecteur de mélodie est un fréquencemètre instantané numérique de grande précision (1 Hz à 1000 Hz), développé pour l'étude des phénomènes microprosodiques [4]. Il est associé à un intensimètre qui mesure le logarithme de la valeur efficace du signal de parole avec une intégration pondérée exponentielle de 10 ms. Sa dynamique est de 80 dB avec une précision de + ou - 1 dB. Le signal de parole est enregistré sur un magnétophone REVOX B77 à partir d'un microphone BEYER B75 placé à 20 cm de la bouche du sujet, nous ne faisons pas de calibration au niveau, les mesures sont donc en dB relatifs (à quelques dB de la référence absolue) mais constantes pour tous les locuteurs. Les signaux de parole, de débit, de mélodie et d'intensité sont acquis simultanément sur un micro ordinateur PC vectra ES12 au moyen d'une carte de conversion DATA TRANSLATION 2801-A. Le signal de parole est échantillonné à 4 kHz, les autres signaux à 1 kHz. Ils sont codés sur 12 bits. Les acquisitions sont gérées par le logiciel PHYSIOLOGIA [5]. L'enregistrement sur l'ordinateur et le magnétophone est télécommandé par le locuteur. Il est demandé à chaque locuteur de réaliser, dans la mesure du possible, 4 enregistrements. 1- La voyelle «A» tenue pendant un temps de phonation de 4 à 5 secondes, à la hauteur et intensité la plus naturelle au sujet. 2- La même voyelle «plus haute». 3- La

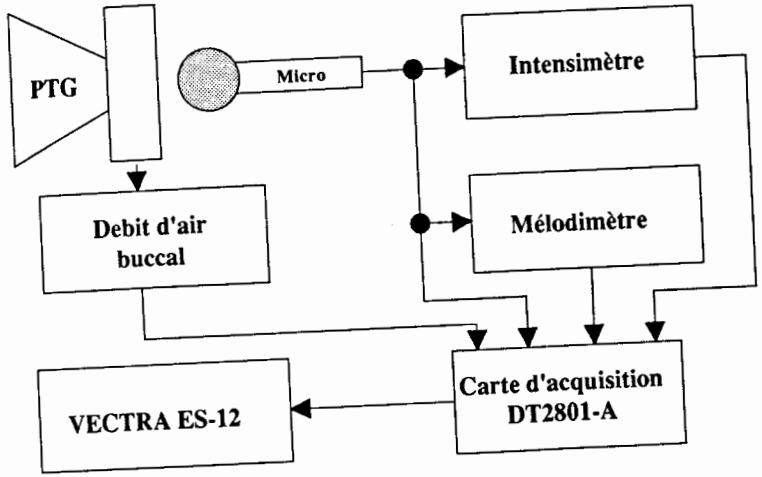


Figure 1: Schéma de principe de la chaine d'acquisition.

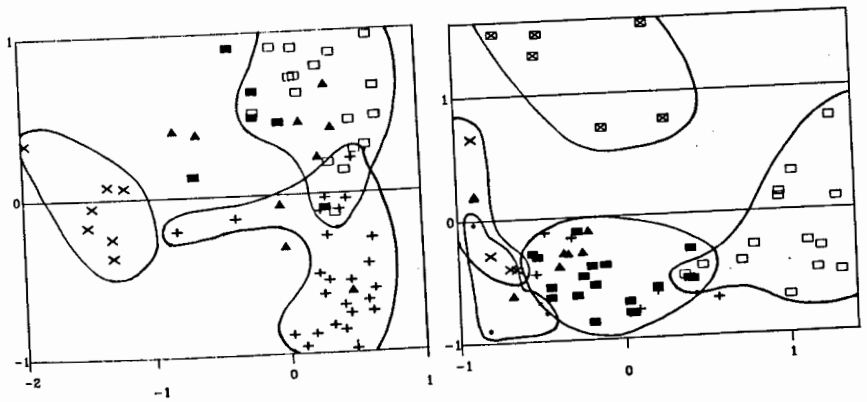


Figure 2: Tableau factoriel des sujets en fonction de l'échelle de qualité de voix

Figure 3: Tableau factoriel des sujets en fonction de l'évaluation pathologique.

même voyelle «encore plus haute». 4- La phrase «elle n'aimait ni maman ni mamie», à la hauteur usuelle.

2.2. Le matériel

Matériel pathologique: 50 patients consultants pour dysphonie dans le service ORL du centre hospitalier universitaire de la Timone à Marseille, dont 24 femmes et 26 hommes, de 6 à 62 ans. On distingue deux catégories de patients. D'une part, ceux atteints par une dysphonie jamais explorée auparavant, d'autre part, ceux qui consultent pour un contrôle après traitement (chirurgical ou médical) ou après une rééducation orthophonique, ou les deux associées. Dans ce groupe,se trouve un sous groupe de patients guéris que l'on peut identifier au groupe des normaux.

Pour tester la validité du classement multiparamétrique nous avons décidé de mélanger au matériel pathologique un matériel de normalité, constitué par un échantillon de la population étudiante de l'Université de Provence comprenant 14 sujets, 7 femmes et 7 hommes de 20 à 45 ans, choisi dans l'environement du laboratoire pour la stabilité de la qualité vocale.

3. LA METHODOLOGIE

Pour chaque patient, la dysphonie est évaluée et la pathologie diagnostiquée. L'évaluation acoustique est faite au moyen de 5 facteurs bipolaires déduits de la méthode de HAMMARBERG [6]. Elle aboutit à une échelle de dysphonie de 4 degrés: légère, moyenne, sévère et très sévère. Le degré 0 représente les voix normales. L'évaluation est faite par un phoniâtre à partir des enregistrements effectués pendant l'experience, et de textes lus avant cette dernière. L'examen clinique laryngé est pratiqué par un ORL au moyen d'un laryngoscope rigide avec étude vidéolaryngostroboscopique pour chaque sujet du groupe pathologique. Les déficiences observées sont classées en fonction de la nature physiologique du mécanisme du vibrateur laryngien. On distingue ainsi 5 classes de pathologies.

1- Altérations morphologiques des cordes vocales:
-Légère (nodule) codée MORL.
-Modérée (inflammation) codée MORM.
-Sévère (Reinke) codée MORS.
2- Altération de la motricité codée MOTR.
3- Séquelles chirurgicales:
-Cordectomie codée CHIC.
-Laryngectomie partielle (Tucker) codée CHIP.
-Laryngectomie totale codée CHIT.
4- Dysfonctionnement avec larynx normal codée DYSF.
5- Patients guéris (voix et larynx normaux) codée NORP.

Les sujets normaux sont codés NORM.

Sur chaque paramètre débit, intensité,mélodie, les mesures suivantes sont effectuées dans une durée normalisée de 2 secondes où les attaques ne sont pas prises en compte: Valeur moyenne, écart type et coefficient de variation,ainsi que le débit intégré pour avoir le volume d'air expiré. Les calculs sont effectués au moyen de l'éditeur de signal de PHYSIOLOGIA [5].

4. ETUDE STATISTIQUE

Les traitements statistiques sont effectués au moyen du logiciel ADDAD [7].On ne retient de ces mesures que 9 variables sur 10, on supprime la moyenne du débit d'air expiré car elle est parfaitement corrélée avec le volume d'air. On ajoute une variable ordinale codée sur 5 niveaux pour la qualité de la voix, ainsi qu'une variable qualitative correspondant aux 10 classes pathologiques. Les données représentent un tableau de 174 lignes sur 11 colonnes: 65 productions du «A» usuel, 60 du «plus aigu» et 49 de «l'encore plus aigu». Un premier traitement de statistique élémentaire nous permet d'étudier la répartition des différentes variables sous la forme d'histogrammes.

4.1 Classification des sujets en fonction de l'échelle de qualité de voix

Il est difficile de traiter le tableau des données dans sa forme initiale par une analyse factorielle, car les variables descriptives sont hétérogènes (écart type, coefficient de variation, moyenne pour des signaux de nature très différente). L'analyse en composantes principales normée, offre la possibilité d'homogénéiser un tableau en divisant chaque valeur par l'écart type de la variable correspondante. Ainsi, les différentes variables présentent toutes la même variance (=1). Dans notre cas, on hésite cependant à accomplir cette normalisation pour des raisons de nature «sémantique». Comment interpréter un coefficient de variation rapporté à son écart type? La procédure suivie par DEJONCKERE [2] n'est peut être pas à cet égard sans inconvénients. On a donc recodé le tableau sous une forme disjonctive complète. Cela consiste à délimiter sur chaque dimension une dizaine d'intervalles caractérisés par des effectifs égaux. Les intervalles sont considérés comme des classes à l'intérieur desquelles se répartissent les sujets. On a donc transformé une variable quantitative en variable ordinale. On «éclate» alors cette variable ordinale, qui comporte les modalités possibles, en une nouvelle variable indicatrice de modalités binaires. Ce nouveau codage permet de rendre les données homogènes et de les soumettre à une analyse des correspondances [8]. Pratiquement, les 9 variables initiales sont découpées en 11 intervalles d'effectifs égaux puis recodées sous forme disjonctive complète. On s'est limité, dans la présente étude, uniquement aux voyelles de hauteur usuelle (n=65). On croise la variable disjonctive de la

qualité de la voix avec les 99 indicatrices de modalité. Ce nouveau tableau est alors soumis à une analyse des correspondances. Les individus n'entrent pas en ligne de compte dans le calcul des facteurs. Cette procédure peut se comparer à une régression multiple au moyen de laquelle la variable «qualité de voix» est expliquée, au sens statistique du terme, par l'ensemble des mesures. Les résultats sont donnés dans le plan factoriel de la figure 2. La répartition des 5 niveaux de qualité de voix est bien différenciée si ce n'est pour les niveaux moyens et sévères qui s'interpénètrent et se répartissent plus largement. Ceci s'explique par le fait que ces niveaux sont plus difficiles à différencier auditivement par un seul auditeur. Les deux exceptions sur la normalité s'expliquent par le fait qu'il s'agit de deux sujets masculins à voix très grave (basse profonde). Le sujet «sévère» perdu au milieu de la normalité est un cas très particulier qui présente une voix au timbre dégradé mais avec une bonne stabilité de la production vocale. Cela montre la nécessité d'une analyse harmonique complémentaire.

4.2. Classification des sujets en fonction de l'évaluation pathologique

On cherche par ce traitement à savoir si la variable pathologique recodée sous forme disjonctive, est corrélée avec les variables de mesure. La procédure statistique est la même que précèdemment. Cependant, faute d'un nombre suffisant de sujets, nous supprimons les modalités MORS, CHIC, et CHIT. Les résultats sont donnés dans le plan factoriel de la figure 3. Il apparaît que la répartition est encore mieux marquée que pour la qualité de voix. On constate la différence entre les sujets normaux avec les «normaux guéris», qui s'apparente pathologiquement à de légères lésions morphologiques.

CONCLUSION

Cette étude fait apparaître qu'il est possible de différencier objectivement au moyen du débit d'air expiré, de l'intensité et de la mélodie, des patients dont l'évaluation du degré de dysphonie et la classification pathologique sont différentes. Nous considérons que ces résultats apportent un argument solide à l'utilisation d'une évaluation vocale assistée par des mesures objectives. Nos résultats ne sont que partiels et portent sur un échantillon de patients assez faible. Cependant, nous les jugeons très encourageants. Le dépouillement continue avec les productions de différentes hauteurs mélodiques. Nous allons ajouter l'analyse harmonique qui nous semble susceptible d'apporter des renseignements complémentaires et approfondir l'importance de l'information contenue dans les différents paramètres.

BIBLIOGRAPHIE

[1] HIRANO, M. (1989),"*Objective evaluation of the human voice: Clinical aspects.*",Folia Phoniatrica, Vol 41, 89-144.

[2] DEJONCKERE, P. H. (1990), "*Bruits de turbulence et aperiodicité dans la voix pathologique. Une approche multifactorielle.* ", Revue de Laryngologie, Vol 3, No 4, 353-357.

[3] TESTON, B. (1983),"*A system for the analysis of the aerodynamics parameters of speech.*", 11 th ICPS, Utrecht, Aug 1983, Sec 5, 457.

[4] TESTON, B. et ROSSI, M. (1977), "*Un système de détection automatique de la fréquence fondamentale et de l'intensité de la parole.*",8 èmes JEP, Groupement des Acousticiens de Langue Française, Aix, 111-117.

[5] GALINDO, B et TESTON, B. (1990), "*PHYSIOLOGIA: Un logiciel d'analyse des paramètres physiologique de la parole.*", TIPA, Vol 13, 197-217.

[6] HAMMARBERG, B.,FRITZELL, B., GAUFFIN, J.,SUNDBERG,J. and WEDIN, L; (1980), "*Perceptual and acoustic correlates of abnormal voice qualities.*", Acta Otolaryngologica, Vol 90, 441-451.

[7] LEBART, L.,MORINEAU, A. et FENELON, J.P. (1979), "*Traitement des données statistiques. Méthodes et programmes*", Dunod, Paris, 254 p.

[8] CAZES, P. (1976), "*Régression par boule et par l'analyse des correspondances*", Revue de Statistiques Appliquées,Vol 24, No 4, 5-22.

A MODEL FOR AUTOMATED SPEECH CORRECTION OF GERMAN VOWELS: A PILOT STUDY

Rudolf Weiss

Western Washington University
Bellingham, WA, USA

Antonio Arroyo

University of Florida
Gainesville, FL, USA

ABSTRACT

A model is provided by which automatic error detection of vowels could be accomplished using predictable and pedagogically pretermined environments and specific analysis routines.

1. INTRODUCTION

Despite advances in modern technology and computerized speech recognition, a device which automatically detects and provides for correction of pronunciation errors is still far in the distance. However, we believe it is possible with the application of basic phonetic principles and pedagogical techniques to create an automated device which may be of use to foreign language teachers and students.

Criteria for automated speech recognition were already discussed nearly twenty years ago [1]. For the last ten years there has been almost a preoccupation with automatic segmentation and labelling of speech sounds, as can be seen by the large number of papers in this area at the last congress. It is exactly in this area of sound segmentation where some of the greatest problems in contemporary phonetics lie. With the advent of digitizing techniques, great strides have been made in voice analysis and synthesis. However, it has not changed the speech act itself. The problem, as has been pointed out for decades and again quite recently, is that of determining segment boundaries [3]. The efforts to find discrete information equivalent to sounds in the myriad of signals emitted by the continuous speech act eludes phoneticians. Ever present elements of co-articulation, compounded by factors of individual production and physiology, as well as suprasegmental elements and changing temporal aspects manifested in the continuous speech signal, confound efforts to find "sounds" particularly in an automated and error-free fashion. Furthermore it has been demonstrated that even we humans have difficulty labelling sounds categorically (absolutely) but must instead rely upon the contrastive relationship of the environment [2].

It is therefore our firm belief that at this point in the development of technology, error detection/correction can more easily be successfully accomplished if efforts are goal-directed to specific predictable errors and if they can be pedagogically "framed."

2. TECHNIQUES

A variety of techniques have been used in speech recognition, particularly for segment labelling.

Basically the incoming acoustic signal has to be broken down at certain intervals and then matched to some preprocessed acoustic criteria. The techniques vary; often a step filter device is used, matching bands of spectral energy above or below a certain frequency to the existence of certain sounds [6,3]. Routines of this nature often resemble a Jacobsonian distinctive feature approach.

3. WEISS MODEL

The model which we propose attempts to avoid some of the pitfalls inherent in conventional speech recognition processing. In application to foreign language teaching (or speech correction), certain assumptions have to be made:

- There are spectrographically identifiable and definable segments (sounds) in natural speech. We will create predictable environments for the ease of processing these targeted sounds.
- There are always features of co-articulation (transitions, alterations, etc.) in natural speech. These are also largely definable and predictable by the environment. By predetermining the environment we will be able to circumvent most of the problems co-articulation features might present.
- There are acoustic characteristics as well as idiosyncratic articulatory habits of each speaker. These are less predictable and no model can totally accommodate them.

A computerized model, using a Mac II and MacSpeech Lab II, or a comparable speech work station would work in the following manner:

- A correct utterance produced by a native speaker has been digitized and stored. The student is given a screen prompt and the digitized utterance is played.
- A prompt appears on the screen to repeat the word.
- The student's response is then digitized.
- The computer processes a hierarchy of matching routines (based on matching the digitized information).
- If an error is made, the student is prompted as to the nature of the error. The correct utterance is given again, and the student is prompted to repeat the utterance.
- The computer reprocesses, matches and gives error statements until the student responds correctly or gives up.

There are two real limitations to the functional success of such a model.

- The processing and computer response (digitizing, analyzing and matching routines) must be very fast (ideally < 1.5 sec) to be of practical use. Otherwise the nature of the student's production is likely forgotten.
- The model will work only if the errors are predefined and predictable and if the environment is completely controlled and chosen to facilitate ease of computer processing.

This model is an outgrowth of previous work on computer assisted diagnosis of vowel perception and a phonetics manual written by the author in which specific anticipated errors and exercises for overcoming these errors are provided for each sound [4,5]. This contrastive and predictive approach can be applied to our model as follows. If the target sound for practice is German [e:], the anticipated errors of production will fall into three primary categories for American learners of German:

- the tendency to produce the vowel with too short duration;
- the tendency to diphthongize;
- the tendency to produce a vowel of the incorrect quality (either too high or too low).

Potential errors in the articulation [e:] and their acoustic manifestations are illustrated in the following chart.

CHART 1: Errors for [eː]

PRODUCTION ERRORS	ACOUSTIC MANIFESTATION
[e] too short pedagogically needs minimal [extended] length.	F_2 of < 1.5 sec.
[eʲ]/[ɛʲ], etc. diphthongization	F_2 change of > ± 50 Hz (in > .15 sec)
quality errors:	
a. [i(ʲ)] too high (result of perception studies) with or without diphthongization	F_2-F_1 = >2K or F_2 = >2.5K (in <.15 sec)
b. [ɛ(ʲ)] too open/low with or without diphthongization	F_2-F_1 = <1.7K or F_2 = <1.9K (in >.15 sec)

4. PROCEDURE

For practical considerations, the utterance ['beː tən] is chosen for emulation. This choice facilitates computer analysis routines since the VOT of [b] corresponds closely to full consonantal release. Digitized samples of correct and incorrect production serve to develop the bases for the matching routines which perform the analysis functions.

Analysis at the first evidence of a harmonic wave (release spike/VOT of [b]) continues until the cessation of the harmonic wave (i.e., onset of [t]). The first and last 30 ms of the vowel are considered transitional and thus omitted from the LPC analysis.

The acoustic manifestations of the errors are processed in the sequence shown on the above chart.

- **Length.** If the wave length is less than 150 ms the computer does not process the signal and the student is prompted that the vowel is too short and is requested to produce the utterance again.

Only if the vowel is longer than 150 ms is the next routine enacted:

- **Diphthongization.** If a shift of more than 100 Hz is detected in the F_2 frequency during the steady state portion of the vowel, a message signaling diphthongization error is given and the student prompted to repeat the utterance.

If the vowel is produced with adequate length and without detectable diphthongization, then the last routine is enacted:

- **Quality.** If the figure F_2-F_1 is more than ±150 Hz. from a predetermined figure, an error message related to quality is given. Messages of too high or too low tongue position would be prompted depending on the type of subroutine triggered by the error.

Each time an error is made, the student is prompted as to the nature of that error according to the error routine triggered by the analysis of the digitized student response. Repetitions are elicited until no error matching routines are triggered at which time the utterance is deemed to have been correctly rendered.

5. IMPLEMENTATION

LPC analysis of a correct model and five potential error types were verified using the MacSpeech Lab II program indicated above. Based on these analyses, a set of criteria for the automated rating of these utterances was developed. Using a file exchange utility, the binary data was converted to the IBM PC format. A FORTRAN program was developed to 1) reverse the two data bytes necessary for software compatibility and 2) trim the data samples to isolate the start and end of the voiced portion of the utterance.

A prototype data analysis system was developed using MATLAB as a development environment. This system is compatible with virtually every popular platform. Data written as a 2-column ASCII array with FORTRAN is first imported into MATLAB using its "load" command. Then MATLAB scripts perform the analysis of the data and error detection according to a hierarchically arranged ranking order. Analysis is carried out of segments or windows of 30 ms in duration on the basis of estimated F_1 and F_2 frequencies.

Although MATLAB uses about 20 seconds to calculate a 256-point LPC, the calculations of formants and error criteria requires less than 5 seconds on a MAC II and a 12 MHz 285 PC-AT. By upgrading to a 386 environment with a 25 MHz clock speed, initiating onset of analysis at onset of voicing, and using "custom" software in assembly language to optimize performance, the error feedback time to students could be reduced to within the 1.5 second time window pedagogically needed.

6. CONCLUSION

Initial results have shown that the described model could be an effective pedagogical tool to enable error correction. Its proven functional success rests upon the predictability and normability of errors and specifically designed error-matching routines. This model does not depend upon full-spectrum matching routines. It may thus be the closest we can come to an automated phonetician at this time.

N.B. Much of the work for this paper was accomplished at IASCP at the University of Florida in Fall 1989.

7. REFERENCES

[1] FLANAGAN, J.L. (1972), "*Speech analysis, synthesis and perception*", Berlin: Springer Verlag.

[2] REPP, B., et al. (1979), "Categories and context in the perception of isolated steady-state vowels", *Journal of experimental psychology, human perception and performance,* 5 (1) 129-145.

[3] ROACH, P., et al. (July 1990), "Phonetic analysis and the automatic segmentation and labelling of speech sounds," *Journal of the international phonetic association,* 20 (1), 15-21.

[4] WEISS, R. (1987), "Computer-assisted diagnosis of perceptual errors", *Proceedings of the XIth international congress of phonetic sciences,* Tallinn: Academy of Sciences of the Estonian S.S.R., 295-297.

[5] WEISS, R. and H-H. WAENGLER. (1985), "*German pronunciation: a phonetics manual*", Bellingham: Western Washington University Press.

[6] ZWICKER, E., TERHARDT, E. AND E. PAULENS. (Feb. 1979), "Automatic speech recognition using psycho-acoustic models", *JASA,* 65 (2), 487-498.

LE FRANCAIS DE GERMANOPHONES DEBUTANTS TOPOLOGIE GRAPHIQUE DE DISTORSION

D. Lefevre

Docteur en Phonétique

ABSTRACT

The aim of this paper is to present an analysis method, of how german speaking natives reproduce french phonemes, as well as the phonemes uttered instead.
It points out, besides several matters, the importance of the differents factors and the difficulties in using different levels of enonciation.
Besides, it emphazises the importance of interlanguage facts and the ability in establishing a synthetical system of the different phonemes.
Finally, it makes an attempt in setting forth some answers to allow a better comprehensiveness of french language by German advanced beginners.

1. INTRODUCTION

Le corpus est un enregistrement de phrases produites par un professeur de français et répétées par un ou plusieurs locuteurs germanophones dans le cadre de groupes d'apprentissage débutants avancés.
Deux groupes ont été ainsi enregistrés permettant la constitution de trois cassettes de 90 minutes.
Le matériel utilisé est un magnétophone à cassette de type MARANTZ C.D. 330 STÉRÉO DOLBY, cassette de type TDK MA90 IEC IV TYPE IV MÉTAL POSITION et de deux microphones unidirectionnels, l'un dirigé vers le professeur et l'autre dirigé vers les locuteurs.
Les locuteurs sont d'origine du Bade Würtemberg. Deux dans le premier groupe et un dans le second, ils travaillent tous dans des entreprises en tant qu'employés.
Le dépouillement s'est fait à l'oreille d'un seul transcripteur. Il y a eu plusieurs écoutes.

2. ANALYSE

2.1. Comparaison de chaque idiome

Les systèmes de l'allemand et du fran çais se différencient de la manière suivante
- Un assourdissement de tout élément consonantique en position final de syllabe [1].
- Une attaque progressive des voyelles en français et brusque en allemand [2].
- La présence d'une seule voyelle [ə] en syllabe inaccentuée en allemand [3].
- La présence de voyelles hautes ouvertes en allemand inexistantes en français [4].
- L'absence de voyelles longues relâchées ouvertes en allemand [4].
- Une unité accentuelle libre hiérarchisée en allemand [4].
- Une unité accentuelle fixe en français [5].
- Un rythme plat sur les syllabes inaccentuées et montant ou descendant sur les syllabes accentuées en français [2].
- Un rythme en escalier sur l'avant-dernière syllabe en allemand [2].

2.2. Étude des distorsions.

Les distorsions relevées sont de quatre ordres:-mémorielle -syntaxique -lexical -phonétique
Dans cet article, seules les distorsions d'ordre phonétique seront évoquées

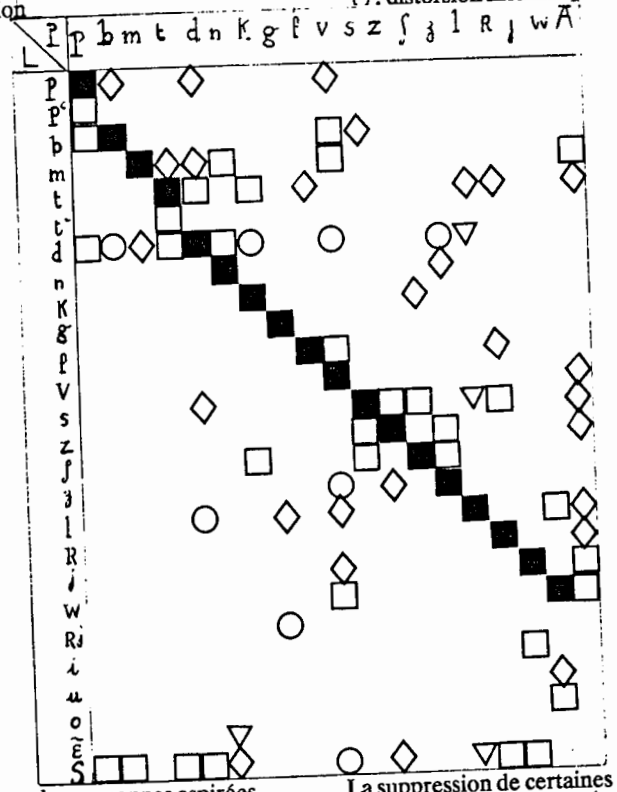
2.3. Répartition des distorsions.

Les distorsions phonétiques suivent l'axe diagonal d'adéquation entre l'énonciation du professeur et la reproduction ou l'énonciation des locuteurs.
Les locuteurs utilisent en moyenne pour les consonnes près de trois autres phonèmes alors que pour les voyelles ils emploient en moyenne un seul autre phonème.
L'établissement des distorsions a permis de dresser deux graphes de distorsion

2.3,1 Le système des consonnes

Légende
L : Locuteurs
P : Professeur
S : Suppression
A : Apparition

■ : adéquation
□ : distorsion phonétique
◇ : distorsion lexicale
▽ : distorsion syntaxique
○ : distorsion mémorielle



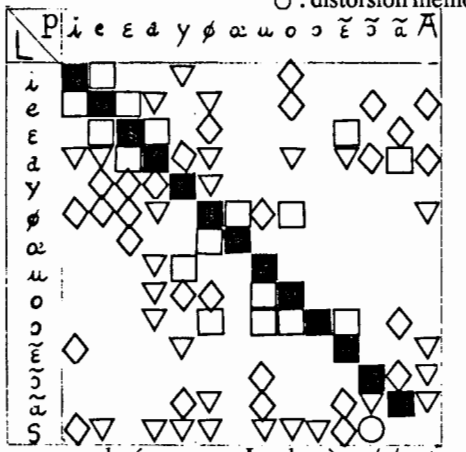
La présence des consonnes aspirées, chez les locuteurs, est due au maintien du système phonologique de leur idiome lors de la reproduction.
La présence de la voyelle /ə/ est due à un problème syntaxique.
La présence de la voyelle /i/ au lieu de /j / essentiellement en finale confirme le constat que les locuteurs germanophones interprètent la finale /j / précédée d'une voyelle comme une finale diphtonguée [2].
La présence de la voyelle /u/ au lieu de /w/ est due à la difficulté de prononcer ce phonème comme d'ailleurs en anglais [5] et à sa palatisation lorsqu'elle est suivie de la voyelle /i/ par le professeur.

La suppression de certaines consonnes par les locuteurs se fait principalement: en initiale de groupe, en position implosive, en position subséquente: soit en initiale soit en finale et en position intervocalique pour le phonème /j/.
L'apparition de certaines consonnes se fait en finale de mots pour les phonèmes /z/ et /s/ par réaction orthographique, en position subséquente pour les phonèmes /v/ et /w/ derrière /s/ et /ʁ/ derrière /b/, en position finale de syllabe pour /m/ lorsque celle-ci se termine normalement par une voyelle nasale.
Le phonème /p/ n'est pas mentionné car comme constaté [6], il est remplacé par la suite /n / + /j / chez les francophones.

2.3,2 Le système des voyelles.

Légende
L : Locuteurs
P : Professeur
S : Suppression
A : Apparition

■ : adéquation
□ : distorsion phonétique
◇ : distorsion lexicale
▽ : distorsion syntaxique
○ : distorsion mémorielle



Le phonème /i/ se trouve remplacé par le phonème /e/ soit en position finale absolue soit par métathèse double pour le couple: e-i devenant i-e pour le verbe /dezire/ qui devient /dizere/ lorsqu'il y a interrogation.

Le phonème /e/ se trouve remplacé par le phonème /ɛ/ en position finale absolue ou par dédoublement de la voyelle /i/ se substituant à la voyelle /e/ pour le verbe /dezire/ qui peut devenir /dizire/ dans le cas de l'interrogation.

Le phonème /ɛ/ se trouve remplacé par le phonème /e/ en position finale fermée soit par la consonne /s/ soit par la consonne /j/ qui chez les locuteurs est devenu /i/. Il peut- être remplacé par /a/ lorsqu'il est suivi de la consonne /j/ par les locuteurs germanophones comme dans /butaje/, /paje/ au lieu de /butɛj/ et /peje/.

Le phonème /a/ est remplacé par le phonème /ɛ/ en position finale fermée comme dans /bɛt/ au lieu de /plat/.

Le phonème /ø/ se trouve remplacé par le phonème /œ/ en position finale fermée de syllabe. Il peut-être dissimilé en /ɑ̃/ comme dans le mot /møsjø/ qui devient /mɑ̃sjø/.

Le phonème /œ/ se trouve remplacé par le phonème /ø/ en position finale fermée chez les locuteurs germanophones comme dans /anøvœ/

Le phonème /u/ se trouve remplacé par le phonème /ɔ/ par dissimilation partielle comme dans /nuvɔdrijɔ̃/ au lieu de /nuvudrijɔ̃/.

Le phonème /o/ se trouve remplacé par le phonème /ɔ/ parfois en position finale de syllabe ouverte et par le phonème /ø/ parfois, en position finale absolue ou par métathèse double comme dans /medølø/ au lieu de /medølo/.

Le phonème /ɔ/ se trouve parfois remplacé par le phonème /o/ en position finale ouverte de syllabe.

Le phonème /ɛ̃/ se trouve parfois remplacé par le phonème /ɛ/ ou par le phonème /a/ en position ouverte absolue.

Le phonème /ɑ̃/ se trouve parfois remplacé par le phonème /a/ en position finale absolue ou ouverte de syllabe.

Les locuteurs germanophones font parfois apparaître le phonème /ə/ en finale de groupe rythmique. Celui-ci correspond sans doute au "schwa" que nous n'avons pas considéré car, s'il appartient au système phonologique des locuteurs, il pose des difficultés au niveau de son intégration dans le système du français, au niveau de son statut phonologique, au niveau de sa stabilité, au niveau de sa perception car même phonétiquement, il ne se différencie pas de /ø/ et de /œ/ [6].

On notera l'absence du phonème /ɑ̃/ chez le professeur car il n'a plus de valeur

3. SYNTHÈSE

3.1. Les consonnes

Les causes phonétiques de distorsion sont:

- La distribution en fonction de la syllabe et la nature inverse de celle-ci, 80% de syllabes ouvertes en français et 70% de syllabes fermées en allemand [2].
- Le besoin de surmonter les difficultés en obtenant une monotonie de rythme par dédoublement ou dissimilation de certaines consonnes [1].
- Le maintien du système phonologique d'origine qui provoque d'une part la présence de consonnes occlusives sourdes où les locuteurs ont tendance à émettre une légère expiration comme pour /p/ et /t/ et un assourdissement des consonnes sonores /b/, /d/, et /g/ neutralisant l'opposition sourde/sonore en finale [7].
- La position des organes différente entre les consonnes du français et de l'allemand, notamment la position de l'apex de la langue ou du dos de la langue en ce qui concerne par exemple les consonnes alvéo-dentales.
- La différence d'énergie articulatoire existe entre les consonnes du français et de l'allemand comme par exemple pour les consonnes occlusives bilabiales.
- La position implosive de certaines consonnes suivies d'occlusives qui provoquent leur disparition comme généralement constaté pour les idiomes romans [8].

3.2. Les voyelles

Les causes phonétiques de distorsion sont:

- La distribution en fonction de la syllabe d'où la neutralisation de certaines oppositions et le remplacement des voyelles ouvertes par des voyelles plus fermées.
- Le besoin de surmonter les difficultés en obtenant une monotonie de rythme par dédoublement en métathèse double notamment pour le couple i-e et en particulier dans le cadre de la phrase interrogative.
- Le maintien du système phonologique d'origine qui provoque surtout la fermeture en position inaccentuée des voyelles ouvertes car en allemand le timbre des voyelles dépend de leur longueur.
- L'absence de certains phonèmes dans le système d'origine notamment les voyelles nasales qui se trouvent de ce fait soit dénasalisées en /ɛ/ pour /ɛ̃/ et /a/ pour /ɑ̃/ et suivie des consonnes /n/ ou /m/ soit renforcées par une consonne nasale.

4. CONCLUSION

Pour améliorer l'apprentissage du français, il peut-être bon de:

- Dès la première leçon faire pendant environ 10 minutes un exercice intonatif pour introduire les phrases interrogatives simples avec "est-ce-que" et les réponses soit positives soit négatives. Cet exercice pourra-être effectué à chaque introduction de phrase plus complexe. On évitera au départ l'intonation expressive.
- Au cours des dix premières leçons, on fera avec les mots appris des exercices de distributions notamment pour les consonnes finales sonores et à l'initial pour le phonème /s/ afin qu'il ne soit pas reproduit en /z/.
- Lors de l'introduction de verbes comme "aller", on pourra faire des exercices d'articulation qui concernent les voyelles nasales en effectuant une gamme d'oppositions du type : "vont", "vent", "vin" et des exercices pour bien opposer les phonèmes /w/ et /v/.

5. RÉFÉRENCES

[1] BOTHOREL A., SIMON P., WIOLAND F., ZERLING J-P., (1986), "*Cinéradiographie des voyelles et consonnes du français.*" Strasbourg Institut de Phonétique

[2] BOURCIEZ E. et J. (1982) "*Phonétique Française. Etude Historique*", Paris Klincksieck

[3] GARDE P. (1968), "*L'accent* " Paris Presse Universitaires de France

[4] KENWORTHY J. (1987), "*Teaching English pronunciation*" London New-York Longmann

[5] LÉON P.&M. (1976), "*Introduction à la phonétique corrective*", Hachette & Larousse

[6] MALMBERG B. (1972) "*Les nouvelles tendances de la linguistique*", Paris Presses Universitaires de France

[7] MALMBERG B. (1974), "*Manuel de phonétique générale.*", Paris Picard A.J.

[8] MARTENS P. (1989), "*Quelques problèmes concernant la réalisation des voyelles et des diphtongues de l'allemand par un francophone*", MELANGES de Phonétique générale et expérimentales offerts à Péla SIMON 563-584.

GERNALIZATION OF NEW SPEECH CONTRASTS TRAINED USING THE FADING TECHNIQUE

Donald G. Jamieson and April E. Moore

University of Western Ontario, London, Ontario, Canada

ABSTRACT

Subsequent to training with synthetic speech using a fading technique, we found substantial improvements in the ability of unilingual francophone adults to identify voiced and voiceless English "th" sounds, presented in a vowel-consonant-vowel (VCV) format. Improvements were seen for the tokens used in training, for natural utterances by two speakers, and when listening in noise as well as in quiet. Smaller improvements occurred for target sounds in other word positions (ie., VC or CV context) or in other vowel environments.

1. INTRODUCTION

While adults may develop a sophisticated command of a new language, difficulties often persist with the perception and pronunciation of phonemes which are foreign to the person's first language. A common example of this phenomenon is seen in Canadian Francophone adults, who have difficulty distinguishing and pronouncing the English voiced (V^+) and voiceless (V^-) fricatives /θ/ and /ð/, often substituting them with /t/ and /d/, which are present in the phonemic repertoire of French.

The present study sought to examine some of the limits of a method by which adults can be trained to perceive non-native contrasts [1][5]. Subjects' abilities to identify and discriminate the synthetic voiced and voiceless English "th" sounds improved significantly after only 90 minutes of training, and generalized from the synthetic stimuli used in training to *natural* speech counterparts of /ð/ and /θ/ in CV syllables recorded by male and female talkers. Training effects transferred across different voices, but did not generalize to the same contrasts presented in different positions within the word, nor to the /ð/-/d/ contrast.

The present work explored the potential to increase generalization through a minor procedural modification: training with a VCV continuum, rather than a CV continuum, to provide acoustic cues associated with the formant transitions for *both* the preceding vowel and the following vowel. It was hypothesized that training would transfer to trained and non-trained word position and possibly endure when vowel context was altered.

2. METHOD

Our training paradigm employs the perceptual fading technique introduced by Terrace [6]. The goal is to train a perceptual contrast with a minimum amount of difficulty and few errors, by beginning training with an exaggerated exemplar of the feature being trained -- in this case, V^+ or V^- frication -- and providing immediate feedback. The distinction being trained is perceptually salient initially, but becomes progressively more subtle as training progresses.

2.1 Stimuli

2.1.1 Training Stimuli. Training used 8 VCV speech segments, synthesized at 20 kHz using an implementation of Klatt's [4] cascade/parallel speech synthesizer for IBM AT computers [3].

The 8 stimuli formed a VCV continuum with the consonant varying from the voiced interdental fricative /ð/ to the voiceless interdental fricative /θ/. The neutral vowel, /ʌ/, was used in both the initial and final positions in all training stimuli. The parameter values used to generate these sounds were based on those used in [1].

Vowel duration was fixed at 135 ms and 210 ms for initial and final positions, respectively; the duration of frication varied from 360 ms in stimuli 1 and 8, and to 90 ms in stimuli 4 and 5, respectively, decreasing by 90 ms for each consecutive stimulus.

2.1.2 Test Stimuli. Three sets of test stimuli were used in pretesting and posttesting. The *synthetic VCV syllables* were the same stimuli used in training. The *natural speech nonsense syllables* included 8 VCVs, 8 VCs, and 8 CVs. Within each subset, each of the English sounds /t/, /d/, /θ/, and /ð/ appeared once with the vowel /ʌ/, (as in training), and once with the vowel /i/ (where formant transitions were dissimilar). All tokens were spoken by 2 native speakers of Canadian English, 1 male and 1 female, and recorded, edited, and stored on disk using the CSRE software [3].

The 12 minimal-pair *word* stimuli contrasted /ð/ with /θ/; /ð/ with /d/; /θ/ with /t/; and /t/ with /d/ -- each, in -initial,-medial and word-final position. The word pairs used were: either-ether; riding-writhing; pity-pithy; wading-waiting; loathe-loath; fraught-froth; bade-bathe; bud-but; this'll-thistle; tinker-thinker; den-then; and doe-toe.

2.2 Subjects

Twenty-one unilingual Francophone college and university students (8 males and 13 females), participating in an English-language summer immersion program at the University of Western Ontario, were paid to serve as subjects. All placed in the lowest third of their class on an English language placement test and an oral language interview, and all passed a hearing screening at 20 dB HTL.

Pretest scores were used to construct one control and one experimental group, approximately matched in both pre-training perception skills, and gender.

2.3 Procedure

Pretesting was conducted one week, with training during the second week, and posttesting during the third week. Instructions were given in French, and only when it was clear that the task was understood did testing proceed.

2.3.1 Pretesting. Subjects were tested individually in a sound-attenuating booth, being seated at a small table facing a monitor on which were displayed response alternatives. Subject's listened to a signal over headphones, then identify the consonant portion as being either /d/, /t/, /θ/, or /ð/. On each trial, 4 words appeared on the monitor: THE, THING, DOG, TIME -- each containing one of the target sounds.

2.3.2 Training. Subjects were trained individually using the configuration described for pretesting. The task was to listen to a signal (one of the 8 synthetic sounds), then identify the consonant as being either /θ/ or /ð/ by selecting the word containing the target (THE, or THING). Immediately follow-

ing each response, the correct choice was illuminated to provide feedback. The initial stages of training were very easy, with stimuli selected from the opposite ends of the continuum. Progress from one stage to the next required 90% correct performance, on three consecutive blocks of trials. As training progressed, more medial stimuli from the continuum were used, so that the task became more difficult. During the final two blocks of training sounds, were presented sounds in a background of speech babble, to simulate many real life listening situations. Subjects were tested for one hour periods with breaks between blocks. Only two subjects failed to complete training within the alloted 4 hours.

2.3.3 Posttesting. Posttesting stimuli and procedures were identical to those used in pretesting.

3. RESULTS

3.1 Synthetic Tokens

Data were first reduced to the proportion of each identification response made by each listener for each stimulus under each test condition. There was a clear bias on the part of all subjects to choose *voiced* responses ($t = -2.353$, $df = 19$, $p = .014$) for pre-test, ($t = -2.958$, $df = 19$, $p = .004$) scores) for posttest. To allow performance to be measured independently of such biases, identification responses were converted to A' scores, using each subject's hit rate with a given stimulus type (e.g., "*TH*ING" responses when /θ/ stimuli were presented), in combination with that subject's overall error rate on all stimuli of the opposite type (e.g. "/θ/" responses to presentations of voiced stimuli) as the False Alarm rate. Pairing was effective in terms of equating the two groups at pretest ($t = -.843$, $df = 38$, $p = .404$).

A' pretest scores were subtracted from posttest scores to arrive at an A' difference score indicating the change from pretest to posttest. The A' difference provided the basis for further analysis. Posttest scores were significantly higher than the pretest scores for the trained group, ($t = 3.814$, $df = 9$, $p = .002$, one-tailed) and for the control group ($t = 2.396$, $df = 9$, $p = .020$, one-tailed), but the trained group improved more than the control group from pretest to posttest ($t = 1.935$, $df = 9$, $p = .042$, one-tailed).

3.2 Natural Nonsense Syllables

Nonsense syllables varied in their similarity to the synthetic training stimuli by syllable structure (VC, CV, or VCV), vowel environment (/ʌ/ vs. /i/), consonant (/ð, θ, t, d/), and talker (male vs. female). The pretest scores of the 2 groups did not differ ($t = .417$, $df = 18$, $p = .681$).

Analyses on subsets of the nonsense syllables demonstrated substantially different degrees of generalization for different types of stimuli. Training transferred directly to the natural /ʌðʌ/ and /ʌθʌ/ tokens, with an overall improvement in performance for both talkers, under both noisy and quiet conditions ($t = 2.68$, $df = 9$, $p = .013$ for the trained group and $t = 2.06$, $df = 9$, $p = .034$ for the control group). A greater improvement occurred for the trained group ($F(1,9) = 4.83$, $p = .055$).

Less transfer occurred when syllable structure changed (ie., with /θʌ, ðʌ, ʌθ, and ʌðʃ) and in other vowel contexts (ie., with /iðɪ/ and /iθi/). Scores for the /ʌtʌ/ and /ʌdʌ/ tokens failed to show an effect of training ($F(1,9) = 2.11$, $p = .1805$), reflecting the good pretest performance which produced a ceiling effect, reducing the *possibility* for a training effect.

3.3 Natural Word Pairs

The 48 natural words varied in terms

of syllable structure (VC, CV, or VCV), vowel environment, consonant (/ð, θ, t, d/), and talker (male vs. female). For example, for the "either, ether" word pair (same syllable position, but in a different vowel context from that used in training), training generalized and trained subjects showed somewhat better performance than did control groups ($F(1,9) = 2.26$, $p = .0928$). However, with the word pair "loath, loathe" (different word-position and vowel environment), training did not generalize, and trained subjects did not differ from control subjects ($F(1,9) = .33$, $p = .570$).

4. DISCUSSION

The present findings display an orderly pattern of results. A' difference scores improved most for the identification tasks involving the syllables /ʌθʌ/ and /ʌðʌ/, which are identical both in structure and in phonemic content to the synthetic training stimuli, next for nonsense syllables and words in which the syllable structure and consonant were held constant while the vowel environment differed from the training stimuli, and least for conditions involving altered syllable structures (CV and VC) and non-trained homorganic phonemes (/t/ and /d/). Consistent with previous research [2], speaker sex did not affect listeners' ability to perceive the non-native phoneme contrasts on which they were being trained.

5. REFERENCES

[1] JAMIESON, D.G., & MOROSAN, D. (1986), "Training non-native speech contrasts in adults: Acquisition of the English /ð/-/θ/ contrast by francophones", *Perception & Psychophysics, 40*, 205-215.

[2] JAMIESON, D.G., & MOROSAN, D. (1989), "Training new, nonnative speech contrasts: A comparison of the prototype and perceptual fading techniques", *Canadian Journal of Psychology, 43*, 88-96.

[3] JAMIESON, D.G., NEAREY, T., & RAMJI, K. (1989), "CSRE: The Canadian Speech Reasearch Environment", *Canadian Acoustics, 17*, 23-35.

[4] KLATT, D.H. (1980), "Software for a cascade/parallel formant synthesizer", *Journal of the Acoustical Society of America, 67*, 971-995.

[5] MOROSAN, D. & JAMIESON, D.G. (1989), "Evaluation of a technique for training new speech contrasts: Generalization across voices, but not word-position or task", *Journal of Speech and Hearing Research, 32*, 501-511.

[6] TERRACE, H.S. (1963), "Discrimination learning with and without 'errors'", *Journal of the Experimental Analysis of Behavior, 6*, 1-27.

ACKNOWLEDGEMENTS

We are grateful to M.F. Cheesman, J. Booth, K. Ramji and W. Allsop for advice and assistance. The project was supported, in part, by grants from the NSERC, URIF, and from Unitron Industries Ltd. Address correspondence to Dr. D.G. Jamieson, Hearing Health Care Research Unit, University of Western Ontario, London, ON, CANADA, N6G 1H1.

PERCEPTION AND PRODUCTION OF ITALIAN PLOSIVES BY AUSTRIAN LEARNERS

H. Grassegger

Abteilung für Phonetik
Institut für Sprachwissenschaft, Graz, Austria

ABSTRACT

Two experiments on the differentiation of bilabial plosives were carried out to test the hypothesis that articulatory difficulties have a perceptual correlate which may function as a diagnostic tool. An identification test on a synthesized 10-stimulus continuum with different VOT-values ranging from -80 to +64 msec showed listener specific subcategorisations. These subcategorisations are reflected in the individual voicing and aspiration parameters of the subjects' realisations of bilabial plosives in initial prevocalic position.

0. INTRODUCTION

0.1. Previous studies on the production of plosives by speakers having an Austrian variety of German as their native language (see [1], pp. 209 - 234) yielded the following facts: a) Lenis plosives (/b,d,g/) in initial prevocalic position were produced without voicing. b) Articulatory differentiation of initial prevocalic lenis and fortis plosives resulted in shorter voicing lags for lenes, longer voicing lags for fortes. Average differences amounted to 10 msec for labials (/b/: 20 msec - /p/: 30 msec), 20 msec for dentals (/d/: 25 msec -/t/: 45 msec), and 55 msec for velars (/g/: 30 msec - /k/: 85 msec).

0.2. It is the differentiation of (voiced) lenis and (unaspirated) fortis plosives which is specifically problematic for Austrian (as well as for German) learners of Italian. Measurements of Italian voiced plosives (here again: in initial prevocalic position) revealed a considerable voicing lead around 80 msec (see [1], p. 149).

1. HYPOTHESIS

1.1. It is now hypothesized that the traditional difficulties which Austrian speakers show in adequately pronouncing Italian voiced and voiceless plosives of the same place of articulation are not only to be explained by differences in the sound pattern of source and target language but that these difficulties have a perceptual correlate. That is to say that an individual learner's perceptual habits would presumably reveal his efficiency in articulation.

1.2. In this study perception and production of the contrast between (initial prevocalic) /p/ and /b/ by Austrian subjects learning

TABLE 1: Perception test

Stimulus nr.	1	2	3	4	5	6	7	8	9	10
VOT (msec)	-80	-64	-48	-32	-16	0	+16	+32	+48	+64
GERMAN TEST GROUP										
Listener 1	48	49	47	46	42	39	*30	*19	7	4
Listener 2	50	50	40	*31	*24	18	15	2	0	0
Listener 3	47	49	47	41	40	36	32	/ 18	16	2
Listener 4	48	44	46	45	39	40	37	/ 18	5	0
Listener 5	46	47	45	43	36	34	*31	*30	10	0
Listener 6	49	48	46	44	41	38	*28	18	3	2
Listener 7	49	50	49	32	*20	18	10	2	0	0
Listener 8	48	47	42	41	42	40	*30	18	0	0
Listener 9	50	50	48	*27	*24	14	13	0	1	0
Listener 10	47	44	44	43	38	34	33	/ 11	2	0
ITALIAN CONTROL GROUP										
Listener 1	50	50	38	*19	12	5	3	0	0	0
Listener 2	49	50	34	/ 16	8	2	4	1	0	0
Listener 3	50	50	35	*22	10	4	2	0	1	0
Listener 4	49	49	32	*20	15	7	3	0	0	0
Listener 5	50	49	37	/ 17	11	3	4	0	0	0

Italian is examined. The acoustic dimension that has been decided to be most useful for this purpose is voice onset time (VOT).

2. PERCEPTION TEST

2.1. Test Structure. A series of 10 /Ca/ stimuli with VOT values ranging from -80 to +64 msec in 16 msec increments (Fo 125 Hz) were synthesized using the synthesis system Flexivox (for details on the system see [2]). Spectrum and intensity of the release burst (320 msec prior to the end of the vowel) represented an average of values used in the common text-to-speech synthesis of bilabial (/b-/ and /p-/) stops. Each synthesized stimulus was recorded on a Revox tape recorder (B 77MK II) 5 times in randomized order with a fixed interstimulus interval of 2 seconds. The resulting block of 50 stimuli was then copied 10 times with a 10-second pause after each block (thus yielding 50 tokens of each stimulus) and administered to ten Austrian learners of Italian, whose experience of Italian ranged from one to four years (German test group) and to five native listeners from Northern Italy (Italian control group). The stimuli were presented binaurally over headphones (PIONEER SE-305) at about 70 dB SPL(A). A practice block consisting of two repetitions of the 10 stimuli in ascending order of VOT was presented before the experiment for familiarization with the synthetic voice. The subjects were told that each stimulus represented the beginning of either Italian 'palla' ("ball") or 'balla'("bale"). They had to indicate their judgement by circling 'palla' or 'balla' on an answer sheet. Moreover, they were asked to guess in case of uncertainty.

2.2. Test Results. The results for both groups of listeners are given in table 1 as the number of /b/-identifications for each stimulus. Values not significant according to a chi-square test ($19 \leq x \leq 31$, $n = 50$, $p = 0.05$) are marked by an asterisk. In case of lacking asterisked values a slash separates significant /b/-judgements (left of the slash) from significant /p/-judgements (right of the

slash). In all other cases numbers preceding asterisked values represent significant /b/-identifications, those following significant /p/-identifications. Diagrams of the individual results rendered the basis for interpolation of VOT-values that represent the 50% crossover of judgement. With all Italian listeners and with three of the German listeners these crossover VOT-values above which more than 50% of the tokens of one stimulus were judged to be voiced lie within the voicing lead, for the remaining German subjects it reaches far into the voicing lag (cf. table 2).

TABLE 2: Interpolated cross over VOT-values

German test group				
L 1	L 2	L 3	L 4	L 5
+22	-22	+22	+21	+36
L 6	L 7	L 8	L 9	L10
+25	-23	+23	-19	+24
Italian control group				
L 1	L 2	L 3	L 4	L 5
-37	-40	-36	-39	-38

It was for this reason that the German test group was finally divided in two subgroups according to their cross over points: subgroup 1 (comprising listeners 1, 3-6, 8, 10) with crossover in the voicing lag, subgroup 2 (listeners 2,7,9) with cross over in the voicing lead. In figure 1 the mean values of /b/-judgments for the two German subgroups and the Italian control group are plotted for comparison.

The subcategorisation of the 10-stimulus continuum is seen to be shifted to the right by German listeners compared to the Italian group, more strongly so for subgroup 1 than for subgroup 2. As a result, stimuli with no voicing lead or even with a voicing lag up to about 24 msec are perceived as acceptable realisations of the beginning of Italian 'balla' by subgroup 1. Subgroup 2 more rigorously rejects stimuli without any voicing lead as 'ba' but does not reach the cross over values of the Italian listeners.

3. PRODUCTION TEST

For the production test the same ten subject of the German test group read a list of ten Italian words beginning with voiced and voiceless bilabial stops which included the two words 'balla' and 'palla' and repeated it five times. The subjects' speech was recorded on a tape recorder (Revox B 77MK II). The five realisations of 'balla' and 'palla' respectively were then analyzed by means of a digital oscilloscope (Nicolet 3091), which allowed easy measurement of VOT-values in steps of 0.2 msec. Measurement reliability was assessed by remeasuring ten randomly selected realisations. The mean difference between the two measurements of each stop was 1.2 msec with a range of 0.8 msec.

The results of the production test are given in table 3 as mean values of the five /b/- and /p/-realisations.

Only those speakers who are members of the German subgroup 2 according to the perception test show considerable voicing leads in their /b/-realisations.

TABLE 3: Production test VOT-values in msec

	/b/		/p/	
Sp	avg	sd	avg	sd
1	+20.6	2.4	+30.8	2.4
2	-27.8	3.2	+27.8	1.9
3	+21.4	2.1	+27.2	1.5
4	+19.6	3.8	+24.2	1.6
5	+26.6	3.2	+38.8	1.6
6	+18.8	2.4	+32.8	1.9
7	-37.6	3.9	+26.4	1.4
8	+23.4	2.0	+30.8	2.4
9	-32.4	2.4	+19.8	1.9
10	+22.4	2.9	+37.6	2.7

4. CONCLUSION

The results of the perception and production test on the one hand show the expected difference in categorisation ability between native speakers and learners of the language and, moreover, clear differences among the learners themselves. On the other hand they demonstrate most convincingly that well-established perceptual categories are more likely to be accompanied by more acceptable production.

Thus, our hypothesis on the interrelationship between perceptual and articulatory abilities seems to be proved. The link between perceptual categories and articulatory differentiation is an encouraging indication that suitable perception tests provide a powerful diagnostic tool in pronunciation teaching.

5. REFERENCES

[1] GRASSEGGER, H. (1988), *"Signalphonetische Untersuchungen zur Differenzierung italienischer Plosive durch österreichische Sprecher"*, Hamburg: Buske Verlag.

[2] OLASZY, G., GORDOS, G. (1988), "Die Anwendungen des Flex-Deutsch Sprachsynthesesystems in phonetischen Forschungen", *Hungarian Papers in Phonetics, 19*, 34 - 46.

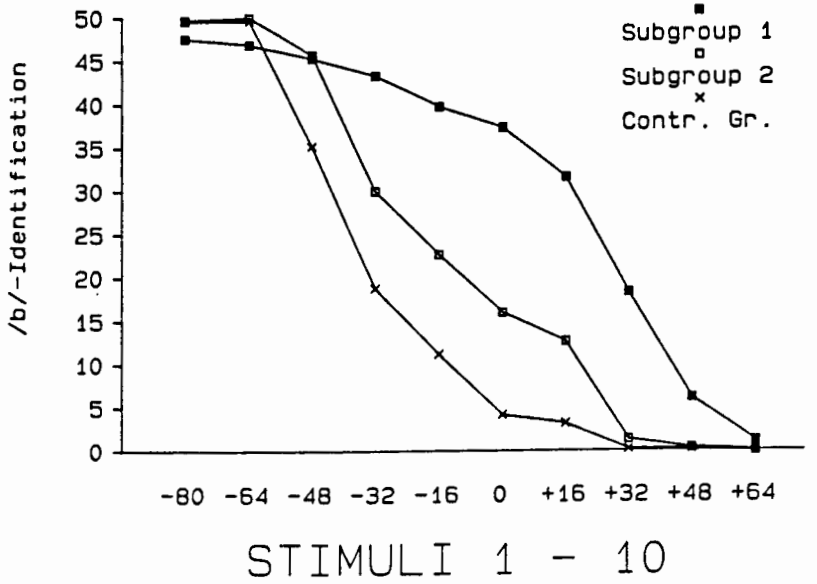


FIGURE 1: Mean values of /b/-judgements

STUDIES OF METHODS FOR THE MEASUREMENT OF SPEECH COMPREHENSION

Robert McAllister and Mats Dufberg

Dept. of Linguistics, Stockholm University, Sweden

ABSTRACT

This paper is a report on ongoing research on the explicit measurement of speech comprehension. Our point of departure was a need for reliable speech comprehension tests with a higher degree of validity than existing meaurement instruments. After several experimental studies which will be summarized here, we have developed a global comprehension test called the Question and Response (QAR) test. Using hard of hearing, second language learners as listeners/subjects we are at present attempting to asses the utility of this test as an instrument for the measurement of comprehension ability.

1. INTRODUCTION AND BACKGROUND

This paper is about the measurement of the ability to understand spoken language. This faculty has, in earlier research, been refered to by means of several different terms. We have chosen the term "speech comprehension" and wish to take a global approach to the definition of this ability. Figure 1 [5] is a graphic representation of some important aspects of one current view of speech comprehension upon which the work reported here is based. It is a very general model of the relationship between two major sources of information used by the listener to interpret a spoken utterance. One of these is the information contained in the speech signal represented in figure 1 as *signal dependent information* and often refered to in the past as "acoustic cues". The other source of information represented in the figure as *signal independent information* refers to the knowledge of the language spoken, knowledge of the world, the current communication situation, etc.. This information creates expectations on the part of the listener as to what meanings are to be communicated and thus greatly facilitates understanding of the message. It should be noted that while we are aware of the important roll played by visual information in the interpretation of spoken utterances [7], only the acoustic/auditory component has been considered in this work.

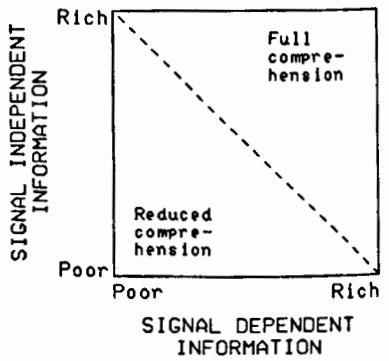


Figure 1. Model of the relationship between two major sources of information used by the listener to interpret a spoken utterance, upon which the work reported here is based.

Our point of departure was the discovery of the need for methods that could be used for the estimation of an individual's ability to understand language spoken in everyday situations. These methods would have several applications both clinical and pedagogical and could also be a useful research tool. There are a number of these tests used in various clinical settings including audiology and logopedics and phoniatrics as well as in the field of foreign and second language teaching where there is a need for methods for the measurement of learners comprehension ability and assessment of the quality of their pronunciation in terms of its comprehensibility. The problem with these existing methods, which is central to the work summarized here, concerns test validity.

The purpose of this work is to develop methods for the explicit measurement of speech comprehension with special consideration of content, construct, and concurrent validity [1]. At present, we are working on the development of a global comprehension test which in its present form is called the Question and Response (QAR) test. The rationale and methods used in the development of this test are summarized below. For a more detailed account of the experiments leading to the present design of the QAR test see McAllister and Dufberg, [7] and Dufberg [2].

2. METHODS

An inventory of comprehension testing methods currently in use motivated an emphasis at the outset on construct and content validity in our early research in the development of the QAR test. The model presented in fig. 1 is the theoretical basis for the structure of the QAR test. This model, however, is very general and vague in terms of specific perceptual mechanisms. It was therefore necessary to consider recent research in comprehension testing in light of linguistic- phonetic theory. It was decided to begin with a method designed by Walker and Byrne [9] which could be said to be a version of a general test paradigm to assess speech reception threshold (SRT). Running speech was presented in noise and the test result was the signal to noise ratio at which the subject/listener "just barely understood the meaning of the text".

With the above experimental configuation (fig 2) we tested several important features of the original SRT paradigm.

2.1 Noise type

Several types of noise sources were used in these experiments. Our pilot studies narrowed these various maskers down to two main candidtates for the QAR test. One was a colored, low frequency modulated noise whose long time average spectrum was approximately the same as a male voice. The other was a "cocktail party noise" achieved by overdubbing one male and one female voice many times to create the effect of a roomful of people engaged in lively conversation. The results of these experiments showed that the "babbel", as we called the cocktail party noise, was

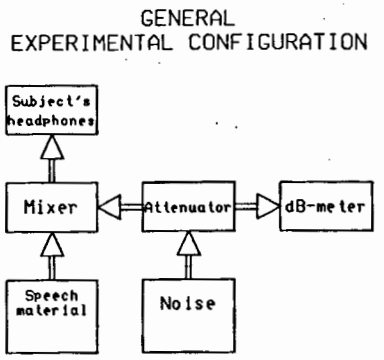


Figure 2.

the most effective masker and therefore we chose this noise for the QAR test.

2.2 Speech material

From a point of view of validity, we judged the original method and therefore the speech material in the SRT test to be inappropriate. Instead of a subjective judgement as to the noise level at which the connected speech was "just barely understood" we chose to ask simple questions whose answers would be clear to anyone who knew the language and had understood the question. It was assumed that this would be a more valid method for determining whether or not the listener/subject had understood the running speech. A further development of this idea resulted in the material for the QAR test. This material consisted of a series of short texts averaging 31 words (5 sentences) per text and constructed systematically according to text linguistic principles of script theory [8].

2.4 Subjects

Three subject groups were used in the experiments which led to the QAR test. These were: Foreigners who used Swedish, the language of the tests, as a second language, hearing impaired persons, and normal native speakers of Swedish as controls.

2.5 The QAR test and test comparisons

The methods summarized above were used in the development of the QAR test. This test, then, used "cocktail party" noise mixed with short texts. Subjects were asked to answer questions, also presented in noise, about the text and the S/N ratio was adaptively adjusted according to the subjects performance in the answering task. The test result is a S/N ration which represents a threshold of 50 per cent speech comprehension.

2.5.1 Tests for comparison

At the time of the writing this paper, the QAR test is being given to new subject groups together with the following tests for comparison purposes.

Bèkèsy pure tone threshold audiometry. This is a standard hearing test used in clinics in Sweden and elsewhere.

Hearing Threshold for Speech. Also a standard audiometric test in swedish clinics. Part of the test repertoire called "Speech Audiometry", this test is supposed to indicate speech comprehension ability.

The Hagerman Test. Also part of the "Speech Audiometry", test repertoire supposed to indicate speech comprehension ability [4].

Sentence Completion Test and *Word Completion Test.* Two paper and pencil tests in which subjects, within limited time frames, had to use syntactic and lexical information to complete various tasks.

Modified Hearing Measurement Scale. A self assessment test based on the Swedish version of the Hearing Measurement Scale [3] in which subjects estimate their comprehension of speech in a number of hypothetical everyday situations.

3. RESULTS

The results presented here are based on only 5 normal and 3 immigrant subjects. At the time of the ICPhS we will have more data to present from these subject groups as well as from a hearing impaired subject group.

The results from the various tests that were run in these experiments were of differt types and could not be directly compared by means of parametric statistics. The subjects were rank ordered for each of the seven tests mentioned above i.e. resulting in seven rank orders. A Spearman rank order correlation was performed between the QAR test and all the other tests given to the subjects. The only test with which the QAR seemed to be correlated was the self assessment test here called the modified Hearing Measurement Scale. This correlation could only be said to be modest with r_s=.52. All the other test showed a very low positive or negative correlation.

4. DISCUSSION

It was not entirely unexpected that the QAR test would not show a high correlation to the other tests used in these experiments. It has been our contention, on the basis of validity arguments, that it could be quite possible that these tests do not, in fact, test speech comprehension as we have defined it here. We would suspect that important components in speech comprehension are tested by them, but that the global aspects of this ability may not be captured. The moderate correlation of the QAR test with the self assessment test could be seen as encouraging if it holds. We would, at present, be very cautious in lending a great deal of credibility to these preliminary results. At the time of the ICPhS we will be able to present similar data based on larger groups of subjects and hopefully be able to make more definitive statements about the utility of the QAR test.

5. REFERENCES

[1] CRONBACH, L.J., (1961) *Essentials of Psychological Testing*, New York, Harper & Row.

[2] DUFBERG, M. (1990), "The measurement of speech comprehension: a progress report of the construction of a global test", *PHONUM* 1, Dept of Phonetics, Univ of Umeå.

[3] ERIKSSON-MANGOLD, M. and HALLBERG, L. (1990), "Hörselskada och upplevt handikapp. Hearing Measurement Scale på svenska" (the Hearing Measurement Scale in Swedish), Report from the Dept of Psychology, Gothenberg University, nr. 3.

[4] HAGERMAN, B. (1984), *Some aspects of methodology in speech audiometry*, (dissertation) Karolinska inst., Stockholm.

[5] LINDBLOM, B. (1987), "Adaptive Variability and Absolute Constancy in Speech Signals: Two Themes in the Quest for Phonetic Invaiance", *Proceedings of the XI International Congress of Phonetic Sciences*, Tallisn, August 1987.

[6] McALLISTER, R. and DUFBERG, M. (1989), "Some Attempts to Measure Speech Comprehension", *PERILUS IX*, Institute of Linguistics, University of Stockholm.

[7] MIDDELWEERD, M. J. and PLOMP, R. (1987), "The effect of speech reading on the speech reception threshold of sentences in noise" *JASA* 82, nr. 6, 2145-2147.

[8] SAMUELSSON, S. and RÖNNBERG, J. (1990), "Script activation in lipreading" *Scandinavian Journal of Psychology* 13.

[9] WALKER, G. and BYRNE, D. (1985), "Reliability of Speech Intelligibility Estimation for Measureing Speech Reception Threshold in Quiet and Noise", *Australian j Audiology* 7:1.

A NEW DICTIONARY OF ENGLISH PRONUNCIATION

John C. Wells

Department of Phonetics and Linguistics, University College London

ABSTRACT. The author's newly compiled Longman Pronunciation Dictionary [5] is not restricted to British RP but also covers American English. Like EPD [2], it gives extensive coverage to inflected and derived forms; unlike EPD, it has entries for affixes and compounds, and offers spelling-to-sound guidelines. Entries incorporate new treatments of epenthesis, syllabic consonants, and "compression" (varisyllabicity).

For nearly a hundred words where competing pronunciations are known to be in use, LPD reports the findings of an opinion poll of speaker preferences - the largest such poll ever conducted.

1. INTRODUCTION

The Longman Pronunciation Dictionary (LPD) [5] was published last year as the culmination of four years' work.

There are three principal things, missing from ordinary dictionaries, that a pronouncing dictionary can offer : information on variants, on inflected and derived forms, and on proper names. All these are in LPD, along with guidance on English spelling-to-sound rules, on the pronunciation of combining forms and affixes (and the effect they have on word stress) and on English phonetics in general.

There was of course already in existence an excellent pronouncing dictionary for English: the classic EPD [2] compiled by Daniel Jones over seventy years ago and more recently revised, first by Gimson and now by Ramsaran. The aim of LPD was to improve upon it.

1.1 Variant pronunciations

Many English words are pronounced in more than one way. As well as the recommended or most usual pronunciation of a word (not necessarily the same thing!), LPD records also the variant pronunciations in common use: not only those considered to fall within RP, but also a limited range of variants from non-RP British English (regional forms restricted to particular parts of the British Isles). Thus as well as again with both /e/ and /eɪ/, LPD also shows one with northern /ɒ/ alongside the usual /ʌ/, and solve with the southern /əʊ/ alongside /ɒ/. Also, more importantly, it gives the General American forms, so catering for those EFL learners who taken AmE as their model. The entry for tomato reads

tə 'mɑ:t əʊ ‖ tə 'meɪt oʊ

--where ‖ introduces an American pronunciation, here with the characteristic American voiced t, judged so salient an allophone as to demand explicit notation. There are also occasional references to Scottish (though), West Indian (Bridgetown), Irish (name of the letter H) and other varieties from outside England.

Extensive listing of variants creates the danger of making the dictionary difficult for the ordinary EFL student, who just wants advice on the pronunciation he should use. LPD helps this kind of user by putting the recommended form in colour (blue), and the other possibilities in less conspicuous black. If the recommended British and American forms are different from one another, then both are printed in colour.

1.2 Inflected and derived forms

These are not always readily inferred by a dictionary user, even if they are regularly formed. He may know that breathe is bri:ð, yet hesitate about breathed and breathes. It is useful to be able to check them explicitly.

1.3 Proper names

Ordinary dictionaries contain few proper names. Yet the spelling is a notoriously unreliable guide to their pronunciation. So LPD offers good coverage of names of

- people (forenames - Angharad, Graham, Ralph, and surnames - Gyngell, Marjoribanks, McElhone, Wheway; literary characters - Gradgrind, Lear, Peter Pan; gods, heroes, and figures of myth and legend - Thor, Hephaestus, Robin Hood);
- places (not only in Britain, as Gloucester and Chiswick, and the less well-known Lympne, Stivichall, and Meols; but also in Ireland (Laois, Drogheda), North America (Poughkeepsie, Spokane), South Africa (Uitenhage) and Australia (Whyalla), as well as many hundreds in non-English-speaking countries); and
- commercial firms and products, such as the Weetabix we eat for breakfast or Exxon whose petrol we buy; and on from Aspro tablets, Armalite rifles, and Araldite glue, through Gauloises, Givenchy and Gucci, Pepsi-Cola, Pernod and Perrier to Winalot dog food and Y-fronts underwear.
- pop music -- not only the Beatles, Lennon and McCartney (all missing from EPD!) but also Bananarama, Sade, and Yazz.

Proper names, too, have inflected forms. A Porsche car, disyllabic in German, is monosyllabic in English, but has a disyllabic plural. A Mercedes, on the other hand, often has a plural identical with the singular, like series.

All told, LPD contains about 75 000 headwords, of which about 15 000, one-fifth, are proper names.

1.4 Foreign languages

Where a word or name is from a foreign language, LPD usually records the pronunciation in the relevant language, too, as well as the anglicization. In this way it includes over a thousand items with their French pronunciation in IPA (au fait, hors d'oeuvres, Sartre); likewise >400 German entries (Berlin, Fräulein, Munich-München), c300 each Italian and Spanish, >240 Welsh, >130 Russian (Chernobyl, glasnost, Gorbachev), >70 Hindi, >60 Japanese (Kyoto, sumo), >50 Dutch, >40 Arabic, and in fact items from a total of 51 languages other than English.

1.5 Speech technology applications

LPD was compiled in machine-readable form and can thus in principle be made available as an electronic database. In speech recognition a pronouncing dictionary look-up can be used in order to match an incoming signal against possible lexical strings; in speech synthesis its usefulness is obvious, given the uncertainty of English spelling-to-sound rules. No comparable machine-readable database exists. (Enquiries should be addressed to the publishers, Longman.)

2. NOTATION

The transcription system is essentially the "EPD-14" IPA notation employed in the current, fourteenth, edition of EPD and by many other writers, particularly in the EFL sphere. This differs from older notations in that the distinction between paired long and short vowels is symbolized both by separate letters and by presence/absence of length marks: leap li:p, lip lɪp; food fu:d, good

gʊd; caught (BrE) kɔ:t, cot (BrE) kɒt. Nevertheless, both theoretical and practical considerations have forced certain minor modifications to EPD-14, as follows.

2.1 Neutralization of high vowels

The weak final vowel in words like happy, coffee, valley is in RP traditionally equated with /ɪ/. Many speakers, however, identify it rather with /i:/. Phonologically we have a neutralization of the phonemic opposition. I have followed others [3, 4] in employing the symbol i (lower case, no length marks) in positions of neutralization, and also u for the corresponding /u:-ʊ/ neutralization: happy 'hæp i, radiate 'reɪd i eɪt; evaluate ɪ 'væl ju eɪt.

2.2 Provision for General American

Since LPD also covers AmE, the EPD-14 system needs certain extensions, chosen in such a way as to harmonize with the notation used for BrE RP while drawing attention to salient phonetic differences. A regrettable consequence is that LPD does not follow any established AmE notation such as Kenyon & Knott, or Pike, or Smith & Trager.

lot lɒt ‖ lɑ:t;
thought θɔ:t ‖ θɒ:t, θɑ:t;
know nəʊ ‖ noʊ;
nurse nɜ:s ‖ nɝ:s;
farmer 'fɑ:m ə ‖ 'fɑ:rm ᵊr;
atom 'æt əm ‖ 'æt əm.

The explicit symbolization of voiced /t/ is helpful to the AmE-oriented EFL learner: few other dictionaries give this information.

2.3 Resolution in prospect

The problem of rival phonetic transcription systems will soon be solved by software. Users of an electronic database will be offered output in any notational system they choose, through an automatic lookup table.

3. UNDERLYING vs. SURFACE FORM

A general problem facing any phonetic transcriber is that of abstractness. How far should the analysis and the corresponding notation attempt to shadow the phonetic details of the utterance, rather than abstracting from this into phonological entities believed to underlie them? The moderate degree of abstractness implied in taxonomic-phonemic analysis is generally acceptable (though even here there may be murmurs about dark l's, glottal stops, and "long" vowels that are physically short). An extreme abstractness such as found in SPE [1] is obviously quite inappropriate for the needs of most potential users of a pronouncing dictionary.

A number of related issues call for discussion. Here are some of them, with the solutions adopted in LPD.

3.1 Assimilation

Include is usually pronounced with a nasal which for some speakers is perhaps always velar, for others occasionally alveolar. Some, but not all, have a psychological awareness of the morphology (compare exclude). Given an analysis that treats /ŋ/ as phonemic, with a corresponding transcription, should the main entry have /n/ or /ŋ/? LPD gives /n/, with an /ŋ/ variant marked "→" ('derived [from first form] by automatic rule'). So too spaceship with /'speɪs-, →'speɪʃ-/.

3.2 Epenthesis

In LPD the established convention of an italicized symbol to show possible omission is applied to capture AmE nt-reduction (painting 'peɪnt ɪŋ), recorded nowhere else. The further convention of using a small raised symbol to show a possible insertion caters nicely for intrusive /r/ (BrE thawing 'θɔ:ʳ ɪŋ) and for other types of epenthesis (fence fenᵗs, spiral 'spaɪᵊr əl).

3.3 Syllabification

Transcribed forms in LPD are divided into syllables by spacing, which is intended to make them easier for the user to process mentally. Syllabification is based on the principle of attraction to stressed syllables, which (I claim) affords a more elegant statement of allophonic distribution than do rival theories.

3.4 Syllabic consonants

Arguably, all syllabic consonants in English are in alternation with a sequence of [ə] plus the corresponding non-syllabic consonant. LPD adopts this assumption, showing syllabic consonants by with either an italic [ə] or a raised [ᵊ] (using the conventions previously discussed). So sudden is written 'sʌd ᵊn, implying a preference for syllabic [n̩], but distant as 'dɪstənt, implying a preference for [ən]. AmE [ɚ] is analysed as syllabic r and written accordingly: AmE manner 'mæn ᵊr.

3.5 Compression and smoothing

The number of syllables in an English word may be variable: we can often compress two syllables into one. Thus listening, for instance, may have three syllables or two; likewise lenient. Choice of variant may depend on stylistic or pragmatic factors. This varisyllabicity is shown by a special convention in LPD, thus 'lɪs ᵊn‿ɪŋ, 'li:n i‿ənt.

The phonological environment for compression is typically a sequence of two weak-vowelled syllables, the first of which loses its syllabicity. The rule is, however, subject to lexical constraints (contrast battery and beggary).

Combining these treatments of potential syllabic consonants and compression, we achieve notably succinct entries for such words as national 'næʃ ᵊn‿əl and liberal 'lɪb ᵊr‿əl: each of these conflates what Jones would treat as six distinct variants.

By **smoothing** I refer to the RP tendency for a stressed vowel to be simplified when immediately followed by a weak vowel. A diphthong in this environment, if smoothed, loses its second element; a high vowel becomes lax. Thus in client the /aɪ/ may be smoothed to [a] before /ə/. In ruinous the /u:/ may become [ʊ]. However, these sequences are also subject to the possibility of compression. LPD's entries read 'klaɪ‿ənt, 'ru:‿ɪn əs. (An actual triphthong in client, i.e. compression without smoothing, is in my opinion rare.)

4. THE OPINION POLL

Reflecting about the problem of authority, I resolved to conduct an opinion poll for nearly a hundred words where competing pronunciations are known to be in use. LPD reports the preferences expressed by panel of 275 native speakers of British English.

The panel consisted mostly of academic phonetics/linguistics specialists, school or college teachers, radio announcers, and speech scientists.

The survey revealed that in zebra /e/ is heavily preferred over /i:/, in accomplish /ʌ/ over /ɒ/, in deity /eɪ/ over /i:/, and in year /ɪə/ over /ɜ:/. In nephew, /f/ has largely displaced /v/; suit is now usually said without /j/. Three out of four respondents prefer often with no /t/.

REFERENCES

[1] CHOMSKY, N. AND HALLE, M. 1968. The Sound Pattern of English. MIT Press. (= SPE).

[2] JONES, D. (1917), Everyman's English Pronouncing Dictionary. Twelfth edition, 1963. Fourteenth edition, edited by A.C.Gimson, 1977. Reprinted with revisions and Supplement by Susan Ramsaran, 1988. London: Dent. (= EPD)

[3] LDOCE = Longman Dictionary of Contemporary English, 1978. Second edition, 1987. Longman.

[4] ROACH, P. 1985. English Phonetics and Phonology. Cambridge University Press.

[5] WELLS, J.C. (1990), Longman Pronunciation Dictionary. Harlow: Longman. (= LPD)

RHYTHM AND THE ALGERIAN SPEAKER OF ENGLISH

Nadia BENRABAH-DJENNANE

Institut des Langues
Université d'Oran, Algérie

ABSTRACT

The paper concentrates on the rhythm adopted by Algerian speakers of English (ASE). The speech sample recorded revealed a number of features which could be held responsible for giving the Algerians' speech a typical staccato unEnglish type of rhythm.

The major rhythmic errors seemed to result from the total absence of weak forms, insertion of glottal stops before initial vowels, lengthening of unstressed syllables, and inadequate stressing.

1. INTRODUCTION

The notion of 'stress-timed' rhythm and/or pure isochrony in English has long been debated [8;10;3; 4]. Experimental evidence has shown that what the human ear perceives as regular beats is far from representing real physical events. In fact, it is the hearer's mind which imposes a regularity which often does not exist [1; 4; 2].

However, for the sake of clarity when describing a given language, one may find it useful to say that in languages like English stress tends to come at more or less equal intervals in time. This results in any succession of unstressed syllables to be 'crushed' or 'compressed' together so as to say them more rapidly. For that purpose, English is well known for its use of vowel elison, reduction or weakening as well as its preference for weak rather than strong forms in most grammatical unstressed words [5; 9]. The combination of those features of connected speech gives the English language its characteristic 'stress-timed' rhythm.

The present paper reports on Algerians' handling of such features. It will be shown how improper use of them by the ASE results in giving their speech this typical 'syllable-timed' rhythm.

2. PROCEDURE

The experiment was originally intended to investigate the intonational proficiency of a sample of twenty ASE. It consisted of ten units altogether which required the informants to read on the one hand (sentences, dialogues and a short narrative) and to speak more or less freely on the other (picture description, guided and free speech).

Note that although this experiment was not specifically designed to study solely the rhythm of ASE, the recorded data provided a substantial amount of information about Algerians' performance at this level. A number of interesting observations were made. Some of them are briefly discussed in the following sections.

3. CHARACTERISTICS OF ALGERIAN ENGLISH RHYTHM

3.1. Weak Forms

The first thing that strikes the observer's ear in the sample of Algerian English is this staccato rhythm more like a 'syllable-timed' language. Every syllable is given the same value. This impression is reinforced by the total absence of weak forms. Nearly all function words are used in their strong forms. As a result, those words are consistently given undue importance. Below are some examples (note: F and M stand for 'female' and 'male' speakers, RP for 'Received Pronunciation'):

(1) What do you think he can do with this computer?

F2: [wat du: ju: θi:ŋk hi: kæn du: wi:d dis kɒmpju:tæ]

as opposed to

RP: /'wɒt djʊ°θɪŋk ɪ kən 'du: wɪð ðɪs kəm°pju:tə/

(2) Then where do you think he is working ?

M10: [ðen wɛ: du: ju: θi:ŋk hi: ʔi:z wɜ:ki:ŋg]

RP: /ðən'weə djʊ°θɪŋk ɪz 'wɜ:kɪŋ/

(3) I am happy

F19: [ʔaɪ ʔæm hæpi:]

RP: /aɪm 'hæpɪ/

(4) What is his job ?

F8: [watʔi:z hi:z dʒɒb]

RP: /'wɒts ɪz 'dʒɒb/

(5) What kind of books does he read ?

M14: [wat kaɪnd ʔɒf bu:ks dɒz hi: ri:d]

RP: /'wɒk °kaɪnd əv 'bʊks dəz ɪ °ri:d/

3.2. Insertion of Glottal Stops

Also peculiar in the Algerian speech sample studied is this tendency to insert a glottal stop before vowels in syllable initial position. This is particularly striking when the word in question is a grammatical item such as 'is', 'am, 'of' as in examples (2) to (5) above. In connected speech those words are never preceded by a glottal stop in RP unless when stressed for particular emphasis. Instead, they are linked together with some kind of liaison as in [aiʲæm] [7]. In order to avoid such insertion of glottal stops, syllable initial vowels tend to be linked to the preceding final consonant as in:

/ðɪs ɪz ə naɪs ɔ:gəst a:ftənu:n/

Glottal stop insertion in Algerian English breaks the

possibility of smooth transition between the words. It could also be held responsible for this staccato type of rhythm.

3.3. Lengthening of Unstressed Syllables

In addition, the long vowel in words like [ʔi:z] for 'is' makes this item sound more like a content word, e.g. 'ease'. Sometimes this can even lead to communication breakdown as shown in the following utterance as spoken by an Algerian informant:

[ðø mɛn ʔi:z ði: onli: li:ŋk bitwi:n ðɛm]

which could mean both

a) The man is the only link between them

or

b) The men ease the only link between them

although the speaker intended meaning a).

Another interesting observation was made. If R.P. makes use of two distinct qualities for the following short versus long vowels:

1) /i: / versus / ɪ /
2) /u:/ versus / ʊ /
3) /ɔ:/ versus / ɒ /
4) /ɜ:/ versus / ə /

it is quite noticeable that similar contrasts (particularly the first two pairs) are absent in Algerian English. The tendency is to use a vowel of the type [ɩ] for the first pair and [ɯ] for the second. Both are usually (though not consistently) long. Thus, in function words like 'you', 'do', etc... [u:] is the only vowel used. The short variety, [ʊ], is never used. Similarly, in function words like 'is', 'his', 'he', 'she', etc... the Algerian preference is for a vowel sharing more resemblance with R.P. [i:] (closer and more front) than [ɪ] which is practically absent, e.g.

[hi: went tu: pæris]

instead of

/hɪ'went tə'pæɪs/

The use of [u:] and [i:] in such grammatical items tends to lengthen those words. The immediate result is that function words in Algerian English are less easily distinguishable from content and stressed words where full/long vowels are also used. Instead, they sound very much similar. It is as if Algerian English consisted of a series of content words with full vowels, hence giving the speech a staccato 'machine-gun' type of rhythm. Even polysyllabic words where R.P. usually stresses one syllable keeping the remaining one(s) weak, e.g. 'comfortable'/'kʌmftəbl/ 'literature'/'lɪtrətʃə/ in Algerian English each syllable is given the same value. However, in order to match the stressed syllables, the so-called short vowels are held longer even if not stressed.
This lengthening phenomenon occured not only in function words, but also in content words. In the next two sentences, the underlined syllables were held longer than should have been by a significant number of ASE.:

(6) 'Bill hasn't washed it, has he?'

(7) 'She brought apples, bananas, oranges, strawberries...'

It has been claimed [6] that utterance final syllables tend to be lengthened. Therefore, one could perhaps argue that the lengthening of the utterance final syllable of, say, 'strawberries' was in fact predictable. However, this view could not hold as it still will not account for the lengthening of other word final unstressed syllables which occur medially in the utterance. This is the case of '-sn't', '-shed', '-pples', '-ba-', '-nas', '-ges', 'be-' '-rries' in (6) and (7) respectively.

3.4. Inadequate Stressing

In conjunction with the above characteristics the data revealed another peculiar aspect of Algerian English rhythm: inadequate stressing. The spoken sample clearly showed that function words were not only used in their strong forms, but they were also very frequently stressed. This is the case of 'do', 'you', 'he', 'can' in examples (1) and (4) above. It may give the impression that those words are emphasized. But it is quite uncommon in English to put so much emphasis on such words with similar regularity.

4. CONCLUSION

So what seems to come out of this pilot study is that the rhythm of ASE presents a number of peculiarities. These seem to work together and contribute in their own way to give the informants' speech this particular Algerian touch.

5. REFERENCES

[1] ALVAREZ DE RUF, H. (1978) *A Comparative Study of the Rhythm of English and Spanish*, M.Phil., University of Leeds.

[2] BENRABAH-DJENNANE, N. (forthcoming), "Rhythm Through the Ages: Towards a Tentative Definition", *Revue des Langues*, I.L.E., Université d'Oran.

[3] CRYSTAL, D. (1969) *Prosodic Systems and Intonation in English*, CUP.

[4] GARNER, D. (1985), "Errors of Timing in Advanced French Speakers of English: a Study of Vowel Duration and Voice Onset Time", *Phon. Lab. Univ. Reading: Work in Progress* 5: 65-82.

[5] GIMSON, A.C. (1980) *An Introduction to the Pronunciation of English*, (3rd ed.), London: E.Arnold.

[6] McNaught, J. (1978), *The Prosodic Competence of a Sample of French Speakers of English*, unpublished MA diss., University of Manchester.

[7] MORTIMER, C. (1985), *Elements of Pronunciation: Intense Practice for Intermediate and more Advanced Students*, CUP.

[8] PIKE, K.L. (1945) *The Intonation of American English*, Michigan.

[9] ROACH, P. (1983) *English Phonetics and Phonology: a Practical Course*, CUP.

[10] SHEN, Y. & PETERSON, G. (1962) "Isochronism in English", *Studies in Linguistics. Occasional Papers* 9: 7-35 (Buffalo University Press).

LA VIE SOCIALE DES SONS, MODELE DIDACTIQUE DE LA PRONONCIATION DU FRANCAIS

F. Wioland

Institut de Phonétique, Strasbourg, France.

ABSTRACT

Habits of pronunciation in French are explained by the rules which govern the society of sounds in the frame of the rythmic group, rules which underlie the hierarchical and individual relations between sounds.

1. INTRODUCTION

Pour l'apprentissage de la prononciation, la réflexion de tout apprenant lettré passe par la représentation écrite des mots. Il s'avère donc nécessaire parce que conforme à la démarche habituelle des apprenants d'utiliser la forme écrite des énoncés oraux en didactique de la prononciation. Mais comme la graphie ne rend compte qu'en partie de la prononciation du français - elle en ignore même des caractéristiques fondamentales -l'écrit en tant que tel ne permet pas d'expliquer les habitudes de prononciation;l'enseignement de la prononciation du français a toujours été l'objet d'un compromis à cet égard.

Pour permettre à l'apprenant comme à l'enseignant de situer n'importe quel problème de prononciation dans un cadre général et de ne pas avoir l'impression de résoudre à chaque fois un problème particulier, nous proposons un modèle qui rende compte de ce qui est important pour " l'oreille " d'un francophone alors que lui-même n'en est pas conscient, en présentant les sons du français comme des individus vivant dans une société dont les règles de vie sont comme dans toute société, arbitraires certes, mais générales, donc peu nombreuses, hiérarchisées et systématiques.

La connaissance des règles de vie de cette société des sons permet d'expliquer l'ensemble des habitudes de prononciation au-delà des diversités individuelles. La prise de conscience par l'apprenant du fonctionnement de l'oral sous forme d'un modèle est une étape essentielle sur la voie de la pratique de l'oral.

2. LE MODELE DIDACTIQUE

Le modèle présenté prend successivement en compte

- le cadre social qui régit la vie des sons,
- les lois qui découlent de la position sociale des sons,
- les relations inter-individuelles des sons en contact,

c'est-à-dire un ensemble autonome dans le cadre duquel la graphie a un rôle à jouer [4] dans la mesure où les positions des sons à l'oral sont définies.

2.1. Le cadre social

Les unités rythmiques qui forment un ensemble homogène de production comme de perception ne correspondent pas au mot du dictionnaire mais aux unités significatives habituelles du discours. Une didactique qui prend en compte le rythme doit tout d'abord permettre à l'apprenant d'identifier ces unités qui sont le cadre social qui détermine l'ensemble des habitudes de prononciation du français [7].

L'identification du nombre de syllabes prononcées par unité est fonction des deux principes de la production rythmique [6], celui d'économie rythmique - un petit nombre de syllabes par unité - et celui de l'équilibre des rapports temporels qu'entretiennent les unités rythmiques successives.

L'élément moteur de chaque unité rythmique qu'est la dernière syllabe prononcée peut être dépouillé de toute ambiguïté graphique grâce à la transcription: il est important de montrer la syllabe la plus importante pour la compréhension dans un contexte signifiant en raison de l'originalité rythmique du français et des paradoxes que présente l'accent rythmique de par sa position finale et la nature de ses paramètres articulatoire et acoustique [2], [1].

Le schéma prosodique proposé favorise le glissando final qui accompagne l'accent rythmique au détriment de toute autre mise en relief étant donné que la hiérarchie syntaxique et la fonction grammaticale sont fidèlement codées dans la structure temporelle à travers les différents degrés d'allongement de la durée des voyelles en position finale d'unité [5].

" Les apprenants doivent très vite comprendre qu'ils ne seront jamais jugés sur un ou deux faits langagiers précis (sauf exception) mais sur une impression d'ensemble " [3].

2.2. Les lois de position sociale

Chaque son occupe, dans le cadre de l'unité rythmique, une position bien déterminée; il se trouve ainsi soumis aux lois qui découlent de la position sociale qu'il occupe. Cette position est soit favorable, soit défavorable.

Le nombre extrêmement limité de positions est d'un grand intérêt didactique: il suffit, en effet, de reconnaître telle position pour appliquer telle habitude de prononciation.

C'est ainsi qu'une syllabe ne peut occuper que trois positions:

- une position accentuée, la plus importante, sur laquelle doit porter l'essentiel de l'apprentissage,
- une position accentuable qui malgré la brièveté de sa réalisation conserve un timbre vocalique comparable à celui de la position accentuée,
- une position inaccentuable dont la structure syllabique majoritairement ouverte ne nécessite pas de différenciation des timbres ouverts ou fermés pour respectivement / E /, / O / et / OE /.

D'autre part les structures ouverte ou fermée de la syllabe déterminent timbre et durée vocaliques.

Enfin pour les consonnes la position finale de syllabe,

quelle que soit la position de cette syllabe, est une position faible qui entraîne un relâchement systématique de son articulation.
Il est de première importance de ne pas calquer en français langue étrangère la prononciation sur la seule forme écrite des mots, mais de lui accorder son indépendance. Les sons du français vivent en société; tout apprenant doit être informé des lois peu nombreuses de cette société et comprendre qu'à une même position sociale correspond une même habitude de prononciation.

2.3. **Les sons en contact**

Les positions respectives des sons étant clairement définies en vertu des lois de position, chaque son se trouve soumis aux lois qui découlent de sa position dans le groupe. Il est des positions privilégiées qui favorisent l'épanouissement du son considéré - son dominant - d'autres au contraire qui l'obligent à subir de multiples influences du fait de sa position défavorable dans le groupe - son dominé. En plus des positions fortes et des positions faibles opèrent également des affinités ou répulsions entre les sons en contact.
La visualisation des sons en contact et de la structuration syllabique de l'unité rythmique est obtenue par une transcription adéquate [8].
Une Consonne au contact d'une Voyelle permet d'expliciter à l'oral la syllabation de base qui entraîne l'élision, l'enchaînement et la liaison.
Une Voyelle au contact d'une Consonne permet d'expliciter à l'oral la limite syllabique et les réalisations nasales ou non des graphies Voyelle + N ou M.
Une Voyelle au contact d'une Voyelle permet d'expliciter à l'oral les graphies vocaliques doubles, les voyelles successives à l'oral soit dans un mot soit séparées à l'écrit par un blanc ou une consonne muette dont le H disjonctif, les graphies I, Y, OU , U suivies d'une voyelle prononcée.
Une Consonne au contact d'une Consonne permet d'expliciter à l'oral les deux types d'assimilation, soit progressive lorsque les consonnes forment groupe, soit régressive lorsque les consonnes ne forment pas groupe, ainsi que l'apparition de nouveaux groupes de consonnes du fait en particulier de la chute de la graphie E.

3. CONCLUSION

Les habitudes de prononciation sont donc le fait d'une société de sons très bien organisée où les relations pourraient paraître comme régies par de simples rapports de force en fonction des positions sociales: en effet certains s'affirment grâce à leur position favorable, d'autres ne peuvent se maintenir qu'au prix d'une évolution qui peut surprendre, la disparition de certains donne naissance à de nouvelles relations inconnues jusque là.
Mais réduire la vie des sons à des lois strictes de positions sociales serait méconnaître l'influence originale dégagée par la personnalité de chaque son. Au contact les uns des autres dans des situations sociales comparables ils

réagissent comme tout individu en fonction de leur nature propre, ce qui est la marque d'une société vivante, gage d'évolution et de créativité.
A partir d'une prise de conscience des lois sociales forcément générales qui régissent les habitudes de prononciation, l'apprenant dispose d'un modèle de prononciation du français qui lui permet d'aller à l'essentiel grâce à quelques références toujours identiques qui, à la différence de la graphie, sont d'une étonnante régularité. La maîtrise souhaitée à l'oral est facilitée par une représentation non ambiguë des contraintes essentielles.

4. REFERENCES

[1] BENGUEREL, A. (1970), " *Some physiological aspects of stress in french* ", Ann arbor, University of Michigan natural language studies, 4.
[2] DELATTRE, P. (1966), " *Studies in French and Comparative Phonetics* ", Paris : Mouton.
[3] FILLIOLET, J. (1985), " Réflexions sur l'apprentissage en français langue étrangère des variations stylistiques du français oral spontané ", *Etudes de Linguistique Appliquée, 59*, 31 - 42.
[4] GAK, V.G. (1976), " *L'orthographe du français* ", Selaf.
[5] PARK, Y.M. (1989), " Aspects syntaxique et rythmique de l'organisation prosodique des phrases en français: étude acoustique des variables temporelles et mélodiques ", *Travaux de l'Institut de Phonétique de Strasbourg, 21*, 1 - 210.
[6] WENK, B.J. & WIOLAND, F. (1982), " Is French really syllable-timed ? ", *Journal of Phonetics, 10*, 2, 193 - 216.
[7] WIOLAND, F. (1983), " La rythmique du français parlé ", *Publication de l'Institut International d'Etudes Françaises, 7*, Université des Sciences Humaines, Strasbourg.
[8] WIOLAND, F. (1991), " *Prononcer les mots du français* ", Collection F, Autoformation, Hachette F.L.E.

FICHES CORRECTIVES DES SONS DU FRANÇAIS: DÉFENSE ET ILLUSTRATION DE LA CORRECTION PHONÉTIQUE PONCTUELLE

Jean-Guy LeBel

Département de langues et linguistique
Université Laval, Québec, Canada, G1K 7P4

The **Punctual Approach** meets *specific* learner needs based on level and L1 by processing errors *immediately* and *analytically* before returning them to the spoken utterance. The role of **Diagnostic Cards** (fiches correctives) is to describe the 36 phonemes of "maximal" French and the 150 errors affecting them, while proposing many corrective techniques for each, as outlined in the *punctual approach*.

1. APPROCHE PONCTUELLE

1.1. Définition

Dans une attitude d'esprit éclectique et pratique, la **méthode ponctuelle de correction phonétique** est un ensemble de 7 GRANDS MOYENS qui englobent, de manière cohérente et intereliée, les divers *procédés, façons, trucs, recettes...* que j'ai inventoriés, modifiés ou imaginés: soit en d'autres mots ce que, à ma connaissance, nous possédons actuellement en correction phonétique comme procédés de correction.[1]

La méthode ponctuelle et ses 7 GRANDS MOYENS est donc la synthèse de vingt-cinq ans de pratique de la classe de correction/travail phonétique et d'enseignement de sa méthodologie aux futurs maîtres, et elle est utilisée selon une stratégie qui sera succinctement exposée dans les lignes suivantes.

1.2. Attitudes fondamentales

Voici ce que propose la méthode ponctuelle de correction phonétique:

a) Elle accorde aux *exercices de discrimination auditive* une prépondérance de tous les instants pour tout phénomène, phonique et prosodique, et ce, de manière **ponctuelle** ou **systématisée.**

b) Elle s'adresse autant aux *débutants* qu'aux très avancés; toutefois, bien qu'un «grand» débutant ne le demeure vraiment que très peu de temps, l'insistance sur l'**atomisation**, sur l'**intellectualisation** et sur la **conscientisation** sera avec ce «débutant» inversement proportionnelle à sa capacité d'expression spontanée. Mais, à toutes fins utiles, la méthode ponctuelle vise tout public apprenant une L2 auquel elle s'adapte en nuances et en intensité.

c) Elle accorde aux phénomènes phoniques et prosodiques un traitement *immédiat* (à l'instant propice), *atomisé* (isolé et détaillé), *intense* (en durée et en répétitions) et, primordialement, *approprié* aux besoins spécifiques de chaque apprenant; cependant elle réintègre vite l'élément travaillé dans la combinatoire de la chaîne parlée, soit dans l'énoncé d'origine, soit dans des exercices appropriés [voir g) ci-après et les **Exercices de conditionnement phonétique** à paraître].

d) Elle favorise la tendance généralisée à l'*intellectualisation* et à la *conscientisation* du processus d'apprentissage [2] de la part de l'apprenant devant tout phénomène phonique et prosodique qu'il a à maîtriser.[3]

e) Elle accorde autant de place à *l'imitation, c.-à-dire à la reproduction* des modèles facilitants qui sont proposés qu'à la *production* spontanée libre ou encadrée (voir **Exercices de conditionnement phonétique**, à paraître).

Elle puise, sans réserve et selon ses besoins, dans *toute discipline* et *toute* idéologie: c'est une problématique ouverte, réceptive, donc *éclectique* qui, néanmoins, n'ira pas pour autant présenter ses connaissances théoriques directement aux apprenants.

f) Elle *privilégie*, d'après tout ce qui précède, une *période spécifique* consacrée à la correction et au conditionnement phonétiques, en plus des interventions sporadiques ou régulières du titulaire de classe quand il s'agit de débutants.

g) Conséquemment, la **méthode ponctuelle** prône de *nombreux* et *longs moments privilégiés* de correction et de conditionnement phonétiques allant de 15% du temps de classe pour un cours extensif (du soir, par exemple) jusqu'à 25% du temps consacré aux cours d'une session intensive (six semaines à raison de 20 h/sem., par exemple); et le travail phonétique personnel de l'apprenant s'ajoute à tout cela.

Ce n'est qu'à ce rythme de travail, sommes-nous plusieurs à le vivre, qu'on atteint agréablement et rapidement la fameuse **zone d'irréversibilité** et, son corollaire, l'**autocorrection active**.[4] Je crois qu'il faut que le correcteur et le corrigé en arrivent rapidement et intensément à vivre ensemble le moment où, en phonétique, *on n'en parle plus mais on parle*, et *correctement*!

1.3. Conditionnement phonétique

Dans cette ère *«communicative»*, et au cours des toutes récentes années, de très nombreux didacticiens de la phonétique tant en Europe qu'au Québec ont parlé du besoin de réformer ou de réinventer ce qu'il me plaît d'appeler la *pédagogie du parler*. Aussi existe-t-il à l'Université Laval (Québec, Canada) un projet qui est la mise en oeuvre d'un nouveau concept, du *processus de conditionnement phonétique* [5] dont la démarche tente d'être incitative, valorisante et efficace. La centration sur l'apprenant oblige évidemment à penser des exercices tout à la fois utiles et agréables à cet apprenant. Nous pensons que, du plaisir stimulant qu'aura ce dernier à les utiliser, découlera une intensité d'implication impressionnante et hautement fructueuse dans son travail de perfectionnement phonétique.

Nous avons conçu le *conditionnement phonétique* par analogie au conditionnement physique. Centré comme il se doit sur l'apprenant, le conditionnement phonétique est un processus dont la dynamique est constamment axée sur la modification de la prononciation en vue d'une amélioration qui pourrait se libeller: *être en forme phonétique* pour accomplir ce qui doit l'être au bon moment.

Les principes du conditionnement phonétique sont les suivants: ***répétition, réchauffement et gradation.*** [5]

1.4. Workout phonétique

Considérant donc plus particulièrement le second principe du conditionnement phonétique, à savoir le **réchauffement**, nous [6] avons transposé dans les cours de phonétique pratique ce que nous faisons au gymnase. À l'origine, le *workout* est de la gymnastique sur musique et *«c'est une activité libre, dirigée. On y enseigne des routines composées d'une variété d'exercices, ayant un caractère éducatif. La performance n'est pas du tout le but visé. C'est avant tout une activité de conditionnement physique plaisante et empreinte de dynamisme.»*

En phonétique, bien que le corps dans sa globalité puisse être sollicité et qu'il le soit de fait, ce sont lesorganes de la parole (avec la tête et les mains) qui sont le point de mire de notre travail. Pour arriver aux routines de *workout* phonétique, nous avons décortiqué le trapèze vocalique ainsi que l'ensemble des consonnes pour aller y chercher les éléments favorables à une meilleure production sonore. Tout est possible et réalisable avec un peu d'imagination et de créativité.

Le but visé par le *workout* est, évidemment, une meilleure prononciation, mais sans stress, sans obligation et sans effort mental de la part de l'apprenant. *Nous ne lui demandons pas de réussir, mais d'essayer.* Il n'a pas à réfléchir sur ce qu'il fait : il se laisse aller. Chacun va à son propre rythme. De plus, la performance n'a pas sa place ici. Elle l'aura petit à petit dans la période ultérieure dite d'exercices qui permet à l'apprenant de *trouver un certain équilibre articulatoire des sons travaillés.*

La **salle de workout** est là où on se trouve! En *classe,* le professeur utilise tout ce qui favorise l'esprit de jeu, de plaisir et d'appartenance au groupe. Le professeur participe aux exercices au même titre que les apprenants, il est leur guide et leur supporteur: il encourage, il exagère, il fait rire. Par ailleurs le *laboratoire de langues* se prête aussi très bien aux séances de workout phonétique, puisque les apprenants s'y sentent particulièrement à l'aise à cause de l'isolement dont ils y jouissent.

Tout le monde peut faire du *workout* phonétique, peu importe son âge et son niveau de connaissance du français. Il revient alors au professeur de composer des routines adaptées à la L1 des apprenants et à leurs difficultés de prononciation, bien que soit laissée une très grande liberté de choisir les musiques et les chansons selon le rythme à la mode, la beauté ou l'intérêt des paroles, etc.

2. FICHES CORRECTIVES

2.1. Description

Il y a tout d'abord les *Fiches descriptives* des 36 sons/phonèmes du français dit "maximal" qui comprennent les éléments suivants:

a) leurs caractéristiques **articulatoires et acoustiques;**
b) leur orthodiagramme, c.-à-d. un **schéma articulatoire idéalisé** de leur prononciation moyenne;
c) leurs **allophones standard** chez les francophones;
d) leurs **graphies courantes;**
e) leurs **fautes éventuelles** par des non-francophones;

Il y a ensuite les fiches correctives proprement dites d'environ 150 erreurs qui affectent les 36 sons/phonèmes du français et qui sont commises par des non-francophones de diverses origines linguistiques. Chaque fiche contient les éléments suivants:

a) la **transcription phonétique** de l'erreur;
b) la **clientèle** non francophone visée;
c) le **diagnostic commenté** de l'erreur;
d) la **correction suggérée;**
e) un **encadré** aide-mémoire qui résume les nombreux *trucs/recettes/moyens* de correction suggérés.

2.2. Utilité et but

Le **diagnostic commenté** de l'erreur est basé sur une phonétique descriptive et théorique qui constitue la 1ère partie de l'ouvrage et qui, aussi et bien sûr, fait de nombreuses références au cadre de travail ponctuel et systémique du **TRAITÉ...** . Le correcteur trouvera dans ce diagnostic les éléments pertinents à la compréhension tant de l'erreur de l'apprenant que de la correction suggérée.

De son côté, cette **correction suggérée** est basée sur l'optique *ponctuelle et systémique* dont le praticien trouvera le cadre théorico/pratique dans le **TRAITÉ** susmentionné.

En résumé, les **FICHES CORRECTIVES** [6] veulent aider le correcteur praticien dans son travail phonétique quotidien où il a à affronter de multiples et diverses erreurs qui exigent un traitement correctif immédiat et où un atomisme et un globalisme appropriés aux besoins spécifiques des apprenants se chevauchent dans un seul but: l'acquisition facilitée d'une prononciation agréable et compréhensible aux oreilles des autochtones francophones.

---

[1] LEBEL, Jean-Guy (1990), **Traité de correction phonétique ponctuelle: essai systémique d'application,** Québec, CIRAL (Université Laval, Québec, Canada), 275 p.

[2] LEBEL, Jean-Guy (1991), **Fiches correctives des sons du français,** Québec, Les Éditions de la Faculté des lettres, Université Laval (Québec, Canada), 385 p.

[3] Il semblait pourtant évident que les êtres humains ne régissaient pas tous leur apprentissage de la même façon, mais, comme pour toute intuition, il fallait des faits et preuves que nous ont heureusement fournis les récents travaux des FELDMAN, de la GARANDERIE, LAFONTAINE, LESSOIL, MEUNIER-TARDIF, ROBERT, TROCMÉ, ETC.

[4] LEBEL, Jean-Guy (1990), **Traité...,** voir chap. 2.3.2...

[5] LEBEL, Jean-Guy (1986 & 1987), **Le conditionnement phonétique: l'enjeu d'une nouvelle pédagogie en correction phonétique,** Département de langues et linguistique (Université Laval, Québec, Canada), 57 p., et dans *Revue de Phonétique Appliquée*, n^{os} 82-83-84: 183-190.

[6] LEBEL, Jean-Guy & LEBEL, Chantal (1989), " **Le workout en correction phonétique** ", *Bulletin de l'AQEFLS* (Association Québécoise des Enseignants de Français Langue Seconde), Montréal, XI,1: 40-46.

ACOUSTIC PHONETIC AND PROSODIC CORRELATES OF HINDI STOP CONSONANTS

Agrawal Shyam

CENTRAL ELECTRONICS ENGINEERING RESEARCH INSTITUTE CENTRE, NEW DELHI, INDIA.

ABSTRACT

This paper presents the result of some studies on acoustic correlates of segmental and prosodic features of Hindi stop consonants.The parameters havebeen used to classify the consonants according to their articulatory manner and place of production and also used to synthesize them using a synthesis by rule technique.Sixteen plosive sounds(voiced, unvoiced, aspirated, unaspirated) were recorded in CVC and VCV syllables using vowel /a/ by ten standard male speakers. The spectral analysis was done using a sound spectrograph, the FFT/LPC analysis programme and the microprosodic feature analysis programmes.The latter analysis was done on a PC(AT) with adequate facilities of A/D and D/A conversion.The segmental acoustic characteristics such as formants and thier transitions with adjuscent vowels(within each 8 or 10 msec frame.),bandwidths, gap(with/without voicebar),burst frequency,aspiration noise etc. and the prosodic characteristics such as variations in durations,fundamental frequency and amplitude were determined.

The individual sounds have been specified on the basis of distinctive acoustic characteristics and classify into various categories such as V/UV,Asp/Unasp etc. These characteristics have been very usefull in synthesizing Hindi words using a digital formant synthesizer based on synthesis by rule technique.

INTRODUCTION

The study of acoustic phonetic and the prosodic charateristics is essential to understand the basic nature ofthe speech sounds as well as to simulate and to develop automatic speech recognition and synthesis systems. These charateristics of speech are highly language dependent and vary considerably from one context to another context.Therefore a data base of the acoustic charateristics reflecting the segmental and suprasegmental characteristics of speech sounds is required for a spoken language under study.

The persent paper describes results of some studies conducted to study the acoustic correlates ofHindi stop consonants which describe their phonetic and prosodic features.

Hindi Stop Consonants

Hindi stop consonants, unlike many western spoken languages are distinguished by the features of aspiration as wellas voicing.There are four distinct places of articulation to produce them.Depending upon the place and manner of articulation, the consonants can be classified as shown intable 1. The paletal sounds are produced with retroflexion and hence they could be classified as retroflexes also.

Procedure

These sixteen stop consonants were combined with vowel /a/ to form CVC type as well as VCV type syllables. In CVC syllables, the final position of the syllable VC was kept the same and only initial cosonant was changed (e.g.pal,bhal etc.)whereas in the case of VCV syllables,the vowel /a/ was kept the same inthe initial as well as final position. These were recorded by 10 male speakers in a studio.

The analysis of the utterances were done using a sound spectrograph as well as a computer having A/D and D/A facilities. The signal processing software used for the analysis included FFT/LPC etc. Acoustic features such as formant frequencies and thier transition, fundamental frequency and duration of different segment of sounds were computed. The time varying display of suprasegmental features of the utterances was obtained using a SNDSYS programme and a prosodic analyzer(Inst.of Phonetics, Aix-en Provence). Microprosodic details such as variations in the fundamental frequency and amplitude was determined from these displays.

Results

The sonogram of the CVC and VCV syllables for a given speaker containing consonant t, th, d, and dh are shown in figurel.It may be seen from these displys that there are distinctive acoustic features which can be used to classify them. The aspiration noise associated with th is different from the aspiration noise associated with d. In the latter case amplitude of turbulance has superimposed upon it and voicing continues through out the spectrum. Similarly there are differences in the nature of plosive burst, voicebar and the formant transitions.

Durational Characteristics

The duration of different segments related to consonatal features have been analyzed. The features selected for measuring duration include the plosive gap/voicebar, the burst,aspiration, vowel transition and the steady portion of the vowel.The figures show duration computed by avaraging the duration obtained from various samples of all the speakers in CVC and VCV context.The following observations were made from this table.

a)Duration of the Gap/VB

Unv. unasp > unv.asp > voiced unasp > voiced asp
Bilab.> Velar > Dental > retroflex

b)Duration of Aspiration
Voiced Asp. >> unv. asp
(Neary same for all places of articulation)

c)Duration of the Burst

Voiced Asp.> Voiced unasp.>
unv. asp > unv.unasp.
Velar > Retroflex >
Dental > Bilabial

d)Voice onset time
Asp. > un. asp.
Velar > Retroflex >
Dental > Bilabial

It is also observed that the rate of first formant transition is much higher than the rate of second formant transition.

The rate,direction and target frequencies of adjuscent vowels play a major role in the classification of consonants.

Target frequencies

The target frequencies of the second and third formant of vowels obtained with different categories of consonants are shown in table 2. .These values are obtained by averaging the initial as well as final values of CVC ans VCV syllables.It may be observed from this table that the target frequencies ofsecond and third formants for the consonants belonging to different places of articulation are quite different. However, the target frequencies of fourth formant are not so distinct.It reflects more the characteristic of an speaker.In the case of retroflex sounds, the third formant indicate two different target frequencies in the VC and CV contexts. In the former case the third formant merges with the second formant whereas in the latter case there isa break in the third formant.

Rate of Formant Transitions

The average rate of formant transitions are much higher for the first formant as compared to the second and third formant transitions.

Fundamental Frequency

Measurements of fundamental frequency of different segments of CV and VC syllable have shown that the F of a voiced plosive is higher than that of the voiced aspirated plosive. Similarly F during voicing of plosion is much less than voicing during the vowel articulation. The fundamental frequency of the vowel in a CV syllable is higher than that in a VC syllable. The value of F of the vowel followed by an unvoiced plosive is higher than the value obtained with voiced plosive. In the former case there is an abrupt rise.This feature can work as a landmark (Cue)for indicating the presence of voiced or unvoiced consonant.

Conclusion

The results obtained in the above experiments show that the stop consonants possess some special and distinctive acoustic features.These can be classified on the basis of the features mentioned in the above results.

ACKNOWLEDGEMENTS

The author is grateful to Dr. W.S.Khokhle,Director CEERI Pilani for encouragement and according his kind permission to publish this work. He is thankful to his collegues, Dr. A.M.Ansari, Mr. M.Ganesan,Mr. S.K.Tyagi and Mr.Anil Kumar for various discussions and help in this work. He is thankful to Prof.Mario Rossi of Intt.Phonetique,Aix-en province and Prof.Ilse Lehiste of OSU, Colembus for assistance in doing part of the studies. Financial Assistance of DOE/KBCS project, Govt.of India is gratefully acknowledged.

REFERENCES

[1] Ansari A. M., Agrawal S. S., Ramakrishanan K.V. (1983),"An Electronic Counter for Frequency Measurament of Spectrographic displays". Jour.I.E.T.E, Vol.29, No.1, 33-34.

[2] Ansari A.M.,Agrawal S.S. (1984),"Control Methods for Acoustical Measurement using a Sound spectrograph". I.E.T.E. Tech. Review, Vol.1, No.2, 29-34.

[3] Blumstien S.E., Stevens K.N. (1980),"Perceptual Invariences and Onset Spectra for Stop Consonants in Different Vowel Environments". JASA, 67, 648-662.

[4] Ohman S.E.G.(1966), "Coarticulation in VCV Utterances, Spectrographic Measurements". JASA, 39, 151-168.

[5] Tyagi Sunil, Agrawal S.S., Pavate K.D.(1988), " Acoustic Parameters of Spoken VCV Syllables", Research Report No. CEERI/ESA/RR-5/88.

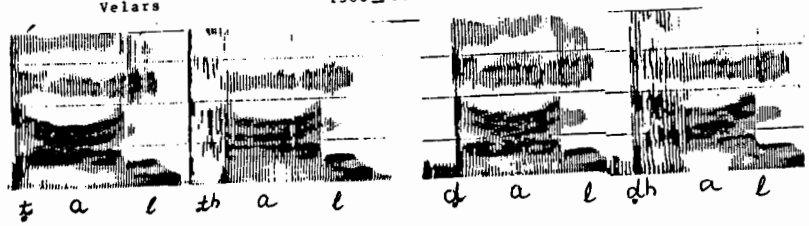
[6] Agrawal S.S.(1988),"Analysis and Synthesis of CV Syllables in Hindi Speech". 116th Meeting of ASA, Honolulu, USA.

TABLE 1

	Bilabial	Dental	Retroflex	Velar
Unasp unv.	p	t	ṭ	k
Unasp voiced	b	d	ḍ	g
Asp. unv.	ph	th	ṭh	kh
Asp. voiced	bh	dh	ḍh	gh

TABLE 2

	Second Formant	Third Formant	Fourth Formant
Bilabial	1000 ± 50	2300 ± 100	3400 ± 100
Dental	1500 ± 75	2600 ± 100	3600 ± 100
Retroflex	1800 ± 50	2700 ± 100 (CV), 1800 ± 100 (VC)	3600 ± 100
Velars	1300 ± 80	2500 ± 100	3500 ± 100



ELECTRONIC EDITION OF *PHONETIC SYMBOL GUIDE*

George D. Allen
Geoffrey K. Pullum and William A. Ladusaw

Purdue University, Lafayette, Indiana
University of California, Santa Cruz

ABSTRACT

Reference and teaching materials for use by scholars and students of phonetics have traditionally included books and audiotapes. A problem with these materials is that books do not contain sound and tapes do not contain text or pictures. However, recently available computer software makes possible the combining of text, sound, and pictures into so-called *hypertexts*, i.e. modular, linked documents which encourage exploration of material along multiple dimensions of information. This paper describes such a version of *Phonetic Symbol Guide* [1], written in HyperCard for Apple Macintosh microcomputers.

1. *PSG* – THE BOOK

Phonetic Symbol Guide [1], hereafter "*PSG*," is an encyclopedia of phonetic symbols. Except for an excursive *Introduction* section, *PSG* is organized as a series of short entries (approximately one-half to two pages in length) for each of the over 300 phonetic symbols "...that linguists, phoneticians, and other students of language are likely to encounter in reading either contemporary books and journals or older works that are important enough to be consulted today" (xvii). These include all of the standard IPA usages (prior to the 1989 revision) as widely employed by the international linguistic community as well as interpretations and symbols from other systems. Along with a large, clear picture of the symbol, each entry contains one or more sections, devoted to *IPA Usage*, *American Usage*, *Other Uses*, *Comments*, and *Source*. In addition to these entries for the individual symbols, there are also several charts, a glossary, a list of references cited, and a complete table of entries. Altogether, *PSG* presents a comprehensive view of the developing world standard established by the International Phonetic Association.

2. *PSG* – HYPERTEXT

The electronic version of *PSG* follows the structure of the book closely. It contains sections corresponding to those found in the book, namely **Introduction**, **Symbols** (corresponding to Character and Diacritic Entries), **Charts**, **Glossary**, **References**, and **Index** (corresponding to Table of Entries). In addition, however, it contains an instructional **Tutorial**, **Exercises** for students, and, most important, **sounds**. The text of the **Introduction**, **Symbols**, **Charts**, **Glossary**, and **References** sections is taken *verbatim* from *PSG*, whereas the other sections have been created especially for the hypertext version. Although the published version of *PSG* is useful primarily as a guide to the interpretation of phonetic symbols, the additions made in the hypertext presentation make it also a useful tool for learning phonetic transcription.

2.1. About HyperCard

HyperCard is a program which runs on Apple Macintosh microcomputers. Based on the metaphor of cards grouped into stacks, HyperCard contains a wide variety of tools for creating and linking modules combining text, graphics, and, most important for our purposes, sounds. It is thus ideally suited to representing so-called "exploratory hypertexts," i.e. large complex documents, such as encyclopedias, which are intended to be used in different ways by different readers. *PSG* is a perfect example of such a document.

2.2. Symbols Cards

As in the book, the most important section in the hypertext is the cards holding the information about individual symbols. Of the 311 entries in *PSG*, 107 were selected for inclusion in this version. (Expanded coverage of *PSG*, taking into account the 1989 revisions to the IPA, is anticipated.) These 107 represent most of the major entries in *PSG* (those with names in all capitals) and the ones of greatest interest to potential users of the document.

Figure 1 shows the screen display for the symbol *LOWER-CASE A* (page 3 of *PSG*). The upper part of the display shows the title, an enlarged picture of the symbol, a scrolling text field, and five buttons which permit the user to select which information appears in that text field. These five choices correspond to the presentation of information in *PSG*. The lower part of the display contains buttons which permit the user to **Hear it** (the sound), to navigate throughout the document (several buttons), and various other actions. This screen display format remains the same for all symbols in the document.

The scrolling text fields are "clickable," i.e. when the user clicks the mouse on selected items of text, appropriate actions occur. For example, Figure 1 shows that the words **Cardinal**, **low**, **front**, and **unrounded** appear in **bold** style. This indicates that clicking on any of them will take the user to the **Glossary** page for that word. Note also that the book title *Principles* appears in *italics*; clicking on it (or on authors' names) will take the user to the correct **References** page. Finally, words which are underlined (e.g. back, patte, pop) or which appear in square brackets (e.g. [pap]) are **pronounceable**, i.e. clicking on them causes their pronunciation to be played.

Among the buttons at the bottom of the display are the **Browse** buttons, consisting of left- and rightward pointing arrows, and buttons which take the user to the main **Menu**, to the overall system **Map**, to the **Charts**, to a **Help** screen, or out of the program (**Quit**).

2.3. Charts

The symbol charts follow as closely as possible the *PSG* format. For example, Figure 2 shows the screen display for the Cardinal Vowels 1–8, found on page 255 of *PSG*. Note that, in addition to the chart and the navigation buttons, there are two additional buttons called **Hear it** and **Find it**. These two buttons are mutually exclusive, i.e. turning one on turns the other off. When the **Hear it** button is "on," then clicking on a symbol in the chart causes its sound to be played. When the **Find it** button is enabled, on the other hand, clicking on a symbol takes the user to the card for that symbol. For example, if the user were to click on [a] (cardinal 4) with the **Find it** button on, as in Figure 2, the screen display would change to that of Figure 1.

All of the *PSG* charts are represented in the hypertext version, although there has been some reorganization due to constraints on Macintosh screen size and legibility. Furthermore, two of the charts (*Bloch and Trager's Vowel Symbols* and *The Chomsky/Halle Vowel System*) are not clickable, since their entries represent a more phonological than phonetic classification of sounds.

2.4. Glossary and References

The **Glossary** and **References** sections are organized alphabetically. All of the entries beginning with the same letter are found together in a scrolling text field.

2.5. Index

The **Index** is a set of three scrolling lists, one for **vowel** symbols, one for **consonant** symbols, and one for

diacritics. By clicking on the name of one of the symbols in these lists, the user is taken to the appropriate **Symbol** card (see Fig. 1).

2.6. Exercises

One of the exciting features of this electronic book is that it contains an infinitely expandable set of exercises. One kind of exercise is traditional listening practice, in which the user clicks on a test sound and has to find the correct symbol. Another exercise is of the symbol-to-feature type, the user being required to match symbols with their phonetic feature descriptions, and *vice versa*. A third kind of exercise, not normally found in phonetics classes, is what we call a "treasure hunt," in which the user is asked to find out certain information from the encyclopedic resources. An example might be "For which symbols do the IPA and American systems agree the most? For which do they disagree the most?"

3. INNOVATIVE FEATURES

The electronic *PSG* incorporates several features which are relatively unusual in instructional software. These relate to the social nature of learning and the advantages which come from collaboration in the classroom.

Recent research with so-called "intelligent tutoring systems" suggests that computers can never have enough "intelligence" to behave as true partners in learning. In other words, they remain artifacts, however complex. Thus, we have incorporated into the *Exercises* procedures for more than one user (e.g. two or more students, or a teacher plus one or more students) to use the hypertext simultaneously. For example, similar to some traditional listening and transcribing practice drills, one user serves as the talker, the others then being listeners. First, the talker views the target phonetic symbol or string and then hides it again. Then, as the listener attempts to transcribe the talker's productions, the talker obtains immediate feedback as the correctness of his or her productions from the sounds played and text viewed by the listener. In this situation, the hypertext *PSG* acts as a non-judgemental authority, available for instant consultation by the students.

Another collaborative aspect of the electronic *PSG* is the **Comments** button, available widely throughout the document (see Fig. 1). By clicking this button, the user is taken to a *Comments* screen, where comments by previous users may be read, and new comments may be entered. In this way, the user communicates with other users, including the authors, and a learning community is established. Frustrations with program content or operation are easily vented, leading to more positive views of the learning experience.

One final innovation which we shall mention here is what we might call the *auto-trace* feature. It is easy to capture the name of each button or field clicked on by users as they navigate the hypertext. This sequence of names is then used in two ways. First, it serves as an aid to the user, who can visit the list in order to recall where they have been recently. They do this by clicking the **See recent history** button (see Fig. 1). By simply clicking on one of the names in the list, they are then taken to the appropriate display. This list also serves a second, pedagogical use. When the user leaves the program (by clicking the **Quit** button), the list is written to disk, whence it can be viewed by the teacher (and eventually the authors) as evidence of the success or failure of the program in serving the needs of that user.

Together, these collaborative and auto-trace features make possible the evaluation and validation of the electronic *PSG* as a teaching tool. As we collect information on the document's use over the next few years, we will thus be able to extend and improve it even further.

4. REFERENCE

[1] PULLUM, G. K. & LADUSAW, W. A. (1986) *Phonetic Symbol Guide.* Chicago: University of Chicago Press.

Figure 1.

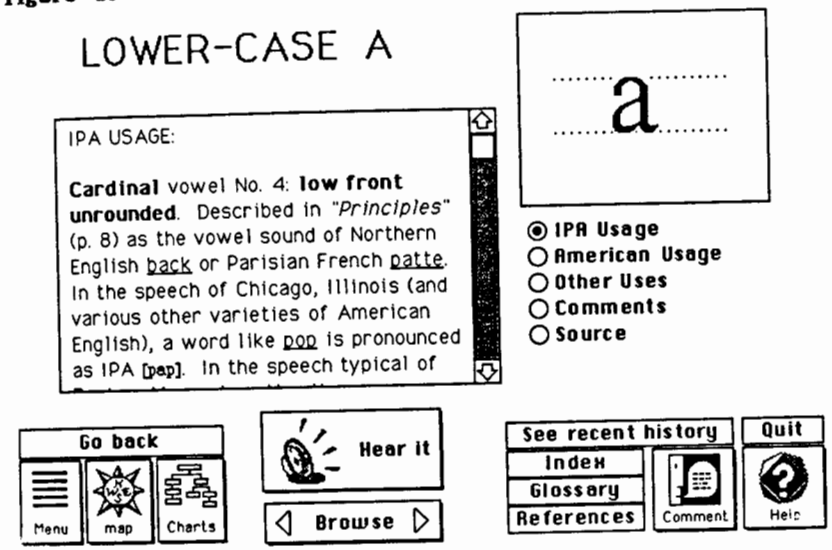
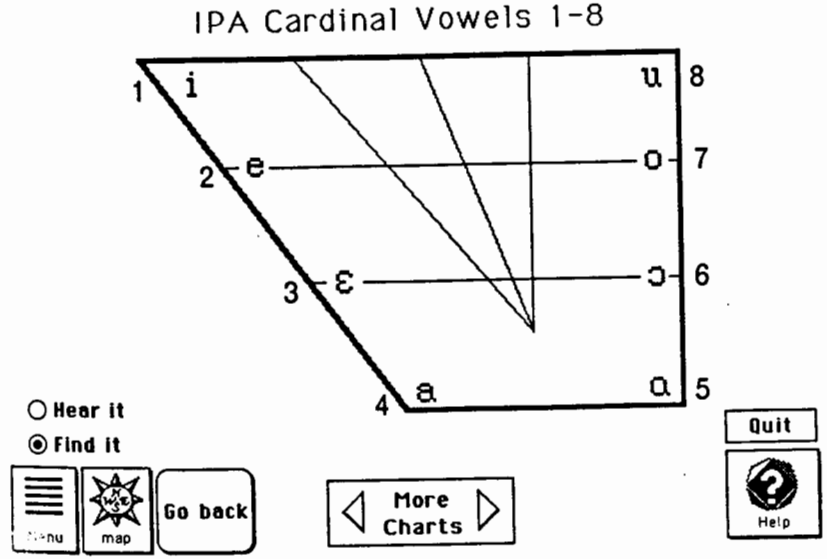


Figure 2.



MICROCOMPUTER-BASED INTERACTIVE PROSODY WORKSTATION

George D. Allen and V. Paul Harper

Purdue University and Harper Associates
West Lafayette, IN

ABSTRACT

One of the difficult problems facing teachers of phonetics is the lack of tools for training prosody (i.e., intonation, stress, and syllable rhythm). Our Interactive Prosody Trainer is a low-cost, microcomputer-based system for interactively teaching speech prosody. Based on a digital signal processing chip and an easy-to-use graphical interface, this device does two different, interrelated jobs. On output from the host microcomputer, it synthesizes models of utterances from stored LPC and prosodic control parameters. On input, it extracts the fundamental frequency and intensity of the user's productions, for comparison with the model. Similarities and differences between the two productions are then highlighted for the user.

1. INTRODUCTION

The teaching of prosody has not taken adequate advantage of modern speech technology. Beyond what is found in books or in the heads of phonetics teachers, there are just two audio-cassette resources and some not-very-interactive hardware. One of the audio-cassettes is the demonstration tape accompanying Cruttenden's text [1]. The author has simply read the examples from the text, as illustrations of the points he makes -- hardly a compelling pedagogy. The other cassette material is Allen's auto-instructional tutorial [2], used successfully for several years in his own phonetics classes but not disseminated widely.

Two (very similar) hardware products exist for training prosody, namely the Visipitch (Kay Elemetrics Corp.) and the PM Analyzer (Voice Identification, Inc.). Both devices extract fundamental frequency (F0) and intensity in real time and display them on a computer monitor. Both permit the visual comparison of a student's response to a teacher's model. Unfortunately, neither one can play back the model and/or the response for auditory comparison by the user, and both require the teacher to be present to evaluate the student's response. And they are both expensive. In other words, these devices are helpful aids for the teaching of prosody, but they require extensive one-to-one interaction with a trained professional, and a majority of training programs can afford to buy at most one.

Interestingly enough, Lane & Buiten [3] showed over 25 years ago that an interactive computer workstation could teach prosody effectively. Their so-called "Speech Auto-Instructional Device" (SAID) required users to match, as closely as possible, either the F0, the intensity, or the syllable timing of a model utterance. Using analog F0 and intensity extractors, plus a DEC PDP/1 minicomputer to calculate the match between model and response, the device cycled users successively among the three prosodic features until all three had converged to an acceptable degree. As successful as the SAID was in training fluent prosody, it is perhaps surprising that its principles have never been extended to modern microcomputers and digital signal processing technology. That time has now come.

2. A SAMPLE SESSION

The user, a young Chinese student who is improving his English, sits in front of the display screen and presses the **Model** button on his keypad. The computer presents the phrase "Good morning" with mid-level pitch on "Good" and a high-falling pitch on "morning." As the utterance is played out, white dots follow along the fundamental frequency (F0) and intensity traces on the screen. If he wishes, he can press the **Model** button to hear the utterance again.

When he is ready to respond, the user presses and holds down the **Talk** key. As he mimics the model with reiterant speech, which consists of repeated /ma/ syllables [4], his F0 and intensity are drawn on the screen in a different color and width of line from the model. Differences between the model and response prosodies are highlighted with shading, and "scores" are generated comparing the F0 and intensity traces and the timing of the syllables.

By pressing either the **Model** key or the **Response** key, the student can then hear "Good morning" uttered with the original prosody or with the prosody of his own response. As the utterance is played out, white dots trace out the prosody curves of the utterance to which he is listening.

At this point, the student has several choices. He may wish to produce another response, which he does by pressing and holding down the **Talk** key. Or, he may wish to review an earlier response, which he does by pressing a left-pointing arrow key the appropriate number of times (once for the previous response, etc.). Data from earlier responses are then presented on the screen, and he can again listen to either the model or the response. Finally, he may wish to move on to another utterance, which he does by pressing the **New** key.

As he works, all of his data are saved in a file, which is then written to disk at the session's completion. Later, the instructor will review this file, evaluating the student's progress.

3. WORKSTATION FEATURES

A prototype has been built using off-the-shelf components and well-documented speech processing algorithms. The hardware consists of a personal computer (PC) with a VGA color graphics display, microphone, speaker, and an audio I/O card with a digital signal processing (DSP) chip. The software consists of F0 and intensity extraction, speech encoding, and human interface modules.

Model utterances are stored in the PC, using linear prediction coefficients (LPC) plus prosodic information. There are two important advantages to using LPC coded models. One is simply the reduction in storage demands, in comparison with digitized speech. The other advantage is more important, however. Effective use of a prosody trainer requires that the model text be presented with either the model or the user's prosody. Since users of our device will respond with reiterant speech, the extracted F0 and intensity data can be substituted for the model data and used to drive the LPC re-synthesis. Thus, the user can easily toggle back and forth between the model as originally presented and the same words uttered with the prosody of the user's response.

Many algorithms exist for extracting F0 from speech [5]. Because reiterant speech is fully and continuously voiced, most F0 extraction algorithms work (equally) well. Speech intensity is usually also obtained as a by-product of this F0 extraction process. Thus, while the user is talking, current estimates of F0 and intensity can be delivered to the PC at the same rate as the model. Before the data are drawn on the screen, they are smoothed, scaled, and sometimes time-normalized. We say "sometimes," because there are situations in which the timing should be corrected, and others in which it should not.

F0 and intensity must also be appropriately scaled, both for the display and for the calculation of the fit between model and response. Both F0 and intensity are scaled logarithmically, to match our perceptions of these acoustic features. Scores for the degree of match between model and response are then

calculated as weighted averages of the differences between the data values. The use of weights permits focusing of attention on important speech features, such as vowel nuclei, without corruption by consonantal gestures.

Adjusting the timing of the response is another case of a difficult task made easier by the use of reiterant speech. Consonant/vowel (C/V) boundaries are relatively easy to locate from the intensity trace of repeated /ma/ syllables. These boundaries can then be aligned successively with pre-stored C/V boundaries of the model. Data points are deleted or interpolated in the F0 and intensity traces, as required, to match up the timing for each segment of the utterance. Finally, a score for the accuracy of timing can be generated as a weighted sum of the adjustments required to align the response to the model. Again, weighting these sums permits the user to focus on important temporal aspects of the model without becoming distracted by irrelevant features that happen to have been mis-timed.

Since the target users are persons with little or no experience in computer use, the interface must be simple and easily learned. As described in §2, above, a session with the Interactive Prosody Trainer requires the user to listen to a spoken phrase, mimic it using reiterant speech, and then view a graphical display showing his response compared to the original. He is then able to listen to the model phrase with its original prosody or with the prosody of his own response. This is similar to the SAID procedure [1], referred to earlier, in which learners alternately mimicked the F0, intensity, or timing of model phrases. In addition, however, our device permits the user to listen to the target phrase with his or her own prosody, in direct comparison with the model. This form of feedback is crucial for successful prosodic training.

4. REFERENCES

[1] CRUTTENDEN, A. (1986) *Intonation.* New York: Cambridge University Press.

[2] ALLEN, G. D. (1984) *Transcribing the Prosodic (Suprasegmental) Features of English.* Short Course presented at the Annual Meeting of the American Speech-Language-Hearing Association, San Francisco, CA, November 17, 1984.

[3] LANE, H. L. & BUITEN, R. (1965) A self-instructional device for conditioning accurate prosody. *International Review of Applied Linguistics*, 3, 205-219.

[4] NAKATANI, L. H. & SCHAFFER, J. A. (1978) Hearing "words" without words: Prosodic cues for word perception. *Journal of the Acoustical Society of America*, 63, 234-245.

[5] HESS, W. (1983) *Pitch Determination of Speech Signals.* New York: Springer Verlag.

A VIDEO INTRODUCTION TO GERMAN PHONETICS

Ursula Hirschfeld

Herder-Institut Leipzig, Germany

ABSTRACT

Learning the phonetics of a foreign language can be greatly facilitated by use of modern technology (video, computer). A video course is described which represents a first attempt at a systematic and concise introduction to German phonetics for students of German as a foreign language. The present summary of motivation, goals, content, and organization of the course will be supplemented at the conference with illustrative excerpts from the video tapes.

1. MOTIVATION

Phonetic research, despite its broad scope, is still not providing much help for those learning a foreign language. Language teachers and authors of instructional books tend to ignore the results of speech perception and production experiments, in part because these results are often reported in a way that makes it difficult to recognize their relevance to problems of second language acquisition. As a result, there is widening gulf between the state of the art in research and the quality of phonetic training in second language teaching. This relative neglect is evident in (a) undifferentiated teaching goals, (b) poorly developed methods of phonetics instruction, (c) insufficient phonetic training of language teachers who, moreover, often speak with a strong foreign accent themselves, and (d) lack of appropriate teaching materials.

Modern technology can aid second language learners in mastering phonetic problems by utilizing multiple information channels and by introducing new instructional methods. Unfortunately, current video language courses do not exploit these possibilities and often proceed without drawing students' attention to the phonetic peculiarities of the target language. Therefore, this author has developed a video course that provides a concise and systematic introduction to German phonetics for students of German as a foreign language [2].

2. GOALS

The course's aims are to enable students to discriminate and identify segmental and suprasegmental units on an auditory and articulatory basis, to lay the foundation for correct articulation of vowels and consonants, and to develop basic abilities of listening, speaking, reading, and writing. It is addressed to both beginning and advanced students.

3. CONTENTS

The course comprises eight lectures which focus on the following:

(1) Intonation (melody and stress).
(2) Unrounded vowels [ɑ:, a, i:, ɪ, e:, ɛ:, ɛ, ə].
(3) Rounded vowels [u:, ʊ, o:, ɔ, y:, ʏ, ø:, œ].
(4) Diphthongs, glottal stop, [h].
(5) Plosives [p, b, t, d, k, g].
(6) Fricatives and affricates [f, v, s, z, ts, ks, pf, kv].
(7) Fricatives and affricates [ç, j, x, ʃ, ʃt, ʃp], [r] sounds.
(8) Nasals [m, n, ŋ], -en suffix.

4. ORGANIZATIONAL PRINCIPLES

The construction of the materials is based on results of phonological and phonetic research, as well as on experiences gained from teaching German as a foreign language. Thus intonation stands at the beginning of the course, so as to enable students to produce all following exercises with correct stress patterns, melodies, and rhythms. This special position reflects the importance of suprasegmental factors in speech perception and comprehension. Vowels and consonants are introduced by their most important phonetic properties in the context of the whole sound system. They are not treated as isolated sounds but always in groups, with reference to systematic feature categories (e.g., "long-short" in vowels and "fortis-lenis" in consonants). Phonological principles are not discussed explicitly but are an implicit component of the course. Articulatory maneuvers are explained in simple terms.

Each lecture consists of three parts. The first part (demonstration) establishes relations within the phonemic system, gives examples, explains correct articulation, states rules, and elucidates sound-letter correspondences. In the second part (listening exercise), word pronunciations close to the norm are used in short interactive scenes, and individual and situational variants are introduced. The third part presents larger regional and individual variants: Various people in the streets of different German cities were videotaped producing examples from each lecture. Such "real-life" variants have so far been totally absent from phonetic teaching materials.

The materials make use both of traditional taeching devices (tables, rules, verbal hints) and of innovative methods (gestures, computer graphics [1]). Orthography and IPA transcription are also included. Explanations are provided in English (with alternative versions in Polish, French, Spanish, Italian, Russian, Hungarian, Slovak, and Czech).

The minimal lexicon of about 200 frequent words and proper names (of persons and cities) enables students at all levels to use the course, which can be combined with any existing language course. Moreover, the course's protagonists, an old Germanic warrior (puppet) who provides the explanations and a charming young woman (live) who speaks the examples, make learning fun.

Additional courses providing contrastive exercises in the nine languages named above are in preparation.

5. REFERENCES

[1] HEIKE, G., PESSARA, R., and STANKEWITZ, A. (1988). "UK-Dynamo, Version 1.0", Institut für Phonetik der Universität Köln.

[2] HIRSCHFELD, U. (1990). "Einführung in die deutsche Phonetik" (video course and booklet, Institut für Film, Bild und Ton, Berlin). München: Hueber Verlag.

DANS QUELLE MESURE PEUT-ON PREVOIR LA SEQUENCE PHONETIQUE QUI SE REALISE EN IMAGE ?
--- LA VOIX INTERIEURE QUI PRECEDE L'EMISSION DE VOIX ---

Keiichi Kojima

Université Seitoku

ABSTRACT

We examined the duration of "phonetic unity realised in an oscillogram analysis of 9 vocal phenomena of French speakers. The "phonetic unity " examined is to be distinguished from the notion of "rhythm group " as it is generally understood. It is phonetic coherence imaged in the brain in spoken language. The examination evinced a " moment of coherence " of 150 - 1000 ms. This is the extent of the moment in which the brain creates the image of an utterance before its vocalisation.

1.INTRODUCTION

La voix est chargée implicitement de sens. Quand donc ce sens est-il réalisé chez le locuteur ? Cela aurait quelque chose à voir avec la voix intérieure qui existe en tête. L'émission de voix est un résultat de <feed-forward> dans la tête du locuteur. Et maintenant nous devons chercher quelques méthodes d'après les quelles nous pourrions tirer des jugements sur la durée de ce temps secret de <feed-forward> qui précède l'émission.

2. MOTIVATION DE CETTE ETUDE

La raison pour laquelle nous nous intéressons à ce problème provient de la phrase française-type suivante : « C'est l'homme dont je vous ai parlé.» où on se sert du pronom relatif <dont>. Selon l'interprétation grammaticale, ce <dont> est nécessairement introduit par le verbe suivant <parlé>. Si la voix est exprimée en poursuivant la pensée, restera à savoir ce qu'est cette pensée au moment où l'on a exprimé le mot <dont>. Si l'on choisit le mot <dont> dans la séquence de la langue parlée, cela indiquerait une prévision du mot <parlé> quand on exprime ce mot <dont>. Nous allons nous demander jusqu'à quel point la voix sans sonorité (voix intérieure) devance l'émission.

3. METHODE D'ANALYSE

Nous allons observer, en analysant les ondes oscillographiques enregistrées par deux Français et un Italien (pas de lecture de texte, parler naturel) à la vitesse lente de 10 cm/sec., combien d'éléments (phonème, syllabe, mot, groupe de mots que nous ne connaissons pas encore) et à quelle place le locuteur cherche au cours de la séquence.

4. OBJET DE NOTRE ETUDE

Lors de l'apprentissage de la langue, se pose la question de l'unité phonétique, c'est à dire plus petit ensemble d'éléments phonétiques. On appelle en général cet ensemble le groupe rythmique. Le groupe rythmique comme signe structurel de langue possède un caractère phonétique, mais la voix intérieure précédant l'émission ne nous semble pas

toujours posséder assez de durée pour nous montrer le signe structurel. C'est plutôt un phénomène physiologique commun parmi les hommes et nous ne pouvons pas encore justifier la valeur linguistique de cette limite. Mais nous pouvons dire que l'analyse de ce phénomène est une révision de la notion de groupe rythmique et nous devons maintenant assouplir cette notion d'unité pour reconnaître d'autres successions phonétiques. Par exemple, dans la séquence qui suit : «En m'excusant encore de ne pas vous l'avoir donné tout de suite ...» nous avons au moins 3 groupes rythmiques possibles : En m'excusant ..., de ne pas vous..., tout de suite. Or, si on la coupe de la sorte, nous sentons quelque chose d'étrange en ce qui concerne la structure : En m'excusant, encore, de ne, pas vous l'avoir, donné tout de suite. Cependant la réalité de ces coupures, faits phonétiques, nous oblige à accepter ce curieux phénomène.

5. INFORMATEURS

1. Français :31 ans, né à Paris, nationalité française. utilisation du japonais pendant 6 ans, 15 ans à Paris, 3,5 ans à Tahiti, 10 ans à Chevreuse, 3 ans au Japon.
2.Italien : monologue à la radio. Pas d'informations détaillées.
3.Français:monologue à la radio. Pas d'informations détaillées.

6. INTERPRETATION DES ONDES OSCILLOGRAPHIQUES

C'est sur l'oscillogramme que nous considérons qu'existent des coupures temporelles reconnaissables.

6.1.Après le prolongement du son.
« En m'excusant / encore / de ne ....(67/100sec)/ pas vous l'avoir / donné...»(informateur 1) [ə] de [nə] a une durée de 67/100 sec. qui nous semble bien exprimer le fait que le locuteur cherche les éléments qui suivent.

6.2. Après l'arrêt du son.
C'est le temps pour chercher les éléments qui suivent, qui est remplacé par l'inspiration ou coup de glotte. « Voilà monsieur. / En m'excusant / encore ... (50/100sec, inspiration) » (informateur 1)

6.3. Après l'insertion du son.
[ə] ou [ø] apparaît dans la séquence. « ... était à peu près [ə] ...(18/100sec) / trois mille ... » (informateur 2)

6.4. Après l'intonation montante rapide. (chez les Japonais, après l'intonation descendante.)
Cela s'accorde avec le changement de hauteur à la fin du groupe rythmique français. « Je suis arrivé / à Londres ... » (informateur 3)

6.5.A la place où est provoqué le décalage d'amplitude au cours de deux mêmes sons.
« En m'excusant / encore ... » (informateur 1) «16e siècle / le nombre ...» (informateur 2)

6.6. Au changement d'un son fort en faible.
Les amplitudes ne se comparent pas en différence acoustique absolue mais relative. « ... pas vous l'avoir / donné tout de suite ... » (informateur 1) [a] de [avwaʁ] étant fort, nous pouvons mettre une coupure devant <donné>. Pourtant si on prononce de suite sans coupure consciente <avoir donné>, où le dernier [a] de [avwaʁ] n'est pas fort en comparaison du [ɔ] suivant, on observe alors que l'amplitude des [ɔ] [e] suivants se rapproche de celle de [a] précédant. Dans cet état, nous considérons ce [a] de [avwaʁ] plutôt comme faible. C'est-à-dire dans cette situation-ci <avoir donné> se groupe en une unité phonétique.

6.7. Au changement du son faible en fort.

«...passer de /bonnes vacances» (informateur 1)

6.8. Après une grande quantité vocalique.

« Si / vous voulez... » (informateur 1)

6.9. Après la détente lente de l'amplitude.

« Voilà, / merci. Merci encore ... » (informateur 1)

7. REVISION DE LA NOTION DE GROUPE RYTHMIQUE

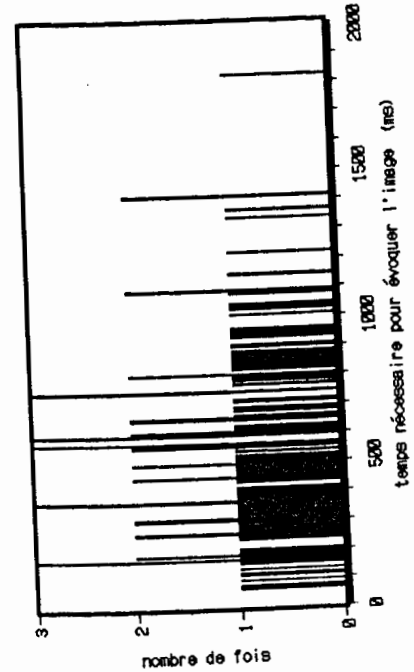
Bien que le groupe rythmique soit le plus petit ensemble phonétique formant une unité de sens, il n'est pas toujours exprimé en unité de groupe rythmique, étant donné que la pensée et la voix sont un amalgame indivisible. La preuve en est que la voix, dans la langue parlée, est émise parfois avant que la pensée développée en tête ne comble l'unité structurelle de sens. L'ensemble phonétique dans ce cas n'est pas l'unité sémantique. Nous appelons cet ensemble "le groupe rythmique physiologique". Dans <... de ne pas vous l'avoir donné tout de suite.>, <l'avoir donné> est censé être un ensemble théorique (grammatical), mais la voix dans la langue parlée fait coupure après <l'avoir>, ce qui est le signe qu'on cherche les éléments qui suivent. Et dans ce cas, <donné (donner) tout de suite> est un ensemble. L'ensemble phonétique que le locuteur retourne dans sa tête avant l'émission nous semble plus court que le groupe rythmique.

A propos, la voix émise réalise-t-elle le mécanisme structurel qui demeure à l'état latent dans les profondeurs? Nous disons que oui. Bien qu'on y sente quelque chose de grammatical, notre opinion est plutôt que la relation entre voix et paroles constitue l'essence même du discours. Dans l'exemple «Le visage visible dévoile le cœur invisible.», la structure peut être analysée en trois unités sémantiques : le visage est visible, il dévoile le cœur invisible, le cœur est invisible. Ce n'est que de la structure. <Le visage visible> est clairement différent de <le visage qui est visible>. Celui-là est en une unité d'image et celui-ci est en deux unités. Sur ce point, dans la tête du locuteur, il est clair du point de vue phonétique que <visible> s'accorde en tant qu'image avec la séquence précédente ou bien s'en sépare. Bien sûr, <le visage visible> peut se séparer ainsi : <le visage> <visible> sans utiliser <qui est visible>. C'est-à-dire soit une unité <le visage visible>, soit deux unités <le visage / visible>. Le contenu <qui est visible> peut être modulé dans celles-ci avec des éléments phonétiques accentuel, intonatif, etc. ... Aussi l'image qu'a en tête le locuteur varie selon le mouvement phonétique bien qu'en surface l'écriture apparaisse identique. Il y a certainement une unité de <groupe rythmique> en français mais le rapport de P. Delattre se rattache à l'image de tête du locuteur : «...toute syllabe finale d'un mot qui porterait l'accent à l'état isolé a tendance à ne pas désaccentuer complètement à l'intérieur du groupe. La désaccentuation est d'autant moins complète que le mot est plus important. » (1966. p.143). Nous trouvons ici la coupure de l'image de tête (feed-forward) et nous pouvons retrouver là les unités émiettées de l'image. Or, le mouvement phonétique dans ce cas-là devient caractéristique sur la partie finale <-sible> pour <le visage visible> et sur la partie <-sage> pour <le visage (qui est) visible>.

8. ETABLISSEMENT D'UNE HYPOTHESE DE FEED-FORWARD

Une certaine image B provoquée au cours d'une certaine image A se réalisera phonétiquement après le mouvement d'émission continue de l'image A. La durée temporelle de B est égale à la longueur du feed-forward qui est formé au cours de A. Il y aurait des mécanismes plus compliqués en réalité dans lesquels interviendraient des changements physiologiques qui sont au-delà de notre connaissance en matière de phonétique. Nous avons besoin de la relation entre la durée de la partie prolongée ou de la partie amuie (qui constituent toutes deux des recherches de la pensée)et celle de l'ensemble phonétique qui les suit. Cette relation est utile pour voir dans quelle mesure l'image avance dans ces parties.

Nous représentons par un graphique l'état de distribution de la durée temporelle.



La partie de suspension (repos d'inspiration) apparaît par endroits mais nous ne considérons pas cet intervalle en tant que le temps mis à chercher les éléments qui suivent. C'est parce que l'on ne peut pas deviner quand le locuteur au cours de l'intervalle a commencé à chercher les éléments qui suivent. Nous trouvons 1830 ms (... tout de suite. --- J'espère ...), dont toute la durée ne pourrait cependant pas être le temps mis à chercher les éléments qui suivent. C'est une partie où il y a un temps d'inspiration qui finit le sujet précédent, et le locuteur en inspirant commencerait à chercher les éléments qui suivent. Nous pouvons enlever une telle durée remarquable. Pour finir, la durée la plus fréquente est concentrée dans la graphique entre 150 ms et 1000 ms. C'est le temps nécessaire pour évoquer l'image. La vitesse avec laquelle le locuteur parle ne serait pas pertinente. La partie prolongée et l'amuie, semble, dans la phrase, plus courte que l'ensemble phonétique qui suit la partie prolongée et l'amuie. Si on peut retrouver ce fait dans beaucoup de cas, cela veut dire que l'image se forme en tête plus rapidement dans le temps que la voix émise d'après cette image.

9. CONCLUSION

La valeur numérique mesurée de 150 ms à 1000 ms est un moment dans une seconde. Il est important donc d'acquérir constamment l'ensemble phonétique réalisé dans une seconde environ et quand on parle, il convient d'utiliser la voix intérieure en découpant l'image la plus proche possible de l'émission.

10. REFERENCE

[1] DELATTRE,P.(1966),"*Studies in French and Comparative Phonetics*", The Hague : Mouton & Co.

ARTICULATION-BASED TACTILE SPEECH FOR THE DEAF: A COMPLETE SET OF TACTILE SEGMENTAL FEATURES FOR GERMAN

H. G. Piroth and H. G. Tillmann

Institut für Phonetik und Sprachliche Kommunikation
der Universität München, Germany

ABSTRACT

This paper presents the definitions for a synthesis of quasiarticulatory tactile speech stimuli and the method of presentation as developed in a German Research Council project concerned with tactile syllable equivalents.

1. INTRODUCTION

Most of the investigations concerned with tactile speech for the deaf use systems that transmit tactile transforms of the acoustic speech wave to the skin. (For an overview see [5]). As opposed to these approaches, the concept of a speech-to-skin transmission system has been proposed by us that fully relies on articulatory information to code tactile speech equivalents [1,6].

Such a system needs to contain two main components. The first has to extract the articulatory information from the speech wave, and the second has to transform it into tactually well-distinguishable patterns. To investigate the general applicability of such a concept our initial experimental research has been limited to the second component.

In several previous investigations proposals for the coding of the articulatory features to construct a tactile system of vowel and obstruent equivalents have been made and modified dependent on the test results (e.g. [2,3]). In preparation of the test stimuli for an investigation of the recognizability of tactually presented words [4] the need for a complete and consistent set of tactile features to code at least the phonemic distinctions of Standard German arose. Only labialization is not included in this version, since lip movements are easily detectable by lip-reading.

2. APPARATUS AND GENERAL METHOD

For the execution of experimental research SEHR ("System for Electrocutaneous Stimulation") has been developed. SEHR enables the computer-controlled synthesis and presentation of electric pulse trains to the skin.

A PDP-11 is connected with a 16-channel stimulus generation device. SEHR produces current-controlled bipolar impulses without a d.c.-component. The impulse forms of the versions SEHR-2 and SEHR-3 are given in Fig. 1. Using

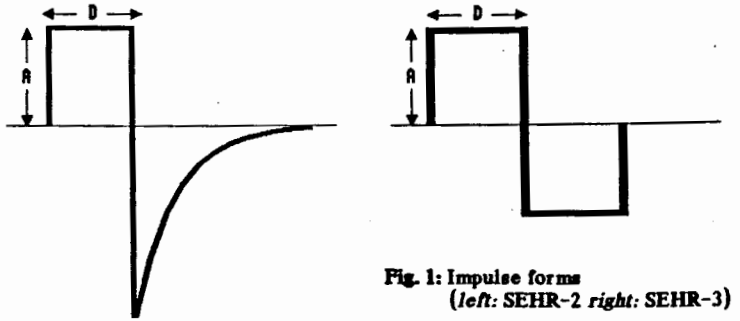


Fig. 1: Impulse forms
(*left:* SEHR-2 *right:* SEHR-3)

version SEHR-2 the pulse repetition rate can be manually adjusted between 100 and 500 pps. The duration of the rectangular part of the impulse is software-controlled and variable between 0 and 500 µs. Impulse amplitude is digitally adjusted to one of 64 steps between 0 and 5 mA. For SEHR-3, also pulse repetition rate is preset via software. Durations are allowed to vary in steps of 32 µs between 0 and 512 µs and intensities in steps of 0.33 mA from 0 to 4.95 mA.

The channel outputs of SEHR are delivered to the subjects via 16 pairs of circular gold-layered electrodes (9 mm in diameter and 1 mm apart from one another). For the experiments on an articulation-based system for tactile speech presentation electrodes are placed as shown in Fig. 2. The PDP-11 is equiped with an AD-converter and a Schmitt-trigger that receive their inputs from a small box with a potentiometer knob and a button. The software package includes several procedures for intensity threshold determination. Thus, during a calibration procedure at the beginning of each test session Ss can adjust impulse intensities separately for each channel by turning the knob and store the desired values by pressing the button. For test presentation, the package contains several modules that take detailed stimulus descriptions (sequence of channels, numbers of impulses, intervals between successive pulse trains) as input as well as the calibration results (pulse amplitudes and durations) and randomization lists for identification or discrimination tests. Some identification test modules allow answering by pressing keys on the computer keyboard and give confusion matrices of the results as output.

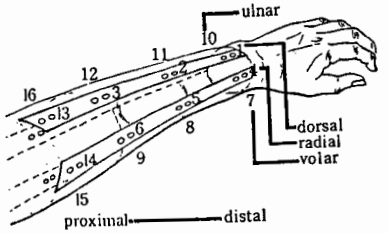


Fig. 2: Electrode arrangement for tactile stimulation

3. THE CODING METHOD

The forearm has been chosen as tactile stimulation area, since it can be interpreted as a mapping of the local relations within the vocal tract. The stimulus patterns are dynamic sequences of pulse trains consisting of three impulses with a duration of 256 µs for each rectangular part and an amplitude corresponding to the subjective mid value between absolute and annoyance thresholds.

Under these conditions, quasiarticulatory patterns are defined on the syllable level: syllables equal in narrow phonetic transcription are represented by equal tactile patterns consisting of subpatterns representing the segmental structure of the syllables. The basic distinction between vowels and consonants is coded by the distinction between longitudinal and circumferent subpatterns (i.e. subpatterns moving along the arm or subpatterns surrounding it).

The coding features of vowels are height (from dorsal to volar) and front/back (distal/proximal). According to this phonetic feature transformation, /i/ is a subpattern oscillating between the two distal electrode pairs on the dorsal side, /u/ between the two proximal ones of the same side, a front /a/ is a subpattern oscillating between the two distal electrodes on the volar side. For a complete coding of vowel height more than two distinctions have to be made: mid vowel height is transformed to a subpattern oscillating both at the ulnar and radial sides of the arm. To create a fourth level, /ɛ/ moves along the radial and ulnar as well as the volar side of the arm. (For details see Tab. 1.)

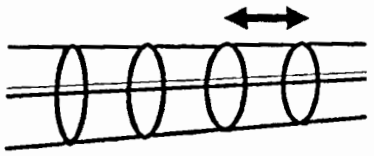
For consonants, the front/back distinction is applicable as well, /f/ surrounds the arm at the distal electrode ring, /h/ at the proximal one, others are in between according to their places of articulation. Since more places have to be coded than electrode rings are available, some fricatives are represented as double rings contrarotating at neighbouring electrode rings. The fortis/lenis feature is coded as a difference in inter pulse train intervals. Fortis fricatives have twice as many pulse trains as lenis fricatives and a shorter inter pulse train interval to preserve equal overall durations. Plosive subpatterns, as opposed to fricatives do not form a complete ring, but after start-

Tab. 1: Sequences of Electrode Pairs Stimulated for Subpatterns (Consonants Initial before /a/)

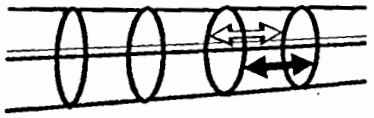
Duration of a single Pulse Train: 5.2 µs
Interval Between Pulse Trains: IPTI [ms]

No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Vowels (IPTI=20)																
/i/	1	2	1	2	1	2	1	2								
/a/	8	9	8	9	8	9	8	9								
/u/	3	13	3	13	3	13	3	13								
/e/	4	5	11	10	10	11	5	4								
/o/	6	14	16	12	12	16	14	6								
/ə/	5	6	12	11	11	12	6	5								
/ɛ/	4	7	5	8	11	8	10	7								
Fricatives (IPTI$_{fortis}$=15, IPTI$_{lenis}$=35)																
/f/	1	1	4	4	7	7	10	10								
/v/	1	4	7	10												
/ʃ/	2	2	5	5	8	8	11	11								
/ʒ/	2	5	8	11												
/x/	3	3	6	6	9	9	12	12								
/h/	13	13	14	14	15	15	16	16								
Fricatives (IPTI$_{fortis}$=7, IPTI$_{lenis}$=15)																
/s/	1	1	8	8	4	4	5	5	7	7	2	2	10	10	11	11
/z/	1	8	4	5	7	2	10	11								
/ç/	2	2	9	9	5	5	6	6	8	8	3	3	11	11	12	12
/j/	2	9	5	6	8	3	11	12								
Plosives (IPTI$_{fortis}$=7, IPTI$_{lenis}$=20)																
/p/	1	4	4	4	4	4										
/b/	1	4	4	4	4	4										
/t/	2	5	5	5	5	5										
/d/	2	5	5	5	5	5										
/k/	3	6	6	6	6	6										
/g/	3	6	6	6	6	6										
/ʔ/	13	14	14	14	14	14										

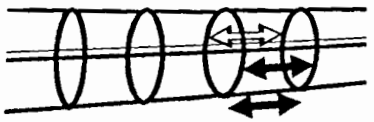
Example: /i/



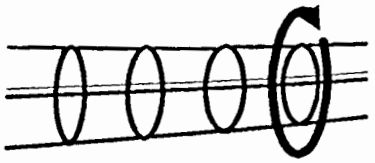
Example: /e/



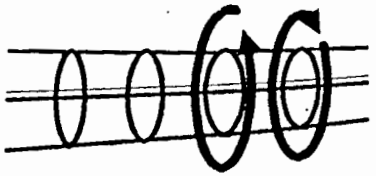
Example: /ɛ/



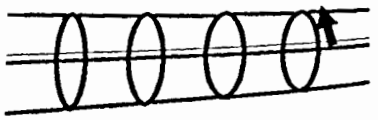
Example: /f/ or /v/



Example: /s/ or /z/



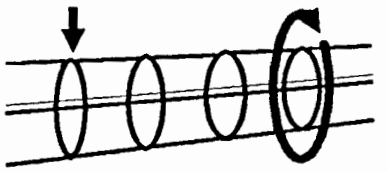
Example: /p/ or /b/



Nasals (IPTI=8)

/m/	1	1	13	13	4	4	13	13	7	7	13	13	10	10		
/n/	2	2	13	13	5	5	13	13	8	8	13	13	11	11		
/ŋ/	3	3	13	13	6	6	13	13	9	9	13	13	12	12		

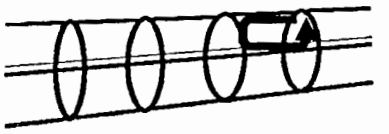
Example: /m/



Liquids (IPTI=15)

/r/	1	2	5	4	1	2	5	4
/l/	2	3	6	5	2	3	6	5
/ʁ/	3	13	14	6	3	13	14	6

Example: /r/



ing at one point of the electrode ring corresponding to their place of articulation, the subsequent pulse trains are statically delivered to the place neighbouring the starting point to simulate their noncontinuous production. Fortis plosives move fast and have a pause added afterwards that resembles the occlusion phase. Lenis plosives move more slowly and lack the pause to keep overall durations of plosives constant.

Nasals are coded like the corresponding lenis fricatives, but with intermediate pulse trains at the dorsal proximal electrode pair to imitate velum behaviour.

Liquids are circular movements using the dorsal and radial side electrodes of two neighbouring rings according to their places as described below. To include an analogue of coarticulation consonantal subpatterns start where the preceeding vowel stops or stop where the subsequent vowel starts. Thus, the exact electrode pairs involved in a tactile consonant equivalent change with the context vowel giving additional hints for the recognition of vowels with the effect that the only patterns that are exactly identical are recurring syllable equivalents, not segments.

In this system overall durations of consonants and vowels are constant (except for some subthreshold variations) for experimental purposes, but durational variations of careful and explicit speech could be implemented without changing the main characteristics of the percept.

4. REFERENCES

[1] PIROTH, H.G.(1985), "Elektrokutane Silbenerkennung mit quasi-artikulatorisch kodierten komplexen zeitlich-räumlich strukturierten Reizmustern", *Forschungsberichte d. Instituts f. Phonetik u. Srachl. Komm. Univ. München 22.*

[2] PIROTH, H.G.(1987), "Incorporation of the fortis-lenis feature in a quasiarticulatory system of tactile speech synthesis by adding temporal variations", *Proc. 11th ICPhS, Vol. 1*, 369-372.

[3] PIROTH, H.G. et al.(1989), "Electrotactile encoding and recognition of a 16-obstruent system", *J.Acoust.Soc.Am. 86, Suppl.1*, S95.

[4] RENZELBERG, G.(1990), "Taktile Analoge", *Phil. Diss.*, Munich.

[5] SHERRICK, C.E.(1984), "Basic and applied research on tactile aids for deaf people", *J.Acoust.Soc.Am. 75*, 1325-1342.

[6] TILLMANN, H.G.(1970), "*Technische Kommunikationshilfen für Gehörlose*", Berlin: Marhold.

SPEECH-PROSODY CHARACTERISTICS OF YOUNG CHILDREN WITH SPEECH DISORDERS OF UNKNOWN ORIGIN

Lawrence D. Shriberg, Ph.D.

Department of Communicative Disorders and Waisman Center
on Mental Retardation and Human Development
University of Wisconsin-Madison

This report displays selected phonetic, phonologic, and prosodic findings for 64 children with speech disorders of unknown origin. Descriptive and inferential statistics provide some support for subgrouping, with associated research attempting to characterize the phenotype(s) for genetically-transmitted speech-delay. [Supported by the National Institutes of Health, NIDCD, No. 26246]

The Phonology Project at the University of Wisconsin-Madison has developed and validated a computer-assisted assessment protocol, administered the protocol to samples of children and adults with speech disorders of known and unknown origin, and conducted cross-sectional, longitudinal, and intervention studies posing questions in description, explanation, prediction, and intervention [e.g., 4, 8, 9, 13, 14]. Project studies are organized conceptually by a seven-category classification of the possible origins of developmental speech disorders of heretofore unknown origin [7]. Long term goals of primary, secondary, and tertiary prevention include the identification of the speech phenotypes associated with genetically-transmitted speech delays and the development of a discriminant function for differential diagnosis.

METHOD

Selected findings for this poster session are taken from a sample of 64 3-6 year-old children with moderate to severe speech disorders of unknown origin. All data collection and analyses procedures have been developed and reported in prior work, including procedures for sampling conversational speech [7], accomplishing narrow phonetic transcription by consensus [5, 10], coding and entering transcriptions for computer-aided phonological analysis [3], and procedures for prosodic analysis of conversational speech samples [11]. Averaged interjudge and intrajudge agreement for the two consensus transcription teams scoring the articulation test and conversational speech samples was 65.5% to 81.1% for narrow phonetic transcription and 86.7% to 95.1% for broad phonetic transcription. These figures are consistent with other reports in disordered child phonology [12]; most of the data are based on findings at the level of broad transcription.

The procedures described in Shriberg & Kwiatkowski [6] were used to assign children to five of the seven putative diagnostic classification categories based on all protocol information other than the speech data: (a) hearing (fluctuating hearing loss secondary to early recurrent otitis media with effusion), (b) dysarthria, (c) apraxia, (d) psychosocial, and (e) noninvolved, reflecting clear non-qualification for any of the other six categories (there were no children meeting criteria for a category termed 'structural', i.e., craniofacial). These children were not frankly hearing impaired, dysarthric, apraxic, or emotionally disturbed; rather, their case history data and responses on protocol tasks indicated possibly subtle involvements in these domains. The children were also classified into three language production involvement groups (at expected level, up to one-year behind, greater than one year behind) based on their structural stage development [1, 2].

FINDINGS

1. The sex distribution in this study was 64% boys-36% girls, compared to previous estimates in our work closer to 3:1, which are consistent with sex-linked or sex-influenced polygenic threshold models of genetic transmission. A more recent estimate based on a database of 212 speech-delayed children yielded a ratio of exactly 3:1. Unlike gender findings in the dyslexia and learning disabilities literatures, ascertainment bias is not likely in these data.

2. A graphic representation of the consonant error pattern for the entire group can be divided into a three-part function for child phonology research. Comparison of the function to mastery data for early, middle, and late-occurring sounds in normal development indicates good concordance (see Figure 1). The few discrepant points in the normal-speech data are readily accounted for by differences associated with citation-form sampling and level of phonetic transcription. The similarity in the two profiles is viewed as support for the nosological term delayed speech (as opposed to disordered speech), and the profile will be tested as a potential phenotype for this classification category.

3. When phonetic, phonologic, and prosodic data are plotted by the three language involvement groups, clear differences are observed in both the severity and pattern of involvement in each domain. Such findings have implications for clinical classification issues--the continuing debate on articulatory vs. phonological disorders, methodological issues--the need for

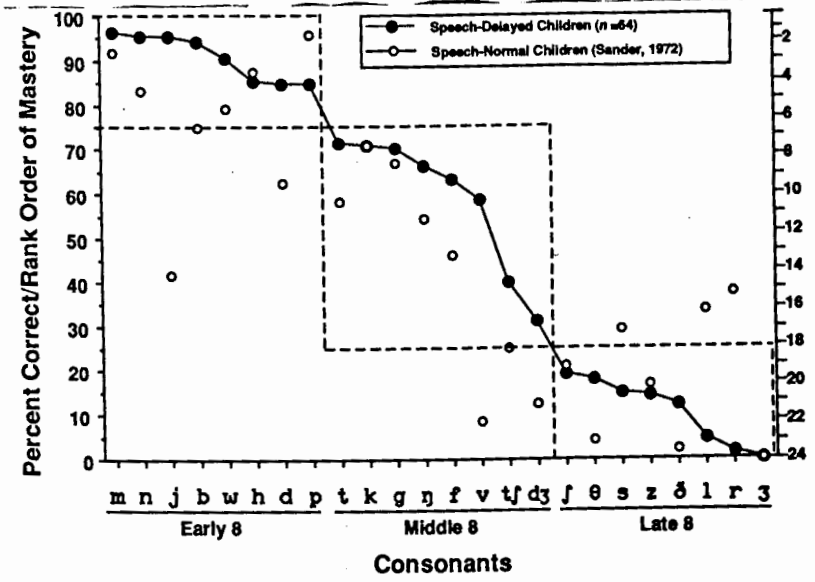


Figure 1. Consonant Acquisition in Speech-Delayed and Speech-Normal Children

comparable subject definitions among child phonology researchers, and clinical issues--the development of instruments for subgrouping, prognosis, and intervention.

4. When divided into the etiological groups on the basis of performance on non-speech measures, most children met inclusionary criteria for more than one category with relatively few 'pure' groups remaining for statistical analysis: hearing, 9 Ss; dysarthria/apraxia, 7 Ss; psychosocial, 7 Ss; no involvement 14 Ss; total n = 37. Phonetic, phonologic, and prosodic profiles for these small groups provide some support for subgroups based on speech-language performance.

REFERENCES

[1] MILLER, J. (1981), "*Assessing language production in children: Experimental procedures*", Baltimore: University Park Press.

[2] SCARBOROUGH, H. (1990), "Index of productive syntax", *Applied Psycholinguistics, 11*, 1-22.

[3] SHRIBERG, L. (1986), "*PEPPER: Programs to Examine Phonetic and Phonologic Evaluation Records*", Hillsdale, NJ: Lawrence Erlbaum.

[4] SHRIBERG, L., & ARAM, D. (in submission), "Speech characteristics of children with lateralized brain lesions".

[5] SHRIBERG, L., & KENT, R. (1982), "*Clinical phonetics*", New York: Macmillan.

[6] SHRIBERG, L., & KWIATKOWSKI, J. (1982), "Phonologic disorders I: A diagnostic classification system", *Journal of Speech and Hearing Disorders, 47*, 226-241.

[7] SHRIBERG, L., & KWIATKOWSKI, J. (1985), "Continuous speech sampling for phonologic analyses of speech-delayed children", *Journal of Speech and Hearing Disorders, 50*, 323-334.

[8] SHRIBERG, L., & KWIATKOWSKI, J. (1988), "A follow-up study of children with phonologic disorders of unknown origin", *Journal of Speech and Hearing Disorders, 53*, 144-155.

[9] SHRIBERG, L., KWIATKOWSKI, J., BEST, S., HENGST, J., & TERSELIC-WEBER, B. (1986), "Characteristics of children with phonologic disorders of unknown origin", *Journal of Speech and Hearing Disorders, 51*, 140-161.

[10] SHRIBERG, L., KWIATKOWSKI, J., & HOFMANN, K. (1984), "A procedure for phonetic transcription by consensus", *Journal of Speech and Hearing Research, 27*, 456-465.

[11] SHRIBERG, L., KWIATKOWSKI, J., & RASMUSSEN, C. (1989, November), "The Prosody-Voice Screening Profile (PVSP): I. Description and psychometric studies; II. Reference data and construct validity", Paper presented at the Annual Convention of the American Speech- Language-Hearing Association, St. Louis, Missouri.

[12] SHRIBERG, L., & LOF, G. (in press), "Reliability studies in broad and narrow phonetic transcription", *Journal of Clinical Linguistics and Phonetics.*

[13] SHRIBERG, L., & WIDDER, C. (1990), "Speech and prosody characteristics of adults with mental retardation", *Journal of Speech and Hearing Research, 33*, 627-653.

[14] THIELKE, H., & SHRIBERG, L. (1990), "Effects of recurrent otitis media on language, speech, and educational achievement in Menominee Indian children", *Journal of Native American Education, 29*, 25-35.

TEMPORAL VARIABLES FOLLOWING UNILATERAL LEFT OR RIGHT HEMISPHERE LESION

P. Bhatt

Experimental Phonetics Laboratory,
University of Toronto, Toronto, Canada.

ABSTRACT

This paper presents the results of an instrumental analysis of five temporal variables in the spontaneous speech output of eighteen subjects with unilateral cerebral lesion. Subjects with left hemisphere lesion are shown to have a lower value for RTATL, articulatory rate, speech rate and a higher number of pauses than subjects with right hemisphere lesion.

1. INTRODUCTION

The majority of previous investigations of this topic (see [1] for a brief review) use a phonetically imprecise definition of temporal variables to describe speech output produce by brain damaged patients. Speech rate, for example, is described as the number of words per minute rather than on the basis of an articulatory measure such as syllables per second. Other temporal variables such as articulatory rate, or number, duration and distribution of pauses are generally not taken into account. Among recent studies of temporal variables in brain-damaged subjects, Deloche, Jean-Louis et Seron [2] studied speech rate, speech percentage and pauses in five Francophone subjects with left hemisphere lesion. Their results indicate that these subjects produce a similar articulatory rate and a similar number of pauses as compared to control subjects. On the other hand brain-damaged subjects produce a greater average pause duration than control subjects.

Klatt [4] studied the relationship between pause frequency and grammatical category in Anglophone aphasics in a reading task. He observed that Verbs were associated with the highest frequency of occurence of pauses followed by Nouns and then by Adjectives.

The purpose of this study is to compare two groups of subjects with unilateral brain lesion in order to determine whether temporal variables provide for differentiation of these groups according to hemispheric lateralization of lesion. The following temporal variables are examined in this study:

a) RTATL; (see [3] and [4]);
b) articulatory rate;
c) speech rate;
d) the total number of pauses;

e) average pause duration.

2. SUBJECT POPULATION

The patients for this study were eighteen right-handed Francophone adults. Each subject suffered from a confirmed, focal, unilateral cerebral lesion. Eleven patients had suffered cerebro-vascular accidents and seven a cerebral tumor. All were in stable neurological condition at the time of interview.

In terms of lesion localization, subjects A-I (Group A) suffered unilateral right hemisphere lesion and subjects J-R (Group B) suffered unilateral left hemisphere lesion. These two groups can be further sub-divided into three further groups according to intra-hemispheric lesion localization:
A, B, C, J, K, and L, had frontal lobe lesion; D, E, F, M, N and O, had parietal lobe lesion and G, H, I, P, Q and R had temporal lobe lesion.

Subjects with right hemsiphere lesion varied in age from 34 to 56 years old with an average age of 44. Subjects with left hemisphere lesion varied in age from 23 to 71 years old with an average age of 52.3.

None of the subjects had a family history of left-handedness and all were strongly right-handed. All subjects were unilingual Francophones, born, raised and still residing in the greater Paris metropolitain area. They had completed at least primary school education and spoke with a relatively standardized Parisian French pronunciation.

3. SPEECH SAMPLE

The speech sample submitted to instrumental analysis was drawn from the spontaneous speech section of the clinical aphasia examination battery in use at the time at the Salpêtrière and St. Anne Hospitals. This section occurs at the very beginning of the battery, thereby minimizing the effect of fatigue. The patients were replying to questions about their illness, their profession, etc.

For each subject, a speech sample of approximately 300 syllables was analyzed. This speech sample was submitted to two parallel instrumental phonetic analyses of Frequency, Intensity and Duration attributes. The first analysis was carried out with a Pitch Machines digital real-time fundamental frequency analyzer and the second with an RT-1000 digital real-time colour spectrograph.

4. INSTRUMENTAL ANALYSIS

The results given in Figure 1 show that subjects with unilateral left hemisphere lesion show a much lower overall value for RTATL when compared to subjects with right hemisphere lesion.

RTATL is lowest for subjects J (39.55%) and L (37.73%) (left frontal lesion). Other patients with left hemisphere lesion produce a RTATL value of between 60% and 80% which is still significantly lower than that obtained for subjects with right hemisphere lesion. There is then a strong correspondance between interhemispheric lesion lateralization and RTATL.

	RTATL	Articu-latory Rate	Speech rate
A	81.12%	4.22	208.20
B	70.19%	5.95	250.80
C	60.90%	6.18	234.00
D	83.09%	6.02	304.20
E	90.75%	5.86	315.60
F	75.81%	4.24	191.40
G	88.87%	3.84	201.60
H	87.37%	3.88	203.40
I	78.82%	5.62	262.20
Avg	79.65%	5.09	241.26
Std. dev.	9.05	0.95	43.01
J	39.55%	3.59	77.40
K	61.41%	2.99	102.60
L	37.73%	4.20	82.80
M	68.86%	4.63	183.00
N	64.31%	4.07	135.00
O	60.76%	3.72	124.20
P	77.47%	4.46	196.80
Q	81.89%	3.74	163.80
R	67.19%	4.36	184.20
Avg	62.13%	3.97	138.86
Std. dev.	14.17	0.48	42.74

Figure 1. RTATL, articulatory rate and speech rate

	Number of pauses	Average pause duration
A	14.00	101.57cs
B	23.00	81.41cs
C	34.00	75.77cs
D	16.00	56.44cs
E	7.00	62.00cs
F	27.00	75.11cs
G	11.00	77.54cs
H	17.00	62.82cs
I	27.00	48.62cs
Avg	19.55	71.25cs
Std. dev.	8.24	14.85
J	98.00	139.09cs
K	100.00	48.43cs
L	139.00	98.38cs
M	47.00	64.80cs
N	63.00	76.02cs
O	71.00	69.41cs
P	38.00	46.76cs
Q	50.00	41.72cs
R	56.00	56.12cs
Avg	73.55	71.19cs
Std. dev.	30.78	29.11

Figure 2. Number of pauses and average pause duration

T.Test Group A / Group B

RTATL
t = 3.124 (16 d.f.) p = 0.0065

Articulatory rate
t = 3.157 (16 d.f.) p = 0.0061

Speech rate
t = 5.066 (16 d.f.) p = 0.0001

T.Test Group A / Group B

Number of pauses
t =-5.084 (16 d.f.) p = 0.0001

Average pause duration
t = 0.550 (16 d.f.) p = 0.995

Articulatory rate was calculated separately for each accentual group rather than for the whole speech signal sample.

The data in Figure 1 show that subjects with unilateral left hemisphere lesion also produce a significantly lower articulatory rate than subjects with right hemisphere lesion.

The other data given in Figure 1 indicate that subjects with unilateral left hemisphere lesion once more produce a much lower value for this variable as compared with subjects with right hemisphere lesion.

Direct statistical comparison of the averages indicates the existence of a very highly significant difference between the two groups of subjects

Pauses were defined as any period of silence in the speech signal, with the exception of consonantal occlusions.

The results given in Figure 2 show that patients with left hemisphere lesion produce an extremely high number of pauses. In particular, those with left frontal lesion produce many more pauses than all the other subjects: J (98), K (100) and L (139). Direct comparison of the averages shows that there is a highly significant difference between the two groups.

On the other hand, and in sharp contrast to the four other variables studied here, average pause duration does not distinguish the two sets of subjects.

5. CONCLUSION

The various results presented here indicate that subjects with left hemisphere lesion produce speech output which is slower and punctuated by a greater number of pauses than subjects with right hemisphere lesion. The most dramatic differences between the two groups are to be found in speech rate and the number of pauses. This finding is significantly different from that reported in [2]. The data presented here indicate that it is the relative discontinuity of the speech output which is the major differentiating factor between these two sets of subjects.

REFERENCES

[1] BHATT, P. (1990) Variables temporelles chez neuf sujets atteints de lésion unilatérale gauche, In Actes des XVIIIe Journées d'études sur la parole, D. Archambault and R. Descout (Eds.) Montreal: S.F.A. 264-67.

[2] DELOCHE, G., J. JEAN-LOUIS et X. SERON (1979) Study of the temporal variables in the spontaneous speech of five aphasics, Brain and Language, 8, 241-250.

[3] GROSJEAN, F. et A. DESCHAMPS (1972) Analyse des variables temporelles du français spontané, Phonetica, 26, 129-156.

[4] GROSJEAN, F. et A. DESCHAMPS (1973) Analyse des variables temporelles du français spontané, Phonetica, 28, 191-226.

[5] KLATT, H. (1980) Pauses as indicators of cognitive functioning in aphasia, In Temporal variables in speech, H. Dechert et M. Raupach, Dirs., The Hague: Mouton, 113-120.

ACOUSTICAL CHARACTERIZATION OF A PALATINE PLATE

F. Plante (*), C. Berger-Vachon (**),
I. Kauffmann (*), L. Collet (*).

(*) Hôpital E. HERRIOT, Lyon, France
(**) Université Lyon I, Villeurbanne, France

ABSTRACT

In this communication, the authors perform a first analysis of the effects of a palatine plate on the phonation. The first four formants of the french vowels /a/, /i/, /u/, /ə/ have been studied.
The acoustical differences between the two configurations are pointed out.

1. INTRODUCTION

During the last few years, the phonatory pathology have been heavily studied. A good cooperation between scientific and medical teams is necessary to achieve efficient investigations. Under this assumption, a voice processing laboratory had been implanted in the otorhinolaryngology department of the Edouard Herriot hospital of Lyon, and in one place all the necessary skills to study these pathologies are met.
Several ways can be taken to perform these investigations [2,3]. The analysis of the acoustic wave had been chosen, as it is harmless for the patient.
The voice pathology is a very wide topic [5] and complex mecanisms are involved. The speech processing method described in this paper had been tested in a very simple situation in order to assess its efficiency : the presence or the absence of a palatine plate in a patient suffering of a palate agenesia should modify the phonation. This acoustical study had been carried out for its connection with velar incompetency. The patient who collaborated with this study is a 54-year old woman. Differences between the two cases (with or without the plate) come only from the palatine plate and are free of the patients' diversity.
The analysis of the influence of the pathology can be done with two techniques :
- the discrimination : the separation of a pathological voice from a normal voice is studied,
- the characterization : parameters quantifying normal and pathological voices are evaluated.
These two means do not imply absolutely the same tools. In a preceeding work [4] we have seen that cepstral coefficients led to the best classification results, but they cannot be easily interpretated when phonatory mecanisms are involved. This present paper is aimed to point out the influence of the palatine plate on some classical parameters of the voice. The first four formants of the voice had been studied for the cardinal and the neuter (the "schwa") french vowels.

2. MATERIAL AND METHODS

To avoid coarticulation problems, vowels had been spoken in a standard context; the sentence is "C'est *x* ça" (this is *x* that), where *x* is the vowel. The patient spoke forty-three times the sentence with the plate, and also forty-three other times without the plate. The recording was made in an anechoïc room.
The signal coming from the tape was then filtered (10 KHz) and sampled (20 KHz) on a data acquisition card. Samples were kept on the hard disk of a micro-computer for futher analysis. The I.L.S software (Interactive Laboratory System) had been used. Firstly , vowels had been marked in the signal and a sliding linear prediction analysis (LPC) was performed; the LPC was made on a 10 ms window, moved by 4 ms steps over the vowel (120 ms). Consequently 28 analyses were obtained and each analysis led to a set of LPC coefficients [1]. The average of the 28 sets gave the final representation of the utterance.
A fundamental question is raised by the way of processing the data : is a shift due to the formant values (coming from the LPC) or is it due to the speech variability ? The problem is even more harder when pathological voices are considered. In this work the utterances which led to abnormal coefficients had been removed out of the study. We have labelled "abnormal" the utterances leading to detached values far from the average; distributions had been plotted for this purpose. Eventually four utterances were removed for the /a/ and the /i/, eight for the /u/ and thirteen for the /ə/.

3. RESULTS

Table 1 shows, for the four vowels, the average of the formant values. In each cell of the table the first figure is for the utterance without the plate, and the second (in italic) is with the plate. The asterisk (*) indicates that the difference between the two figures is not significant.

Table 1 : Average values of formants

	F1	F2	F3	F4
/a/	832 *571*	1647 *1445*	2776 *2332*	4158 *4057*
/i/	*279 **305*	*1621 **1584*	2613 *2859*	3778 *4137*
/u/	319 *282*	*1229 **1229*	2616 *2513*	*4069 **4007*
/ə/	567 *435*	2025 *1884*	2990 *2647*	*4251 **4182*

4. DISCUSSION

Only /a/ leads to a significant difference for the four formants. This result agrees with a preceeding study which stated that the /a/ gives the best discrimination [4].
The behaviour of the vowel /i/ is unlike the three other vowels. The value of the fomants are lower without the plate and the biggest difference occurs with the third (F3) and the fourth (F4) formants. For the other vowels /a/, /u/ and /ə/ the third, mostly, and the first formants are also modified.
It must be kept in mind that none of the two situations is normal; when the patient wears her palatine plate, the nasal and buccal cavities are parallel leading to a nasalisation with formants

and antiformants. Figure 1 shows a representation of the vowels in the (F1,F2) plane. The utterances with the plate lead to formants close to a normal nasality (300-500,1000 Hz).

4.1 Vowels /a/, /u/ and /ə/

The decrease of the first formant is more important for the /a/ (261 Hz) and the /ə/ (132 Hz) than for the /u/ (26 Hz). This drop can be attached to the nasal formant, N1 (250-300 Hz), and there is a confusion between F1 and N1. The same situation occurs for the /u/; the resulting peak is still in the same range. On the contrary, for the /ə/ and the /a/ the merging of N1 and F1 lowers strongly the position of the peak.

The drop of the second formant, can come from the antiformant lying in the 2200 Hz range. This drop is lighter for the /u/ than for the /a/ (202 Hz) and for the /ə/ (141 Hz) probably because the second formant is not in this range.

The third formant indicates, for the vowels, the rounding of the lips; dramatic drops can be underlined for the /a/ (440 Hz), and for the /ə/ (330 Hz). The diminution wich was observed is, so far, difficult to explain. This drop could be linked to the opening of the vowels (as it is more important for the /a/).

Generally the fourth formant is said to be "speaker dependant". The drop is significant only for the /a/ (101 Hz).

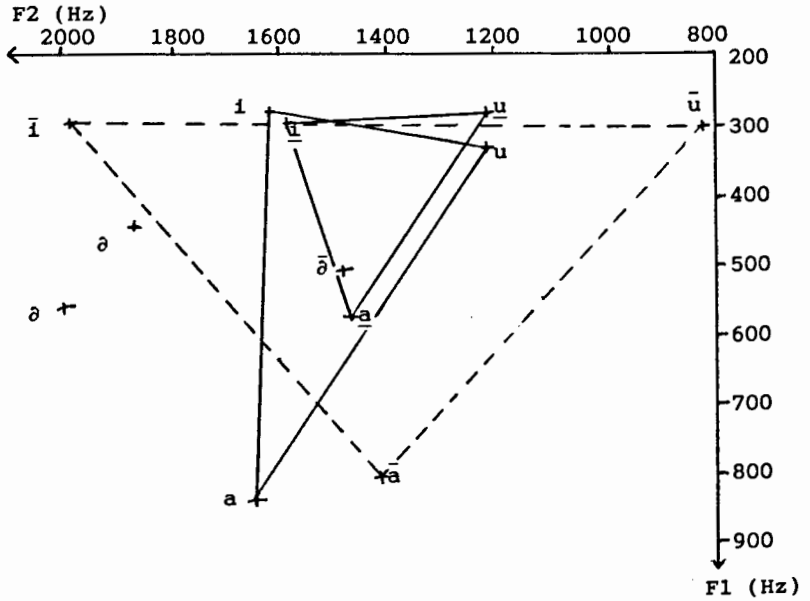


Figure 1 : Schematic vowels representation in the (F1,F2) plane. Vowels spoken with the plate are underlined; vowels overlined are for normal phonation. The third case is without the plate.

4.2 Vowel /i/

The behaviour of the /i/ is much peculiar. The first and the second formants are not significantly modified. The first nasal formant is in the range of the first buccal formant; therefore their merging does no lead to a significant modification of F1. The 2200 Hz antiformant does not have a great influence on the value of F2. It can be pointed out that the second formant of the /i/ is not as high as usual (normal values are about 2000 Hz). This /i/ has the values of a normal /Y/.

The third formant is pulled up (246 Hz) under the influence of the 2200 Hz antiformant.

5. CONCLUSION

The presence or the absence of the palatine plate do not have the same influence over all the vowels formants, but the two situations can be acoustically distinguished. Some more work is needed to provide a more accurate analytical description of the influence on the spectrum of the mecanisms involved in the phonation with a palatine plate : higher formants can be considered and the bandpass of the formants have their own information too.

The best discrimination, on an acoustical ground, can be performed with the utterance of the vowel /a/; this is worthy to be more deeply studied, when velar incompetency will come under scrutiny.

6. REFERENCES

[1] ATAL, B.S. HANAUER, S.L. (1971), *"Speech Analysis and Synthesis by Linear Prediction of the Speech Wave"*, JASA, 50, 637-655.

[2] CALLIOPE (1989), *"La parole et son traitement automatique"*, Masson.

[3] HIRANO, M. (1991), *"Objective Evaluation of the Human Voice : Clinical Applications"*, Congress "New ways of the Voice", Besançon, (14,16 February).

[4] PLANTE, F. BERGER-VACHON, C. GACHE, N. (1991), *"Choix d'un espace de representation pour la discrimination d'une plaque palatine"*, submitted for publication.

[5] Société française d'ORL (1984), *"La voix humaine et ses troubles"*, Arnette Publisher .

AN ELECTROPALATOGRAPHIC STUDY OF SIBILANTS PRODUCED BY HEARING AND HEARING IMPAIRED SPEAKERS

Nancy S. McGarr, Lawrence J. Raphael, H. Betty Kollia, Houri K. Kaloustian and Katherine S. Harris

Haskins Laboratories, New Haven, CT, U.S.A.

ABSTRACT

This experiment examined the production of the sibilants /s/ and /ʃ/ followed by the vowels /i/ and /u/ produced by four normally hearing and four hearing-impaired college students who each wore a Rion semi-flexible palate. Electropalatographic and acoustic data were collected simultaneously. Analyses show production pattern differences between hearing and hearing-impaired talkers.

1. INTRODUCTION

The use of palatography as a measurement technique in speech production studies has a long history in experimental phonetics [1,2,4]. Recently, a semi-flexible palate [4] has been developed in a variety of sizes for use with morphologically normal adults and children. Research [5] indicates that these semi-flexible palates can be fitted to many subjects and that high correlations can be obtained in repeated measure comparisons of custom-fitted and semi-flexible palates thus permitting studies of a broader population that might previously have been examined with custom-fitted prostheses.

We have suggested in our on-going work [3] that speech produced by hearing-impaired persons has characteristics in common with that of young normally developing children, namely, that it is slower, more variable, and hence unskilled. Hearing-impaired talkers are often said to articulate on a segment-by-segment basis. Sibilant production is known to be particularly problematic. Although there are a number of competing coarticulatory theories regarding skill development in normal children, there have been essentially no studies of coarticulation in hearing-impaired speakers. This study was therefore conducted to examine sibilant production using electropalatography as an objective index of production.

2. PROCEDURES

2.1 Subjects

The subjects were four normally hearing and four severely-profoundly hearing-impaired college students attending a university in New York. The university provides academic support services for the hearing-impaired students. Two of the test subjects sustained adventitious hearing losses post-lingually due to meningitis; two subjects had congenital hearing losses. All subjects used speech as their primary mode of communication.

2.2. Stimuli and instrumentation

The subjects produced twenty repetitions each of the utterances "see, she, sue, and shoe" while wearing a Rion semi-flexible palate. The palate embedded with 63 electrodes arranged in rows was calibrated against plaster casts of the subject's dental arch to insure correct size and placement.

2.3. Analysis

Audio and electropalatographic data were collected simultaneously on FM tape.The data were analyzed using computer programs at Haskins Laboratories. Palatographic records were analyzed on a frame by frame basis (15.6 msec. per frame) for approximately 600 msec. before and after the onset of voicing for the vowel., essentially the entire utterance. In addition, the records of the 63 electrodes were also grouped in "rows".

While various combinations of electrodes were analyzed, the areas of particular interest for production of the phoneme contrasts were defined from the palatal records as the back side rows and the front side rows.

Acoustic measurements of the subjects' fricative productions were done using discrete Fourier transforms on selected intervals of the acoustic file [6] which corresponded to points of interest in the palatographic records.

2.4. Listener Judgments

The audio recordings of the subjects' productions were presented to panels of listeners unfamiliar with the speech of the hearing impaired for identification in a closed-set response task. Preliminary results for the hearing and three hearing-impaired subjects show that the productions were nearly always identified as intended and relatively high identification scores (greater than 90% correct) were achieved. Analysis of one hearing-impaired person's productions showed somewhat poorer scores (71%) due to consonant confusions. .The /s/ was primarily identified as /ʃ/. This speaker is identified as EW below.

3. RESULTS

Fig. 1 is a plot of a hearing subject (JR) that shows the averaged percent of electrodes contacted over time for each of the four stimuli. The vertical axis is the averaged percent of electrodes contacted; the horizontal axis is the time in frames from -40 to +40 where each frame is 15.6 ms. The 0 frame is the onset of voicing for the vowel. Each of the stimuli were produced over twenty times and the data plotted represent the average. This normally hearing control produces /si/ (solid line) with two well defined peaks corresponding to the /s/ and /i/ respectively; when contrasted to production of /su/ (dashed line), there is only one peak that decreases nearly coincidentally with the onset of voicing of the vowel, in this case /u/. The /ʃ/ contrast shows a greater number of electrodes contacted in the front side rows but the same double vs. single peak pattern associated with vowel differences (dotted versus thin lines respectively). Fig. 2 shows that essentially the same plots can be described for the electrode patterns contacted in the back rows during production of the /si / vs. /su/ and /ʃi/ vs. /ʃu/. These patterns were fairly typical of the normal talkers.

These patterns described for one hearing control are contrasted to those for the least intelligible hearing-impaired speakers Fig. 3 is a plot for the front side rows of electrodes. The pattern for /si/ shows the two peaks associated with consonant and vowel respectively; the plot for /su/ shows a decrease in the number of electrodes contacted for /u/. This is as the normals. However, this speaker does not differentiate /s/ and /ʃ/; the plots are essentially overlapping. In Fig. 4, the plots for the back side rows show a more diffuse undifferentiated pattern across the four stimuli types until about frame +10 (approximately 150 msec) after onset of voicing for the vowel.

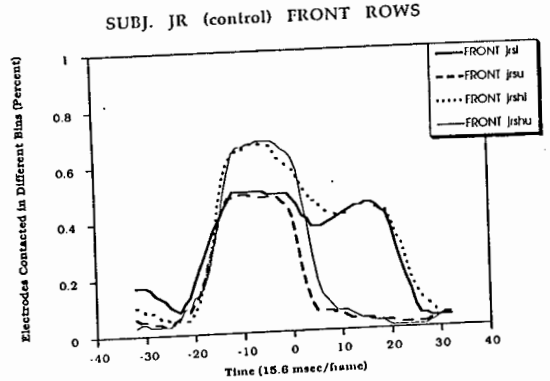


Fig. 1 shows data for a hearing control. See text for explanation.

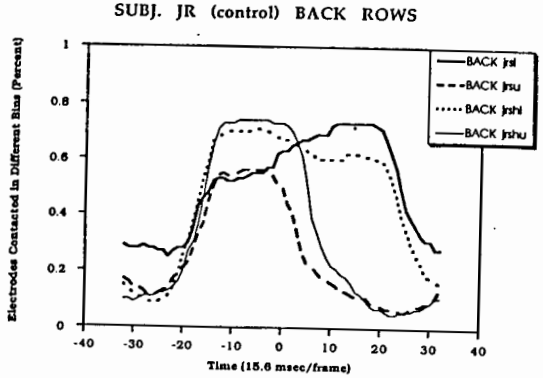


Fig. 2 shows data for a hearing control. See text for explanation.

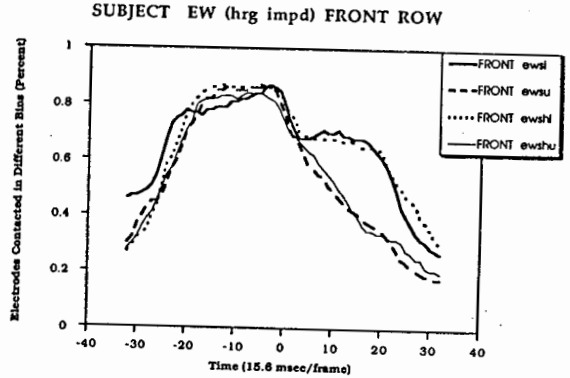


Fig. 3 shows data for a hearing-impaired subject. See text for explanation.

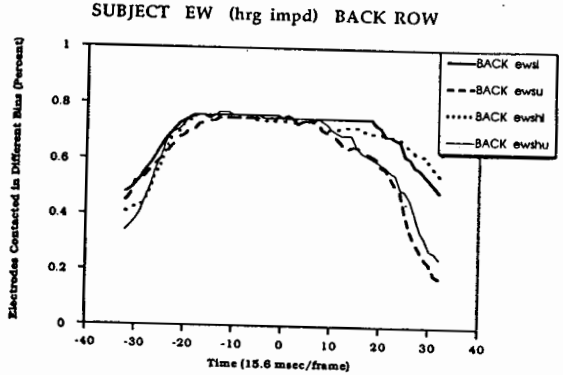


Fig. 4 shows data for a hearing-impaired subject. See text for explanation.

Plots for the other hearing-impaired speakers differed in some instances from normal controls and also from other hearing-impaired talkers. For example, in production of /si/, three of the hearing-impaired subjects produced a single pattern extending over a considerable time frame and did not differentiate the consonant and vowel. In still other instances, e.g. production of /su/, the hearing-impaired subjects were like the normals although the timing of the change in electrodes contacted from the sibilant to the vowel production occurred late relative to the onset of voicing. Additional electropalatographic patterns of contact and the corresponding acoustic measures will be presented in greater detail.

4. DISCUSSION

Electropalatographic contact differed for /s/ and /ʃ/ and for /i/ and /u/. Normal hearing controls had similar patterns of production. Specifically, there was more anterior palatal contact (front side rows) for the vowel /i/ than for /u/. This difference was related to sibilant context, both vowels evidencing less contact following /s/ than /ʃ/. There was more palatal contact for /ʃ/ than /s/. Moreover, there was considerable articulatory movement that occurred over a relatively short time frame around the onset of voicing for the vowel. In contrast, some hearing-impaired speakers achieved relatively correct electropalatal contact for the target phonemes but that "correct" pattern was often extended inappropriately over time.

These data show evidence of coarticulation in speech produced by hearing-impaired persons and do not substantiate the notion of segment-by-segment production. Coarticulatory studies of speech produced by young normally hearing children suggest that development is skilled based and that children should coarticulate less than adults because the phonemic elements of speech are not appropriately blended. Speech would be viewed as slower, more variable and unskilled. Other work [6, 7] suggests that units of speech produced by young children are somewhat larger than the single phoneme size. Because these units are not differentiated, coarticulation across syllable size units is greater in children than in adults, and decreases as skills develop. The data in this study suggest that while hearing-impaired persons coarticulate, the units are not fully differentiated and in that regard, resemble the unskilled productions of young normal children.

5. ACKNOWLEDGEMENTS

This work was supported by NINCDS Grant DC-00121-29 to Haskins Laboratories.

6. REFERENCES

[1] FLETCHER, S., MCCUTCHEON, M.J., & WOLFE, M.B. (1975), "Dynamic palatometry". *Journal of Speech and Hearing Research, 18,* 812-819.

[2] HARDCASTLE, W., JONES, W., KNIGHT, C., TRUDGEON, A., & CALDER, G. (1989), "New developments in electropalatography: A state-of-the-art report", *Clinical Linguistics and Phonetics, 3,* 1-38.

[3] HARRIS K.S., RUBIN-SPITZ, J., & MCGARR, N.S. (1985), "The role of production variability in normal and deviant developing speech". In J. Lauter (Ed.), *ASHA Reports (Proceedings on the Planning and Production of Speech in Normal and Hearing Impaired Individuals: A Seminar in Honor of S. Richard Silverman.* Rockville, MD, ASHA.

[4] HIKI, S., & IMAIZUMI, S. (1974), "Observation of tongue movement by use of dynamic palatography", *Annual Bulletin, Research Institute of Logopedics and Phoniatrics, 8,* 69-74.

[5] MCGARR, N.S., TSUNODA, K., & HARRIS, K.S., (1990), "Palatography: A comparison between custom-made and "flexible" artificial palates for speech production", *Journal of the Acoustical Society of America, 86 (S1),* S115 (A).

[6] MCGOWEN, R. & NITTROUER, S. (1988), "Difference in fricative production between children and adults: Evidence from an acoustic analysis of /ʃ/ and /s/", *Journal of the Acoustical Society of America, 83,* 229-236.

[7] NITTROUER, S., & STUDDERT-KENNEDY, M. (1987), "The role of coarticulatory effects in the perception of friactives by children and adults", *Journal of Speech and Hearing Research, 30,* 319-329.

COMPUTER-ASSISTED RUSSIAN PHONETICS LEARNING

N.V.Bogdanova, I.S.Panova-Jabloshnikova,
S.B.Stepanova

Leningrad State University, USSR

ABSTRACT

This paper dealing with the experience of using computers in teaching Russian phonetics describes a working program of a teaching automatical transcriptor and a number of controlling phonetic programs for students of Russian, foreign teachers of Russian phonetics who wish to raise their qualification and other of users. These programs allow to study theoreical phonetics, reading rules and work out practical transcription skills.

Computer-assisted means of teaching already familiar in foreign language learning (at the Department of Philology of the Leningrad state University such programs have been worked out on two languages and are used in teaching process), can be successfully used in different courses for students studying Russian philology. At the Department of Philology of the Leningrad state University a whole series of such programs has been worked out and is actively used in teaching process. In the first place they are intended for doing exercises which demand a student to have some definite analisys skills and automatism. Such skills often can be acquired only long training and its' control is very boring to a teacher and takes much class time.

So at the course of Russin phonetics original work of students is provided at a computer class where they can do exercises on transcription, syllable division and phoneme description in terms of differential features. Working out practical transcription skills takes particularly much time. The teaching automatic transcriptor (AT) created at the Leningrad state University is a kind of the universal AT which had been worked out for the Computer Russian language Fund as a part (or one of structural blocks) of the Phonetic Fund. This one is able to translate any orthographic text into a sequence of transcription signs with different degree of detailazation - phoneme, phonetic and detailed phonetic which is used in synthesized speech creating. The teaching AT has other aims - teaching and control of transcription skills. The AT program includes 3 series of exercises: phoneme transcription, phonetic transcription and a multivariant phonematic transcription taking into account different modern phonological conceptions.

The phoneme part of the teaching AT includes 12 exercises located according to increasing of their complexity. Each exercise introduces concrete phonetic rule: pronunciation of jotted and unjotted vowels in the stressed and non-stressed position, assimilation of consonants in different features, pronunciation of different consotant sequences including dubble and "non-pronounced" . In all cases a word displayed on the screen corresponds to its' transcription presentation, it means that the rules are formulated "from letter to sound" or to be more correct - to its' transcription designation . Separately taken exercise contains words-exceptions pronunciation of which should be learned by heart (что, конечно, мягкий etc.). The last exercise in this part contains trnscription of combinations of words and a summary of all considered phonetic features in Russian speech. This part of the teaching AT has rather wide sphere of application. Besides its ability to control the acquiring of phoneme transcription skills within the framework of Sheherba's phonological conception it can carry out other functions. Thus, the series of exercises on phoneme transcription can serve as a peculiar testing of non-Russian students' knowledge of Russian pronunciation. It is known that getting information about the level of student's training at the beginning of studies helps teacher to choose methodics and lexic material and plan the teaching in a way to reach the best results in this given audience. Phoneme transcription based on Shcherba's phonology fully can serve as this test as it reveals a real pronunciation.

Another sphere of using the phoneme part of the teaching AT is teaching pronunciation people with difficulty of hearing. Generally such a methodics may be applied to the people who lost their hearing at a rather early age. It is known that if inborn hearing skills are lost before 12 years of age many of pronunciation skills concerning phonetic active laws of the Russian language - opposition of voiced and voiceless consonsnts, difference between vowels with different degrees of reduction etc.- are not formed yet. When there is no natural way: perception of sound - production of sound - people with difficulty of hearing have the only way of gaining orthoepic skills left: orthographic word image - its' sound image. The intermediate link in last case is the transcription notion of sound word image which is provided by the AT. In this sense methodics of the speech training of foreigners and people with bad hearing is very monotype.

Computer-assisted teaching and the AT programs are supposed to be organized in the following way. A student or a patient is communicating with computer in a computer class. Words the transcription of which should be typed on the keyboard are displayed on the screen according to the booted program. The user types transcription according to his own idea of the word sounding. Pronunciation mistakes which are characteristic to the student and

are the results of wrong or simply specific Russian speech perception will be reflected in the transcription. In the case of a wrong answer "?" is displayed on the screen in the place of the mistake and the program suggests user to transcribe this word once more. If it is necessary the student can make an inquiry to the computer memory and get a rule corresponding to this case. On the stage of training in the Russian Literary (standard) pronunciation this rule can be considered as orthoepic rule. On the advanced stage of teaching students can get knowledge of Russian orthoepic standard variancy. At last if the right answer is not received the program suggests the right answer and goes to next word. Words specially selected for exercises and step by step passing from simple rules to more complicated ones should help people with difficulty of hearing and patients with temporary violation of hearing and speech to eliminate mistakes in pronunciation and acquire right orthoepic skills.

Next part of the teaching AT - phonetic - consideres all main laws of Russian speech: the law of quality and quantity vowel reduction, mutual influence of sounds in speech flood, positional distribution of Russian phoneme allophones. This part of the teaching AT contains 17 exercises which are located according to their increasing complexity. Each new exercise introduces next phonetic rule and suggests a new transcription sign. All these rules are stored in computer memory and can be displayed on the screen at the user's inquiry on any stage of work with the program. The inquiry part of the AT also contains information on the Russian language phoneme system, its' structure, differential features and their correspondence within the framework of the system. Any part of this information can be demanded by the user either before work - in this case doing exercises is ablolutely original without teacher's help - or in course of work if some difficulties appear. This part is supposed to have more limited sphere of application - it is intended for philologists who study Russian phonetics. However it is acquiring phonetic transcription skills that takes particularly long training; but replacing auditory classes with work at the computer-equipted class saves much teaching time.

The third part of the AT is even more specific. It contains exercises on phoneme transcription within the framework of different phonological conceptions. According to the Russian phonetics teaching curriculum students of the Department of Philology get acquainted with the point of view of Moscow Phonological School supporters and the theory of weak and strong pnonemes by R.I.Avanesov. Serious knowledge of these phonological statements expect students to be able to give a phoneme structure of any word from the point of view of both these conceptions. In this part of the AT each word displayed on the screen should have 5 corresponding transcriptions: phonetic, Shcherba's, Moscow School phoneme and 2 transcriptions by Avanesov - word-phoneme and morph-phonetic. Such skills also take much teaching time to be gained. When using the AT teacher only has to prepare student for this work and control its' fulfillment which is fixed in computer listing.

The program of the AT, as it is clear from the previous part of the report, is organized according to the pattern when student enters the needed form of the given word. The exercises on syllable division in the Russian language have the same structure. In this case practical work also improves the knowledge gained at lectures. Students acquaint themselves with all existing theories on syllable division and use them in practice. As in all previous cases computer checks the correctness of the answers, appreciates the results and memorizes mistakes.

Checking theoretical knowledge of Russian phonetics also takes much of teacher's efforts and time, it is realised with the help of a series of phonetic exercises. The main attention is paid here to the phoneme system of the Rusian language, its differential features and phoneme modification in speech flood. This program is based on the "menu" principle when the user is to choose the right answer among several suggested variants. The student is only to press one key to answer the question on Russian phoneme differential features, their difference and similarity. The student of the Department of Philology T.Taalmann took part in working out the recent phonetic programs.

All considered programs are intended for the students of Russian (including foreigners) and for those who perfect their Russian. They can be of great help for foreign teachers of Russian who teach Russian phonetics. The programs are open, can be added and work under all operational systems.

REALISATION OF EXPRESSIVE FUNCTION OF INTONATION IN DIFFERENT NATIONAL VARIETIES OF ENGLISH

Paul Kremel

State Teacher's Training University,
Moscow, USSR

I.INTRODUCTION

I.I.General Considerations on the Problems of Stability and Variability of English Intonation System.

Opposition "stability-variability" is one of the most essential in the study of the movement of any particular language in space and time.Although a comparatevely large number of studies have been dedicated to this problem,some of its aspects remain,however,unclear.

I.2.Territorial Variability of Intonation System.

Constant and variable features of the prosodic system of English intonation in its different national varieties have been, in large part,looked at and investigated at the phrasal level,which,understandably,could not have resulted in giving us a full,clear picture of national,territorial variability of intonation.Moreover, some national intonational varieties(e.g.Canadian) have not been adequately described even at the level of phrasal units,though,as many authors point out(2,7, 9),it is the pecularities in intonation that make CanEng different from AmE.

Therefore,the experimental study that we undertook deals with some constant and variable features of English,Scottish,American and Canadian national varieties of English intonation system at the t e x t u a l level.

I.3.Style

Samples of national intonational varieties were represented by short monologue texts,realised in two kinds of speech activity:reading and speaking.Stylistically, all texts represent informational formal style.Emotional neutrality of this style serves as a reliable basis for singling out even slightly marked emotional and attitudinal meanings and provides safe ground for further comparisons.

I.4.Emotional-attitudinal and Intellectual Levels.

The largest number of variable features is recorded at the emotional level.At the so-called intellectual level we seem to observe a greater degree of stability.Statistically,it seems quite possible,therefore,to approximately define and compare the amount of both, emotional-attitudinal and intellectual information in the overall semantic structure of texts,realised by native speakers of the above-mentioned national varieties of English.It seems also possible tofind out which prosodic features

are responsible for conveying certain attitudinal meanings.

2.EXPRESSIVE(EMOTIONAL-ATTITUDINAL) FUNCTION OF INTONATION

2.I.Linguistic Status

The question of linguistic status of this function of intonation has been traditionally a hot ly debated one.Some authors distingwished between the communicative and emotional aspects(IO),associating the latter with modality,or between the logical and attitudinal(emotional)functions of intonation(I).On the other hand,some linguists consistently excluded the emotional(attitudinal)function of intonation from the scope of linguistic analysis(4).

2.2.Reasons for Different Approaches to Linguistic Status of Attitudinal Function of Intonation.

a).dichotomy:attitudinal (emotional)function - grammatical function;

b).structural approach(2, 3).Some advocates of this approach,however,do not deny the fact that intonation can perform attitudinal(modal)function,but consciously exclude it from the sc scope of their study(4).

c).opposition:"intellectual" - "emotional" information,frequently used in the theory of communication and in psycholinguistics(5).

Obviously,such diohotomies and oppositions are rather artificial and a little too far-fetched, since one can hardly find enough evidence for "absolutely neutral"contours in most sentences in speech (8).To overcome the above-mentioned scholastic dichotomies,D.Crystal suggested the use of scales with minimum and maximum values of attitudinal function at the extremities.Such scales seem to give a much more adequate picture of the real process of expressing various attitudinal meanings in speech.Dichotomies should be seen only as a means of analysis.They are obviously very relative, since in the course of the communication intellectual and emotional information is conveyed simultaneously. Besides,in terms of psychology,there are no clear-cut boundaries between logical and emotional spheres, which,on the one hand,results in the fact that certain expressive,emotional, attitudinal elements penetrate the logical structure of an utterance,and on the other hand,explains why attitudes happen to be basic elements in the individual's cognitive activity. Hence,there seems to be a great deal of overlap and interaction between the intellectual and expressive (emotional,attitudinal) levels.

2.3.Intergrated Approach.

Analysing the above-mentioned approaches,Ladd et al. (6) quite rightly point out that the aim of the first approach(approachA) is to "produce quantifiable descriptions of both the medium(the non-segmental part of the speech signal) and the message(the various types of affective information conveyed),and to attempt to state correlations between the two". Within the framework of the second approach(approach B) "acoustic measurements should not be correlated

directly with attitude judgements,but should be taken as evidence about the phonetic correlates of the intonational contrasts posited in the linguistic description".
They also note that although "shortcomings and methodological difficulties in both approaches to "intonation and attitude" are quite apparent even to enthusiastic proponents of one view or the other,... insofar we find evidence both for categorical distinctions in intonation and for the direct expression of attitude."
Ladd et al.(6)nonetheless conclude that an approach that carefully distinguishes intonation from paralinguistic cues and designs its studies with that distinction in mind will be the most productive way to investigate the role of intonation in expressing attitude".

3.CONCLUSION

3.1.Approach

In the present study we primarily adopted the first approach(approach A) with the above-mentioned distinction in mind).

3.2.Parameters

For convenience,we grouped the attitudinal,emotional, modal meanings,studied in the course of the comparative analysis into five scales:
-certain - uncertain;
-surprised - not surprised;
-approvingly-disapprovingly;
-friendly - not friendly;
-concerned - uninvolved.

Secondly,the task was to determine which audible prosodic characteristics correalate with the above-mentioned modal meanings. Therefore,the following measurements were taken: delimitation,placement of communicative centres(logical and emphatic),contours(type of head,pitch and range),type of nuclear tone(its final section, angle,pitch and range characteristics,in particular) loudness,tempo,rhythm, timbre colouring(esp.on the emphatic segments of the texts).
Thirdly,comparative analysis of national intonational varieties of English is carried out(in two kinds of speech activity: spontaneous reading and speaking,both male and female versions)

3.3.Results.

a).Constant features: the given national varieties of English exibit a great deal of stability primarily in expressing more "intellectual",logical information.
b).Variable features: the given national varieties of English reveal a certain amount of variability of intonation system in expressing attitudinal,emotional, modal meanings(see supplement).

4.REFERENCES

I. Antipova A.M. Sistema angliskoy retchevoy intonatsii.-Moscow,Vishaya Shkola,1979.

2.Bronstein A.J.The Pronunciation of American English.-New York.

3.Brown G.,Currie K.L., Kenworthy J. Questions of Intonation.-London,1980.

4.Halliday M.A.K. Intonation and Grammar in British English.-The Hague: Mouton,1967.

5.Jones D. An Outline of English Phonetics.-Cambridge.-Cambr.Univ. Press,1956.

6.Ladd D.R.,Scherer K.R., Silverman K. An Intergrated Approach to Studying Intonation and Attitude// John - Lewis,C."Intonation in Discourse",-London 1986.

7.Pyles T. The Origins and Development of the English Language.-New York 1971.

8.Uldall,E.T. Attitudinal Meanings Conveyed by Intonation.//Lg.Sp.3,1960.

9.Wells J.C. Accents of English.vol.2.-Cambridge, 1982.

10.Zinder L.R. Obschaya Phonetica.-Moscow,Vishaya Shkola,1979.

UNE NOUVELLE CONCEPTION DE LA PLACE DE LA PHONETIQUE EN DIDACTIQUE DES LANGUES

Abdelazim YOUSSIF et Elisabeth LHOTE

Laboratoire de Phonétique - Besançon - France

ABSTRACT

The authors propose a new approach in oral learning of a foreign language with pragmatic and phonetic basis.
The natural situation of communication is integrated to learning process and the three dimensions of oral communication are not dissociated : Production, Preception and Comprehension.

L'APPRENTISSAGE D'UNE LANGUE ÉTRANGERE

L'apprentissage d'une langue étrangère prend en compte l'ensemble du processus oral de la communication.
Pour qu'un échange verbal soit réussi, un certain nombre de règles sont à respecter, en particulier celles qui concernent les deux pôles de la communication, l'émission et la réception, c'est-à-dire le locuteur et l'auditeur. Le locuteur se caractérise déjà par sa voix qui facilite ou non l'échange ; il a un passé avec lequel l'auditeur a - ou n'a pas - de points communs à partager ; il a une expérience et des connaissances qui se révèlent dans sa conversation. Quant à son message proprement dit, il peut être influencé par l'attente de l'auditeur par ses réactions, ses connaissances, son opinion, ses croyances ainsi que par un accord ou un désaccord entre les deux partenaires. Dans toute communication fonctionne un processus d'interaction. Et la communication orale n'y échappe pas.

L'ORAL D'UNE LANGUE

L'oral d'une langue se définit par l'utilisation d'un code phonétique et phonologique mais aussi par des habitudes langagières qui varient avec la langue et la culture.
On peut se contenter de se référer aux deux niveaux d'analyse, segmental et suprasegmental,qui servent habituellement de repères en phonétique. A condition d'appliquer ce modèle d'analyse à toutes les phases du processus, c'est-à-dire à la production, au message lui-même, à la perception et à la compréhension par un auditeur donné.
Dans l'enseignement et l'apprentissage d'une langue, l'attention est portée sur le code de la langue-cible mais pas sur le passage d'un code à un autre qui ne peut pas ne pas poser de problèmes d'interférences.
Une approche phonétique de l'oral doit prendre en compte et le changement de code et les problèmes liés à ce changement dans l'ensemble de la communication. Ceci signifie que l'approche **contrastive**, encore beaucoup utilisée ne peut à elle seule servir de cadre d'approche.

L'ACTIVITE DE LANGAGE

L'apprentissage d'une langue étrangère se fait en général par référence à une autre langue acquise antérieurement - exception faite du bilinguisme précoce - Quand l'apprenant adulte a besoin de communiquer dans la langue qu'il apprend, il utilise sans y penser les habitudes langagières qu'il a développées

en L1 : dès le plus jeune âge, il a, comme quiconque, appris la variabilité des réalisations phonétiques en situation naturelle ; il a appris à écouter sa langue en prenant en compte tous ces changements. L'activité de langage ainsi développée en L1 est un potentiel très riche : tout individu apprend à utiliser un code dans la variabilité et il construit son écoute et son langage en découvrant (implicitement) un code en s'adaptant à tout moment à des individus, à des situations, à des perturbations de la communication.
Cette approche de l'oral dans la langue maternelle mérite certainement d'être exploitée plus tard car elle est typique de la façon dont l'humain se construit tout en s'adaptant.

L'EVOLUTION DES METHODES D'APPRENTISSAGE

L'évolution des méthodes d'apprentissage depuis 15 ans a redonné sa place à la communication. Ce courant est allé de pair avec le développement de la communication internationale : la situation d'échange, pour réussir un marché par exemple, rend la communication orale indispensable.
On a ainsi vu se développer des approches dites communicatives qui se sont centrées sur l'apprenant. L'objectif n'est plus un objectif d'enseignement, mais d'apprentissage.
De nombreuses méthodes communicatives ont vu le jour. Le concept-clé est devenu l'Acte de Parole (Searle, 1972 et Hymes, 1972) et a servi d'unité didactique de base (Dominique et col. 1982)

LA PRISE EN CHARGE PHONETIQUE EST TOUTEFOIS RESTEE EN RETRAIT

On constate, qu'alors que l'Acte de Parole devient l'étendard des approches communicatives et des didacticiens, la discipline habituellement la plus apte à étudier la parole, la phonétique, n'a pas été saisie par la mutation.Ceci se manifeste par un maintien de l'approche traditionnelle contrastive qui s'exprime encore actuellement dans les manuels les plus récents (Kaneman-Pougatch et Pedoya-Guimbretière, 1989 - Pagniez-Delbart , 1990). On ne conçoit pas en effet d'aller d'une langue L1 à L2 sans passer du simple au complexe de la langue. Pour être plus précis, on peut donner comme exemple la progression phonétique utilisée par Kaneman-Pougatch et Pedoya-Guimbretière qui commence par les oppositions [i-y-u] puis [-e-]
Cette soi-disant simplicité n'est qu'une apparence. La réalité montre que pour des arabophones, des hispanophones et des anglophones, ces oppositions sont une réelle difficulté qui résiste longtemps à l'apprentissage

QUELLE EST LA PLACE DE LA PHONETIQUE?

La phonétique définie comme l'étude des sons du langage a laissé une distance entre phonéticien et didacticien, une distance de nature disciplinaire qui a eu pour conséquence de réduire son efficacité dans la didactique des langues étrangères. De la transcription phonétique, pour éviter la graphie, à la méthode verbo-tonale de correction phonétique, on l'a affectée essentiellement à la production de la parole. Dans les méthodes d'apprentissage ou dans les manuels qui les accompagnent, on cherche à décrire et à enseigner le système de la langue au moyen d'exercices dits de répétition, de discrimination et d'identification. Ces activités ne sont pas des objectifs en tant que tels : elles visent à fixer des acquis linguistiques.
Quant à la correction phonétique de la prononciation qui s'appuie sur des contextes dits facilitants, nous constatons qu'elle ne parvient pas toujours à résoudre les difficultés de perception et de production. De plus elle laisse de côté les problèmes phonétiques liés à la compréhension orale. Une des causes majeures de cette inadaptation est due au fait que le travail porte sur des modèles de communication peu authentiques et peu naturels.

UNE APPROCHE PHONETICO-PRAGMATIQUE

La démarche phonétique que nous proposons s'inscrit dans le courant actuel des études didactiques et plus particulièrement dans la perspective d'une centration sur l'apprenant. Celui-ci, porteur d'une culture, d'habitudes langagières et communicatives, devient à

la fois le **point de départ** et l'**objectif principal** de la formation. D'un point de vue phonétique, ceci signifie qu'on ne donne plus la priorité à la langue et que la **progression** d'enseignement et d'apprentissage **ne se construit pas** - ou pas seulement en tout cas - **sur une progression allant du simple au complexe de la langue**, mais sur la construction personnelle d'une **nouvelle structuration perceptive** et d'une nouvelle façon de comprendre.
En d'autres termes, on peut dire que l'on va aider l'apprenant à **changer d'habitude d'écoute**. Ceci doit l'aider à capter autrement l'information sonore parmi les multiples réalisations individuelles en L2 et au sein de situations de communication variées.
On ne dissocie plus la perception phonétique de l'acte de langage. L'approche phonétique dans ce cas **s'intègre** à l'ensemble du processus d'apprentissage. On n'apprend plus seulement à **parler**, mais aussi et surtout à **entendre** et à **comprendre** dans une autre langue que la sienne.

PRODUCTION - PERCEPTION ET COMPREHENSION

Les difficultés de production sont souvent étroitement corrélées à des difficultés de perception et de compréhension.
Un exemple permet d'illustrer ce point : soit la situation suivante entre une Française "J" et une Soudanaise "A"
"J" - Tu vois! là-bas c'est le bateau qui va en Corse.
"A" - Oh là là ! mais c'est grand!
"J" - Oui, mais il y des étages là-dedans.
"A" - Des otages! pourquoi?
La mauvaise discrimination perceptuelle entre [e] et [o] a conduit l'auditrice soudanaise à faire une mauvaise hypothèse linguistique qui s'explique d'autant mieux quand on sait qu'elle était entrain de discuter d'un événement politique. Cette hypothèse erronée est liée à des attentes de type pragmatique. Le passage d'un niveau perceptuel à un niveau pragmatique explique ici une erreur de compréhension.
Une analyse plus approfondie de cet exemple corrélé à d'autres du même type met en évidence le fait général suivant : le système vocalique arabe réduit aux trois pôles vocaliques /a/-/i/-/u/ ne prédispose pas un arabophone à percevoir de faibles variations de timbre. Cela entraîne par exemple une confusion quasi systématique entre "il" et "elle" à la fois dans la perception et dans la production des arabophones. Cette faiblesse dans la discrimination perceptuelle des voyelles françaises est souvent à l'origine de mauvaises compréhensions entre francophone et arabophones.
Cet exemple illustre assez bien la corrélation permanente nécessaire qui est faite entre les trois pôles phonétiques de la production, de la perception et de la compréhension orales (Lhote, 1990).

LES PRINCIPES DE LA DEMARCHE NOUVELLE

On peut esquisser les principes directeurs de cette démarche :

Premier principe : toute forme d'oral en situation naturelle est bonne pour l'apprentissage.
Le premier objectif à respecter est ici l'apprentissage de relations perceptuelles stables quelle que soit la variabilité individuelle ou situationnelle.

Deuxième principe : l'étude d'un document oral doit partir de la situation dans laquelle il est produit. Ceci suppose l'analyse des composantes pragmatiques qui permettent de comprendre, ceci dès le début de l'apprentissage.

Troisième principe : l'analyse de problèmes d'interférences fait appel à l'analyse contrastive.
Le sens de la démarche est ici important : on n'applique pas l'analyse contrastive à l'enseignement. On se sert de l'outil d'analyse afin de mieux expliquer les structures sous-jacentes et d'orienter la progression.

Quatrième principe : la progression d'enseignement est soumise à la progression d'apprentissage.
Selon la langue d'origine, les difficultés et les interférences ne sont pas les mêmes. Ceci est connu de tous les didacticiens. Les priorités sont donc à définir en fonction de la langue d'origine, des besoins de communication de l'apprenant et enfin des conditions de l'apprentissage.

En conclusion, on peut dire que l'approche phonético-pragmatique que nous proposons n'est ni une théorie, ni une méthodologie nouvelle. Ce qui est beaucoup plus important, c'est une nouvelle façon de considérer l'acte de langage oral.

BIBLIOGRAPHIE

[1] P. DOMINIQUE, J. GIRARDET, M. VERDELHAN, M. VERDELHAN,
- Sans Frontière, 1982
- Le Nouveau Sans Frontière, 1988.
CLE. International

[2] D. HYMES, 1972, On communicative competence. In J. B. Pride and J. Holmes (Eds). Sociolinguistics : Selected Readings, p. 269-293, Harmonds worth, England : Penguin.

[3] M. KANEMAN-POUGATCH et E. PEDOYA-GUIMBRETIERE, 1989, Plaisir des sons, Alliance Française, Hatier - Didier.

[4] E. LHOTE (Ed), 1990, Le paysage sonore d'une langue, le français, Ouvrage collectif. Chapitres 1,2 et 3, Helmut Buske Verlag Hamburg, p. 1-77.

[5] T. PAGNIEZ-DELBART, 1990, A l'écoute des sons, les voyelles, CLE International.

[6] S.R. SEARLE, 1972, Les actes de langage, Hermann, Paris.

PRODUCTION OF STOP CONSONANTS : NEUROLINGUISTIC STUDY IN A CONDUCTION APHASIC

I. Clavier-Pinek B. Pinek J.L. Nespoulous

Laboratoire Jacques Lordat, Université Toulouse Le Mirail
France.

ABSTRACT

The stop consonant transition durations of a french-speaking conduction aphasic and a control subject, have been compared in a word repetition task with and without masking noise. Contrary to the control, the voiced and unvoiced transition durations of the aphasic tended to be poorly differentiated, and in noise, when auditory feed-back was precluded, all his transitions became shorter. The error rates of the aphasic were also analysed. It was much higher for unvoiced stops than for voiced stops. It was higher in masking noise than without noise. The results are discussed with respect to the possible roles of auditory and kinesthetic control of speech production.

1. INTRODUCTION

It is clinically widely accepted that conduction aphasia is associated with posterior cortical damage centered on the parietal lobe and with linguistic deficits that are focused on production mechanisms. The impairments are especially important in word repetition but also in spontaneous speech. Segmental errors and substitutions are very common. On this basis, it has been proposed that a phonological level of speech production is deficient [4]. A few VOT studies have also shown some phonetic disturbances [1, 7]. Blumstein et al. have suggested that these disturbances are related to a timing deficit. However, its origin was not well defined. Kent [2] has proposed that conduction aphasia is related to impairments of phonetic sequencing and deficient central integration of peripheral sensory informations that result in disturbances of sensory trajectories and poor motor control of speech. McNeilage [3] has insisted on the importance of motoricly defined acoustico-articulatory targets during speech production and their implication in a premotor level of speech control. On this basis, Poncet et al. [6] have suggested that the targets are part of an internal linguistic model involving a representation of the buco-phonatory system, and that within it, they are specified spatially and kinesthetically as premotor schemes. At the neurological level, such schemes have been associated with the left parietal lobe which is the cortical area classicly considered to be implicated in conduction aphasia. Internal laryngeal and supra-laryngeal buco-phonatory coordinates may play an essential role in these schemes. However, buco-phonatory gestures are displayed in time. Temporal coordinates must also be taken into account. Auditory temporal discrimination and feed-back are likely to play an important role in the control of the articulatory gestures involved in speech production. At the premotor level, the timing of speech sounds could be coded both kinesthetically and auditorily. The respective roles and interplay of these two types of temporal coding of speech production are not well known. The study of timing factors is possible through the investigation of voiced and unvoiced stop consonants. Contrary to voiced stops, at the articulatory level the realization of an unvoiced stop necessitates the inhibition in time of the vibration of the vocal folds. In addition, differences in transition length are important acoustic elements for the distinction between voiceness and voicelessness. In order to study production timing factors in conduction aphasia, we investigated the stop transition durations of a conduction aphasic and a control subject. We used the stop transitions rather than the VOT because, contrary to English, in French, all voiced stops are always prevoiced and voicing occurs over the entire duration of the closure, and usually also during the plosion. Moreover, in order to investigate the respective roles of the kinesthetic control and the auditory control mechanisms, we tested our subjects under normal condition and under speech masking noise condition. In this latter condition, the noise precluded the auditory control leaving only the kinesthetic control available.

2. SUBJECTS

Two right-handed men with normal audiograms took part in the experiment. They were matched for educational level and pronunciation characterisitics. The control subject was thirty two years old. The Aphasic subject was fifty years old. At the onset, his impairments included global aphasia and right hemiparesia. Within a few days, they evolved to right hemianesthesia and conduction aphasia associated with discrete reading, and writing problems as well as acalculia. MNI showed a hypodensity in the left sylvian territory suggesting a sylvian ischeamic accident. The damage was centered on the inferior parietal lobule and extended to the superior parietal lobule, to the *plis courbe* and the periphery of the somesthesic paracentral parietal areas. In depth, the lesion extended to the upper surface of the posterior part of left lateral ventricule.

3. PROCEDURES

The subjects were tested in an anechoic chamber on a single word repetition task. An experimenter who could not be seen, read each word before the repetition. The subjects did the task two months after the accident of the aphasic. They were tested three times with the same list of words. During the third repetition of the list, a 100 dB spl loud speech masking noise was diffused through audiometric ear-phones. In this last condition, the subjects were specificly asked to speak normally in order to avoid a raising of the intensity of the voice. The list consisted in 552 words with 90 target words randomly distributed among them. They were chosen in order to study and compare all the six stop consonants of French. The stops were either at the initial position for monosyllabic words or at initial and intervocalic positions for bisyllabic and trisyllabic words. A total of 108 stop consonants were included in the target words. Out of these stops consonants, 18 came from monosyllabics, 36 came from bisyllabics and 54 from trisyllabics. The vocalic surrounding was /ɑ/ and in a few cases /ɛ/ or /ɔ/. In our acoustico-phonetic analysis, we considered only the correct productions of the aphasic, and only the second repetition (with no speech masking) and the third repetition (with masking noise) were taken into account. The segmentation was done with the phonetic Bliss Sytems program [5]. The stop transition durations were taken between the beginning of the stop plosure and the first noiseless vocalic period following the stop. These measures constituted the dependent variable. A completely between analysis of variance was performed for each type of word length: monosyllabic, bisyllabic and trisyllabic. The individual stop consonants constituted the random variable. The experimental factors that we studied, were the type of subject (aphasic or control), the voiceness of the stop (voiced or unvoiced) and the repetition condition (with or without speech masking noise). An additional factor, i.e., the position of the stop in the word, was studied for the bisyllabic words (2 positions) and the trisyllabic words (3 positions). The three parametric analyses were performed using a SAS general linear model procedure for analysis of variance. In the phonological analysis, we considered all the target words and the number of errors. The control subject did not make any error. Only the data of the aphasic subject were considered. We compared his performances on the basis of the voiceness or unvoiceness of the stop consonant in target words and the presence or absence of speech masking noise during the repetition. Performance differences based on word length were also examined. The analysis was performed using a SAS categorical analysis procedure for dichotomous variables.

4. RESULTS

The results of the phonetic analysis showed significant effects of the interaction between subject and voiceness for the three types of word length ($F's \geq 5.30$, $p \leq 0.03$). In all cases, the differences between voiced and unvoiced stop transitions were minimal for the aphasic. On the contrary, the voiced stop transitions of the control subject were much shorter than his unvoiced stop transitions (Table I).

Table I Voiced (V+)and unvoiced (V-) transition durations in milliseconds

	Aphasic		Control	
	V+	V-	V+	V-
Monosyllabics	10.4	15.3	6.8	20.2
Bisyllabics	10.8	15.7	6.8	19.4
Trisyllabics	10.6	16.4	7.1	19.6

These results seem to indicate that voiced and unvoiced stop consonants are poorly differentiated by the aphasic subject. The phonetic analysis also showed significant effects of the interaction between subject and noise condition for the three types of word length ($F's \geq 3.85$, $p \leq 0.05$). In all cases, the stop transition durations of the aphasic became much shorter in noise while those of the control subject tended to remain the same in noise and without noise (Table II). These

results seem to indicate that the aphasic has more difficulty maintaining the duration of stop transitions without auditory feed-back.

Table II Stop transition durations in noise (N+) and without noise (N-)

	Aphasic		Control	
	N+	N-	N+	N-
Monosyllabics	7.6	15.9	13.7	13.7
Bisyllabics	10	16.1	14.4	12.3
Trisyllabics	9.9	15.1	13.8	12.9

In addition to these effects, in the bisyllabics and trisyllabics of the aphasic, significant differential effects of noise could be detected with respect to voiced an unvoiced stops (F's $\geq$ 4.6, $p \leq 0.04$). The unvoiced stops transitions always became shorter in noise while the voiced stop transitions did not change much in noise. On contrary, in monosyllabics both voiced and unvoiced stop tansitions became smaller in noise (Table III).

Table III Interaction between noise and voicing for each type of word length

	Noise +		Noise-	
	V+	V-	V+	V-
Mono Aphasic	5.3	10.4	13.8	17.8
Mono Control	6	19.9	7	20.4
Bi Aphasic	9.3	10.7	13.1	18.9
Bi Control	6.6	21.3	7	17.6
Tri Aphasic	8.5	11.3	11.8	20.6
Tri Control	5.7	21.8	8.1	17.9

In the bisyllabic words of the aphasic, a significant effect of the position of the stop consonant in the word was found in conjunction to the effect of noise on voicing ($F(1, 107) = 7.59$, $p<0.007$). Voiced and unvoiced patterns of change in noise and without noise differed only for the intervocalic stop position. At that position in noise, voiced stops did not change while unvoiced stop durations shortened dramaticly and tended to become undifferentiated from those of voiced stops (Table IV).

That level of interaction could not be detected in trisyllabics, even though patterns of changes for initial and inter-vocalic stops did not seem to differ from those of bisyllabics.

Table IV Effect of the position of the syllable in the bisyllabics of the aphasic subject

	Noise +		Noise -	
	V+	V-	V+	V-
Syllable 1	9.3	12.4	16.3	17.3
Syllable 2	9.3	9.6	8.1	21.3

At the phonological level, the analysis showed a significant effect of the word length ($X^2 = 12.55$, $p<0.02$). The error rates increase slightly from monosyllabics (25%) to bisyllabics (27%), and becomes much higher in trisyllabics (43%). These effects further suggest that longer words which include intervocalic stops, are more difficult to realize for the aphasic.The results of the phonological analysis also showed that the error rate for unvoiced stops (33%errors) is significantly higher ($X^2 = 4.83$, $p<0.03$) than that for voiced stops (22%errors). This seems to indicate that, for the aphasic, unvoiced stops are more difficult to realize. The phonological results also showed a significant higher error rate $X^2 = 7.29$, $p<0.007$) in the condition with masking noise (43%) than in the condition with no noise (26% errors). This seems to further indicate, that when auditory feed-back is suppressed, phonation is more diffcult for the aphasic.

5. DISCUSSION

At the acoustico-phonetic level, the poor differentiation of voiced and unvoiced stop consonants of the conduction aphasic, suggests that his timing coordinates are affected. The differential duration parameters for the realization of voiced and unvoiced stops seem to be blurred. This deficit suggests that temporal coding is likely to be an important part of premotor speech production schemes. In the phonological analysis, the higher error rate of the aphasic for unvoiced stops, further suggests that timing could play a specific role. The realization of unvoiced stops is likely to involve a temporally programmed inhibition of the vibration of the vocal folds. This is not necessary for voiced stops and this could be

why they are less prone to errors than unvoiced stops.The shortening of the stop transitions of the aphasic in masking noise, when auditory feed-back is not possible, suggests that articulatory timing can be controlled auditorily. The preclusion of this control could explain that durations are not maintained. The results further seem to indicate that non-auditory control of phonation, i. e., kinesthetic control, is deficient in conduction aphasia, but auditory control could compensate for it to a certain extent. The higher error rate of the aphasic in masking noise also support this view. The results showed a very peculiar effect of masking noise for the aphasic. Unvoiced stops transitions are significantly shortened in bisyllabics and trisyllabics,while voiced stop transitions tend to remain similar in noise and with no noise. It suggests, that in words with more than one syllable, adequate vocal folds inhibition is diffcult to control audio-temporally. This could be specificly related to intervocalic stops because their presence constitutes the major difference between polysyllabic words and monosyllabic words. The role of auditory timing control of phonation might start to increase with increasing complexity and length of words. In addition, the very high error rate of the aphasic in trisyllabics seems to further indicate that successful word repetition is more difficult in longer words that necessitate important cross-syllabic integration. The audio-motor speech control system could be more specific of inhibitory control, in some cases at the intrasyllabic level, as for the production of unvoiced stops, but mostly at the cross-syllabic level for integrated cross-syllabic production. Conversly, the kinesthetico-motor sytem could be linked to an excitatory system, which might be more devoted to the command of the articulators for the initialization of speech and for the successive intra-syllabic integration during ongoing phonation. The performance of the kinesthetico-motor system for the production of a complete word might rapidly decrease with increasing integration needs such as for words with more than one syllable. In normal conditions, the two systems could operate synergistically for the control of the articulatory and acoustic characteristics of speech sounds. They both might encode speech timing but each specializing at a different level : (i) with an intra-syllable focus for the kinesthetic control system which is likely to be more automatic and removed from volontary control, (ii) and with a cross-syllabic focus for the auditory control system with some possibility to overtake intrasyllabic control especially when the kinesthetic system is deficient. Within that context, conduction aphasia can appear as a deficit centered at the intra-syllabic level of speech production and both auditory and kinesthetic timing coordinates seem to be affected. Secondary effects might result at the cross-syllabic level, especially when auditory control is removed.

6. REFERENCES

[1] Blumstein S. E., Cooper W.E., Goodglass H., Statlender S. and Gottlieb J. (1980). Production deficits in aphasia : a voice-onset time analysis. *Brain and Language, 9* , 153-170.

[2] Kent R.D. (1990). The acoustic an physiologic characteristics of neurologically impaired speech movments in *Speech production and Speech Modelling* , 365-401. Hardcastle W.J. and Marchal A. (eds.). The Hague: Kluwer Academic Publishers.

[3] MacNeilage M. (1970). Motor control of serial ordering of speech. Psychological Review, 77, 182-196.

[4] Nespoulous J.L., Joanette Y., Ska B., Caplan D., Lecours A.R.(1987). Production deficits in Broca's Aphasia and Conduction Aphasia : Repetition versus Reading in *Motor and Sensory Processes of Language*. Keller E. and Gopnik M.(eds.)

[5] Mertus J. (1989). Barus Lab Interactive Speech System (Bliss) Software. Department of Linguistic and Cognitive Sciences, Brown University.

[6] Poncet M., Ali Cherif A., Brouchon M. (1980). Néologismes: rôle du cortex pariétal dans le contrôle moteur des organes buco-phonateurs. *Grammatica VII, 1*, 115-130. Université de Toulouse le Mirail.

[7]Tuller B. (1984). On categorizing aphasic speech errors. *Neuropsychologia* , *22*, 547-557.

This work was partly supported by a Lavoisier fellowship awarded to the first author. The authors wish to thank for their help, Sheila Blumstein and the Departement of Linguistic and Cognitive Sciences of Brown University , as well as Daniel Auteserre and the Institut de Phonétique of The Université de Provence.

A CROSS-LANGUAGE STUDY OF VOICING CONTRASTS OF STOPS

Katsumasa Shimizu

Dept. of Linguistics, Univ. of Edinburgh, Edinburgh, Scotland.
Dept. of English, Nagoya Gakuin Univ. , Aichi-ken, Japan.

ABSTRACT

The present study is concerned with cross-language phonetic differences of voicing contrasts of stops in six Asian languages. Such features as voice onset time (VOT), fundamental frequency (Fo) and the contour, spectral analysis, and the onset F1 frequency were examined, and their cross-language comparisons were made. The results of acoustic analysis showed that each language uses several acoustic dimensions in different ways for distinction of voicing categories and "same" sounds in these languages show some language-specific properties as well as features which are common to many languages.

1. INTRODUCTION

The present study is concerned with cross-language phonetic differences of voicing contrasts of stops in six Asian languages, and to explore the ways to describe some cross-language characteristics of stops. The languages investigated are Japanese, Mandarin Chinese, Korean, Burmese, Thai, and Hindi. The examination is mainly based on the acoustic analysis of initial stops in these languages, several numbers of subjects in each language took part in the experiments for recording linguistic materials. The main issues of investigation is to clarify the phonetic characteristics of voicing contrasts of stops in these languages, to examine the language-specific properties and to make cross-language comparisons on these features.

2. ACOUSTIC ANALYSIS

2.1. Voice Onset Time

Voice onset time (VOT) is usually defined as the time interval between the onset of voicing and articulatory release of the stop consonants, and is a timing dimension of the onset of voicing to articulatory oral release [1]. Based on the acoustic analysis in the present experiment, it has become clear that VOT functions for distinguishing voicing categories of initial stop consonants if they are based on the timing event of glottal and supralaryngeal movements. However, as has been noted in previous studies [2], VOT is not sufficient for distinguishing the categories of stops in such languages as Korean and Hindi which use such laryngeal features as glottal tensing or glottal stricture. In these languages, VOT is unable to distinguish tense stops from lax stops, and voiced stops from breathy voiced stops. This means that VOT is not sufficient if laryngeal features other than glottal timing are involved in the voicing distinction.

Among the three major voicing categories of voiced, voiceless unaspirated and voiceless aspirated stops, it was found that the VOT values of voiceless unaspirated stops show a wide range of variability, while those of voiceless aspirated stops show little variability in these six languages. This implies that voiceless unaspirated stops are articulated with language-particular characteristics and have some flexibility in choosing the articulatory timing region in the VOT continuum. The reason that voiceless aspirated stops show little variability is that aspiration requires a carefully adjusted timing event of glottal width and articulatory release; i.e., the timing when glottal width reaches its maximum opening must be adjusted with articulatory release.

2.2. Fundamental Frequency (Fo) and its contour

As to Fo and its curve, it was demonstrated in the present study that voiced and voiceless distinctions have a different effect on the Fo perturbations of the following vowels, and voiceless stops are generally associated with a higher Fo, while voiced stops are associated with a lower Fo. In Korean, all stops are phonemically voiceless, and the distinction between tense and lax stops affects the Fo perturbations, so that voiceless lax stops show the lowest Fo values compared to voiceless tense stops and voiceless aspirated stops. In Chinese, there was no marked difference between the voiceless unaspirated and voiceless aspirated stops, but in Burmese and Thai there were significant differences between these categories, and the magnitude of differences in Fo values were not different from those of non-tonal languages such as Korean.

It was also demonstrated that there is a difference in the Fo curve from the onset to steady-state portion. Voiceless stops tend to show a lowering pattern, and voiced stops a rising one. In Japanese, the effect of voiceless stops is not apparent, and a level pattern was observed. In Korean, a clear-cut distinction between voiceless tense stops and voiceless lax ones is observed, and the tense stops show an abrupt falling Fo curve. Furthermore, in Hindi, Fo curves of the breathy voiced stops show the lowest values and demonstrated a characteristic Fo pattern of fall - rise.

2.3. Spectral Analysis

The examination of power spectra in each language reveals some differences in intensity level and spectral characteristics. The spectral characteristics can be examined in the regularity of peak energy distribution, level of intensity, bandwidths and the spectral shape. Although the degree of regularity is difficult to measure, it can be said as a general trend that voiced stops show more regularly distributed energy peaks, while voiceless aspirated stops tend to show less regularly distributed energy peaks. Furthermore, although it is generally known that voiceless stops show a greater articulatory force; i.e., higher rate of airflow, than voiced stops, this trend was not consistently observed, and some languages such as Japanese did not show any marked differences.

2.4. The Onset Frequency of the First Formant

As to the onset frequency of the first formant, it was found that there is a difference in F1 onset frequency between voiced and voiceless stops; the onset frequency is in most cases higher in voiceless stops than it is in voiced stops. Differences in the onset frequency reflect differences in the speech production of these types of stops. It was also found that the difference in F1 onset frequency is affected by the ones of following vowels, and stops followed by low vowels show greater changes than those followed by high vowels.

3. SUMMARY

The languages in the present study use several acoustic dimensions for voicing categories in different ways, and the "same" voicing categories which are represented by the same phonetic symbols have a language-specific variability, as well as features which are common to many languages. VOT functions for the distinction if the voicing categories are based on the laryngeal timing in relation to the oral release. If other laryngeal features are involved, other dimensions are needed for distinction. Among the major voicing categories, voiceless unaspirated stops show a wide range of variability in the glottal and supralaryngeal timing events, while voiceless aspirated stops do not. The Fc at the vowel onset represents the initial state of the glottal adjustments and is significant for characterizing the voicing categories which involve a change of initial glottal gestures as found in Hindi breathy stops. Spectral analysis such as intensity level and spectral shape does not appear to provide a useful cue for distinguishing major vicing categories. Finally, the F1 onset frequency is useful for distinguishing the voicing categories of stop consonants.

4. REFERENCES

[2]Han, M.S. and Weitman, R.S.(1970) "Acoustic features of Korean /P, T, K/, /p, t, k/ and /p^h, t^h, k^h/" *Phonetica*, 22, 112-128.

[1]Lisker, L. and Abramson, A.S. (1964), "A cross-language study of voicing in initial stops: Acoustical measurements." *Word*, 20, 384-422.

ACOUSTICAL CUES FOR VOICED AND BREATHY FINAL STOPS IN GUJARATI LANGUAGE

Christine Langmeier

München, Germany

The problem of recognizing and discriminating stops is well known to those working in the field of speech recognition. To discriminate between voiced and breathy final stops of monosyllables several acoustical parameters of the preplosive vowel have been examined as predictor variables. Classification scores up to 90% in judged groups proved as highly tolerable performance.
The results confirm Fisher-Jørgensen [1] who claimed that the amplitude of the first harmonic (H1) is most selective to voiced versus breathy phonation.
On the other hand we have to consider the working relations of several acoustical parameters as there are the overall intensity, the first formant and the bandwidth of the first formant.
In one case making use of the difference measure between the first and second harmonic , a technique discussed by Ladefoged et al [2] generates better classification scores than computing on the measurements of H1.

1. INCTRODUCTION

Intensive research has been carried out to investigate the acoustical cues of different stop categories.
For initial stops a great impact on the following vowel has been admitted. For example Schiefer [3][4] found that in Hindi language fundamental frequency and overall intensity are cues to differentiate four stop categories in word-initial position: 1) voiceless aspirated 2) voiceless 3) voiced and 4) breathy stops.
We might doubt that comparable oppositions hold also for a word-final position because the contrasts could be in some way neutralized. Can we notice a regressive assimilation on the foregoing vowel?
A word-final distinction of voiced and voiceless stops has been discussed in the anglo-american literature as a phenomenon of vowel and stop closure duration, e.g. by Wardrip-Fruin [5].
A voiced final stop should lengthen the foregoing vowel and should be characterized by a relatively short closure duration (it is understood that duration measurements are always relative and give evidence only in controlled surroundings).
Lately there appeared also spectral arguments by Van Summers [6]. The first formant was found to stay somewhat lower in a voiced environment.
That there is indeed a physiological basis for a fourfold stop distinction in indo-arian languages even in word-final position has been shown by Yadav [7].
Employing a fiberoptic technique he found four grades of glottis opening when a speaker of Maithili produced the test words. At the moment of stop release the glottis is 1) still closed for voiced final stops 2) slightly open for breathy stops after a roughly periodic phase during the closure 3) slightly open for voiceless stops after an open phase during the closure and 4) "widely" open for voiceless aspirated stops.
The proposed investigation examines several acoustical cues of two final stop categories: voiced and breathy final stops in mono-syllables of Gujarati language.

2. METHOD

The recordings of five native Gujarati speakers were made in Baroda by L. Schiefer and Bh. Modi.

Modi (female, 53 years, grown up in Bombay and district Surat, living for 29 years in Baroda)
Dalal (female, 47 years, grown up in Baroda)
Shah (male, 40 years, grown up in Ahmedabad, living for 14 years in Baroda)
Patel (male, 29 years, grown up in Baroda)
Vora (male, 26 years, grown up in Bombay and Baroda).

The following meaningful words were produced several times (it came up to a sample size of about 120 words for each person with 8 to 9 occurances of the same word):

rab	a̲b^h
sad	sa̲d^h
vaδ	ma̲δ^h
rag	va̲g^h
mɔd	mɔ̲d^h
rɔδ	rɔ̲δ^h
d^sɔg	d^sɔ̲g^h

porridge	sky
voice	accomplish
fence	musical form
raga	tiger
joy	honey
cry	stubborn
world	fragrance

You may read the vowel with the underscore as a breathy vowel; the sigma-letter as the vowel schwa; the delta-letter as a retroflex dental stop and "d^s" as a palatal affricate.

3. SEGMENTATION

The computer aided segmentation of the single word oscillograms was as follows:
The initial consonant, the vowel, the closure and the burst of the final stop were cut by visual and auditive inspection of the digitalized acoustical signal.
There was no need to define a transition phase between the vowel and the stop closure part of the signal, the cut was set as soon as the amplitude was small enough to corresponde with the articulatory closure phase and the typical oscillating pattern of the vowel blurred.
The beginning of the burst then was defined as a small but abrupt vertical impuls of the signal. The final phase of the stop melted into the tape noise and therefore its duration couldn't be measured. The duration of the closure was measured proportional to the duration of the initial consonant and the following vowel.

4. ACOUSTICAL ANAYLSIS

The acoustical analysis was performed with the facilities of the Institute of Phonetics, München.
For all speakers a pitch synchrone discrete Fourier transformation of the vowel part of each word was calculated and for three speakers also a formant analysis by LPC-procedure.

Alltogether the following acoustical parameters computed for each word could enter as canonical variables into the discriminant analysis:

- overall intensity
- fundamental frequency
- intensity of H1
- intensity of H2
- intensity difference between H1 and H2
- relative closure duration
- first formant F1
- bandwidth of F1

5. DISCRIMINANT ANYLSIS

Working with the statistical package SPSS-PC+ a discriminant analysis was computed for every speaker.The jackknife-validation-procedure (one-leaving-out-method) was used to reduce the misclassification bias.
For three samples the assumed multivariate normality and equality of group covariance matrices didn't hold. Consequently the degrees of freedom had to be reduced.
One speaker (Modi) showed different characteristics for vowel /a/ and /ɔ/, therefore the sample had to be devided.
For another speaker (Shah) only one variable remained in the analysis.

6. DISCUSSION

A view of the discrimination score figures reveals that the first harmonic is the most selective variable for all speakers except one. In that case a transformation measure of H1 (the H1-H2-difference measure) shows best results. Indeed 75% of speaker Patel's correct discrimination scores are expressed by the Diff-variable.

All formant measurements showed up to be quite selective between the voiced and the breathy group.

In the cases of Modi and Patel the hight of the first formant is inversely related to breathiness, in the case of Vora the broader bandwidth within the breathy group makes the difference.

Looking at the canonical discriminant function coefficients of the Dalal-sample the overall intensity contributes to the discrimination function because it correlates negatively with the first harmonic in the breathy group but not in the voiced group. The coefficients can be seen in the figure showing the discrimination scores of the source groups.

The impression that the closure duration in breathy final stops could be somewhat longer than in purely voiced stops could't be verified for the whole samples. Indeed closure duration is strong enough to discriminate the phonation type in samples of retroflex dental articulations.

7. FINAL REMARK

In minimal pairs the mean maximum intensity differences of H1 between the two groups of phonationtype are around 3 db, the Diff-measure which corrects in a sense for overall intensity variation shows even lower means.

If we consider that Bickley [8] used 3-db-steps to test perception differences between clear and breathy vowels and found a 75% discrimination score at the 9-db-borderline, we may find that discrimination analysis (or let's say a computer) is much more sensitive than human perceptionists.

8. LITERATURE

[1] FISHER-JØRGENSEN, E. (1967).
Phonetic analysis of breathy (murmered) vowels in Gujarati.
Indian Linguistics 28, 71-139.

[2] LADEFOGED, P.; ANTONANZAS-BARROSO, N. (1985).
Computer measures of breathy voice quality.
UCLA-WP 61, 79-86.

[3] SCHIEFER, L. (1987).
The role of intensity in breathy voiced stops: a close link between production and perception.
Proc. 11th ICPhS Tallinn, Vol. 5, 362-365.

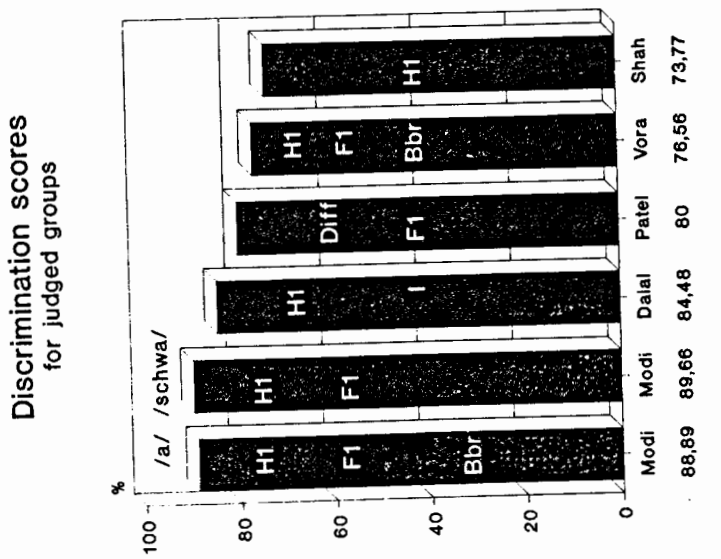
[4] SCHIEFER, L. (1987).
F0-Perturbations in Hindi.
Proc. 11th ICPhS Tallinn, Vol. 1, 150-153.

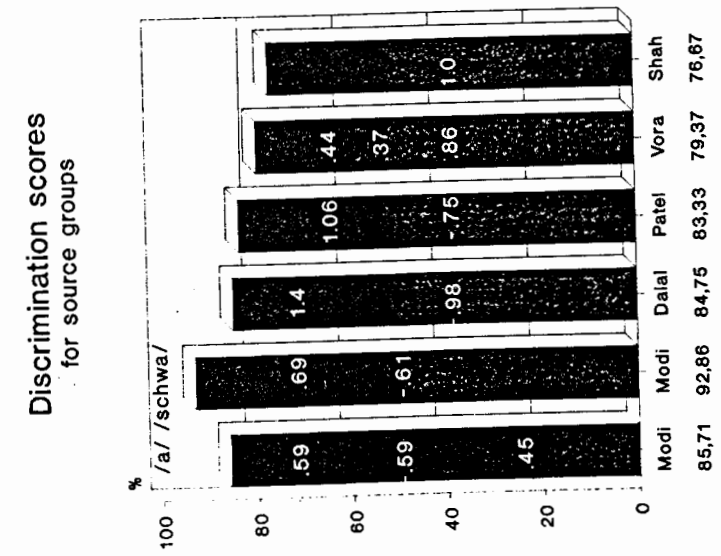
[5] WARDRIP-FRUIN, C. (1980).
On the status of the temporal cues to phonetic categories: preceding vowel duration as a cue to voicing in final stops.
PhD Diss. Stanford University

[6] VAN SUMMERS, W. (1988).
F1 structure provides information for final-consonant voicing.
JASA 84, 485-492.

[7] YADAV, R. (1984).
Voicing and aspiration in Maithili: a fiberoptic and acoustic study.
Indian Linguistics 45, 1-30.

[8] BICKLEY, C. (1980).
Acoustic analysis and perception of breathy vowels.
MIT Speech Communication Working Papers Vol. 1.





EQUIVALENCE PERCEPTUELLE ET DIFFERENCIATION ACOUSTIQUE

P.DURAND

Institut de Phonétique, URA 251 CNRS,
Aix en Provence, France

ABSTRACT :
The aim of this paper is to show evidence that in speech there are continuities that prevent sounds to be mere illustrations of the sound paradigmatic unit, the phoneme. The study of a set of nonsense items points out that consonant environment does not explain entirely spectral shape of vowels. The vowel of the previous syllable influences also its acoustic achievement. So, the actual spectral values of a sound unit reflect also this syntagmatic aspect and are not only the reflect of some paradigmatic invariance.

1. INTRODUCTION :

Beaucoup de linguistes africains ont pu observer, en écoutant des séquences de type " comme bois" , "scène deux", que le français, à l'instar de l'anglais "some boy", du serer , du peul, ou de l'éwondo utilisait des occlusives prénasalisées. Et on se demande pourquoi les spécialistes de l'acoustique ne semblant prendre en compte que l'aspect paradigmatique des unités sonores, ou de ne les considérer que comme des illustrations de phonèmes, se sont souvent contentés d'affiner les valeurs centrales moyennes au lieu de prendre compte la variabilité inter-contextuelle de ces unités, telle qu'elle se manifeste, opposée à leur variabilité face aux unités susceptibles d'apparaître dans le même contexte.

2. BUT DE L'ETUDE :

Nous nous proposons, à partir des occlusives sourdes françaises et de quatre voyelles françaises proches tant du point de vue phonologique que du point de vue articulatoire et acoustique de montrer que les unités sonores ne sont pas l'illustration pure d'entités paradigmatiques mais plutôt des réalisations syntagmatiques qui ont pour fonction, dans la perspective de communication où elles s'insèrent, de s'opposer à la fois aux unités contiguës et aux autres unités susceptibles d'apparaître dans un contexte identique. Cependant, compte tenu des contraintes spécifiques du tractus vocal et compte tenu de la vitesse de la communication orale, les valeurs des unités sonores isolées ne sont qu'approchées. Ce fait ne pose pas de problème pour la transmission de l'information dans la mesure où coexistent dans le message à la fois des phénomènes de continuité , de rupture de cette continuité et que l'unité sonore à décoder s'oppose dans le contexte où elle s'inscrit à toute unité sonore susceptible de commuter avec elle dans le même environnement et dans la même langue.

C'est dans cette perspective, que doivent être intégrés les phénomènes très tôt signalés d'influence bilatérale [5], de chevauchement [4,8], et les interrogations sur les limites de la continuité[8-9] dans la production [9,10,11,12], qui n'ont été intégrés, à de rares exceptions près, nous semble-t-il, que dans le cadre de la syllabe ou de la demi-syllabe [6], vu le nombre de paramètres à intégrer.

Aussi serait-il vain dans les limites d'une étude de ce type, de prétendre apporter plus que quelques orientations sommaires face aux questions ainsi soulevées, mais plus simplement de montrer, dans un cadre contraignant, que, dans le signal observé, coexistent les trois éléments mentionnés à savoir continuité du message, rupture dans ce même message et indices dans la substance acoustique suffisant au décodage de cette unité et d'elle seule. La composition du corpus ne rendra pas nécessaire l'analyse du second facteur (rupture), mais permettra de mieux cerner les phénomènes de continuité, et la présence dans la substance d'indices suffisants à la discrimination des éléments vocaliques étudiés.

3. PROCEDURE :

Une série de logatomes de forme /CV'CVC/ a été enregistrée en chambre anéchoïque. Les éléments consonan-tiques étaient /p,t/ et /k/, et vocaliques /i,e,y/ et /ø/. Les séquences enregistrées ont été l'objet d'une analyse spectrogra-phique (0-8000 Hz, 300 Hz), susceptible de faire apparaître des éléments établissant d'une part, les éléments de continuité dans le signal, de l'autre, des indices spécifiques de chaque élément vocalique en contexte identique.

4. LES ANALYSES

4.1 Les éléments de rupture : Du fait même de sa composition, le corpus étudié oppose les sons entre lesquels existe la plus grande discontinuité possible: des occlusives sourdes et des voyelles. De ce fait, la rupture entre deux éléments consécutifs de la chaîne sonore est maximale, et il est inutile d'insister sur ce point.

4.2 Les éléments de continuité Pour la même raison, parler de continuité à l'intérieur de séquences ainsi construite peut sembler paradoxal. Cependant nous porterons notre attention sur les points suivants :

4.2.1. : La continuité vocalique: Une certaine continuité vocalique subsiste malgré les "accidents consonantiques". Dans le cadre qui est celui de cette étude - des logatomes, et les voyelles du corpus - cette continuité sera particulièrement difficile à mettre en évidence.

4.2.2.:Les consonnes occlusives: Elles introduisent leur propre continuité sur la trame vocalique. Comme nous l'avons montré antérieurement[2], les effets de l'occlusion atteignent les valeurs centrales du spectre et les influencent de manière plus ou moins forte en fonction des éléments vocaliques et consonantiques en contact. Ici, l'influence bilatérale sera maximale.

4.3 Les éléments spécifiques : Il s'agit à terme d'examiner si, existent dans le signal des éléments spectraux spécifiques susceptibles d'opposer une variabilité intra-contextuelle à une variabilité intercontextuelle.

5. LES RESULTATS

5.1 Les éléments de rupture Voir supra 4.1

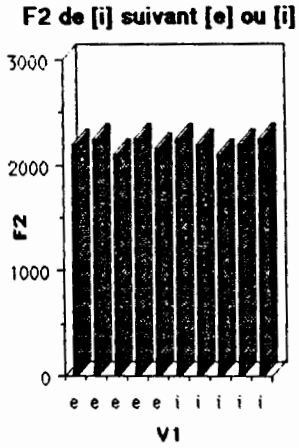
5.2 Les éléments de continuité
Outre les éléments supra segmentaux (accent, intonation),dont l'influence est importante dans l'unité de la séquence sonore, mais qui n'entrent pas dans le cadre de cette étude, les analyses ont porté essentiellement sur la valeur centrale des spectres vocaliques dans la mesure où c'est le centre de l'élément vocalique qui est le moins sensible à la coproduction au niveau de la séquence sonore.

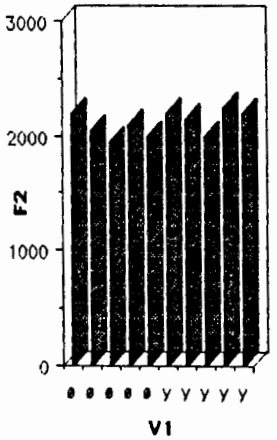
5.3. Les éléments spécifiques.
Les résultats du dépouillement font apparaître pour chaque réalisation l'influence de la labialité de la voyelle précédente, même si celle-ci est pondérée par la consonne intervocalique.

6. DISCUSSION

6.1. Le problème du corpus
Deux phénomènes peuvent relativiser les quelques points observés. Tout d'abord l'utilisation de logatomes. Comme nous l'avions remarqué précédemment [2-3], l'utilisation de logatomes - comme le choix de mots rares - conduit à une articulation plus

précise des séquences articulatoires. De ce fait, les données obtenues ne ne possèdent pas une variabilité analogue à celle observée dans des séquences de parole continue.





Figures 1 a et b : Influence de la voyelle précédente sur la valeur centrale de F2 de [i] en syllabe tonique.

D'autre part, comme le soulignent R.A.Cole and al. [1], et comme le prouvent les résultats en intelligibilité et en reconnaissance de la parole, il

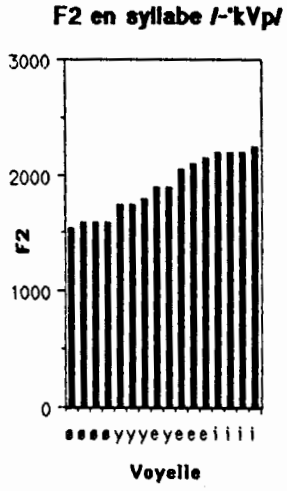


Figure 2 : Répartition de la fréquence du second formant des voyelles étudiées dans une syllabe /-'kVp/.

existe des différences spectrales considérables entre les mots isolés et ceux produits en parole continue.
Paradoxalement ceci ne peut que renforcer l'optique dans laquelle nous nous plaçons, même si, dans le cadre de cette étude, il n'est pas possible d'aborder des séquences sonores de taille supérieure à deux syllabes et qu'une étude prenant en compte une multitude de facteurs serait alors nécessaire [6].
Cependant, la mise en évidence d'indices intra-contextuels opposés à une variabilité intercontextuelle est difficile à montrer étant donné le phénomène de sur-articulation qui caractérise les logatomes. Cependant ce type de corpus est le seul qui permette de montrer avec une rigueur suffisante les effets possibles de variables sélectionnées. L'hypothèse de départ reçoit des résultats encourageants mais peu significatifs statistiquement du fait de la taille du corpus, des analyses actuellement exploitées et du petit nombre de sujets enregistrés dans cette phase préliminaire.

6.2. Les résultats :
Comme l'indiquent les Figures 1[a] et [b], ils montrent surtout l'influence de la nature de la voyelle prétonique, et plus particulièrement de la labialité de cette voyelle sur la voyelle accentuée.
En effet, les mesures effectuées dans le même entourage consonantique révèlent que subsiste à des degrés divers un certain nombre de chevauchements (Ex. : Figure 2), difficilement réductibles au niveau intra-syllabique. L'arrondissement ou l'étirement de la voyelle pré-tonique semble avoir une influence sur la fréquence centrale du second formant, et ce même dans des logatomes. Elle permet d'expliquer pour une large part, sans que des règles puissent être établies, les chevauchements que la pondération des indices spectraux vocaliques en fonction de leur entourage consonantique avait déjà diminué de façon notable.

7. CONCLUSION

Il est donc possible sur la base des résultats obtenus, qui ne sont, comme nous l'avons indiqué, fondés que sur un corpus limité composé de logatomes bisyllabiques, d'avancer que la réalisation des unités sonores est plus fortement dépendante de la chaîne sonore où elle s'insère qu'on ne le suppose souvent. Compte tenu du type de corpus utilisé et des mesures effectuées, ce travail ne peut représenter que l'ébauche d'études mettant en évidence, dans la parole continue les relations qui caractérisent chaque unité à partir du moment où elle est insérée dans une séquence sonore. Elle montre que, du point de vue acoustique, la coarticulation possède des conséquences qui ne se bornent pas aux limites de la syllabe. Même si les modifications articulatoires imposées par la coarticulation ne se traduisent pas par des modifications importantes du spectres, elles ne sont pas pour autant à négliger.

BIBLIOGRAPHIE

1 - Cole. R.A., Jakimik,J , Cooper W.E. Perceptibility of phonetic features in fluent speech. *J.A.S.A.* 64/1, 1978. pp.44-56

2- P.Durand. *Variabilité acoustique et invariance en français consonnes occlusives et voyelles*, Paris. Ed. du CNRS, Collection «Sons et Parole», dirigée par M.Rossi, vol.4,1985, p.25, 165.

3 - P.Durand. Les invariants relatifs : occlusives et voyelles diffuses, *1er Congrès français d'Acoustique*. colloque C2, pp.519-522

4- G.Fant : Descriptive analysis of the acoustic aspects of speech, *Logos*, 5, 1962, pp.3-17 ;

5-E.Fischer-Jørgensen :[p,t,k] et [b,d,g] français en position inter-vocalique accentuée, Papers in Linguistics and Phonetics to the memory of P.Delattre, The Hague, Mouton, 1972, pp.143-200.

6- Fujimura, O., Methods and goals of speech Production research, *Language and Speech*, 1990, 33/3, pp.195-258.

7- M. Grammont, *Traité de phonétique*, Paris Delagrave,1933, 9 ème ed. 1971, p.378 ; "les seules occlusives que nous supposions devant une voyelle sont celles qui ont le même point d'articulation qu'elle".

8- A.S. House, G.Fairbanks, The influence of consonant environment upon secondary acoustical characteristics of vowel, *JASA* 25,1953, pp.105-113

9 - Kent, R.D., Minifie, F.D., Coarticulation in recent speech production models, *J. of Phonetics*, 1977, 5/2, pp.115- 133.

10 - Mattingly, I.G., The global character of phonetic gestures, *J. of Phonetics*, 1990, 18, pp. 445-452.

11- D.H.Whalen, Coarticulation is largely planed, *J. of Phonetics*, 1990, 18, pp.3-35.

12 - Zerling, J.-P., *Aspects articulatoires de la labialité vocalique en français*, Thèse de Doctorat d'Etat, Université de Strasbourg, 1990.

THE CROSS-LANGUAGE VALIDITY OF ACOUSTIC-PHONETIC FEATURES IN LABEL ALIGNMENT

Paul Dalsgaard[1], Ove Andersen[1] and William Barry[2]

[1] Speech Technology Centre, Aalborg University, Denmark
[2] Dept. of Phonetics, University College London, United Kingdom

ABSTRACT

Results are reported from a contrastive study in which the validity of articulatory based acoustic-phonetic features were analysed across Danish, English and Italian. The features are derived by means of a Self-Organising Neural Network, trained and calibrated solely on Danish training data. Test material from each of the languages was then used to obtain language-specific distributions for corresponding articulatory features. These were subsequently used to model allophones in the three languages.

Inter-language dependencies in the allophone models were examined in a label-alignment task using techniques from speech recognition to position allophone-transistion boundaries of large speech corpora. The results are stated in terms of accuracy relative to manually placed boundaries.

The work is in part funded by the Danish Technical Research Council and in part by the ESPRIT project 2589, Speech Assessment Methodology (SAM).

1. INTRODUCTION

During the last decade there has been a growth of interest in modelling sub-word units and their phonetic feature definition. This interest stems from the assumption that such an approach to continuous speech recognition will limit the need for large amounts of new speech data each time a new vocabulary is defined. In addition, the feature-based approach takes into account the co-articulation effects inherent in natural speech, and enables the use of well-established search techniques.

'Phonetic-unit' modelling, using triphone models together with whole-word models of function words has been successfully used in e.g. the SPHINX speaker-independent recognition system [1]. Allophonic modelling has been used in a HMM approach [2], where the individual models serve to identify the string of allophones constituting single words. In both approaches the models are based on cepstral coefficients.

In the research presented here, we have developed a new approach in which cepstral coefficients are transformed into a set of articulatory based features by a Self-Organising Neural Network (SONN), and subsequently used to model the 'phonetic units'. The approach has been applied in a continuous-speech recognition system [3], and in the task of label alignment of large speech corpora [4,5]. The latter task was chosen because of the urgent need for labelled speech databases to use in the training of more robust 'phonetic unit' models.

Previously, we have worked individually with a number of European languages. In this paper we examine the cross-language validity of a set of features by applying them to three languages in a cross-language label alignment task.

2. ACOUSTIC-PHONETIC FEATURES

We will only present the main characteristics of the process of transforming cepstral coefficients into a set of articulatory features (e.g. frontness, backness, closeness, dentalness and fricativeness to mention a few). Details can be found in [6].

The technique used is that of a Self-Organising Neural Network [7] consisting of a structure of 20*20 neurons. During the stimulation process each neuron is assigned a vector of adapted cepstral coefficients, and partly due to properties of the SONN and partly to the updating strategy used during the stimulation process, the cepstral vectors describing individual allophones organise themselves into 'clusters' which cover the network in an orderly way. Acoustically different allophones of the same phoneme will stimulate different neurons in the network.

Following the stimulation process, the entire reference training database is once again presented to the SONN for the purpose of calibration, and the number of firings of each neuron associated with each individual phoneme is counted. The normalised vector of counting rates for each neuron is multiplied with a matrix describing the inventory of phonemes for the language under investigation in terms of a set of distinctive phonetic features [8] giving a vector Φ of acoustically based feature values. The absolute value of each element of Φ is equivalent to the probability that the corresponding acoustic-phonetic feature is involved in the articulation process causing the specific neuron to fire. A neuron fires when the speech frame cepstral vector is the closest to the adapted cepstral vector of the firing neuron taken over all neurons of the SONN.

In each speech frame t, the SONN output is a vector $\Phi(t)$ which in principle could be used as the basis for modelling the individual allophones by a multi-dimensional probability density function [4]. However, some of the features are highly correlated, and therefore the features are submitted to a Principal Component Analysis, the output $\beta(t)$ of which is subsequently used for modelling the allophones. Details may be found in [5].

3. MODELLING OF ALLOPHONES

Each allophone j is modelled by the multi-parameter function

$$f_j(\beta(t)) =$$

$$L_j^{-1} \bullet \exp(-0.5 \bullet (\beta(t)-\mu_j)^T \Sigma_j^{-1} (\beta(t)-\mu_j))$$

where $L_j = (\|\Sigma_j\| \bullet (2\pi)^{\vartheta})^{1/2}$, ϑ the number of parameters in β, Σ_j the covariance matrix and μ_j the average vector for allophone j as given from the training data.

4. LABEL ALIGNMENT SYSTEM

The functionality of the Label Alignment System (LAS) is based on the assumption that the speech production process can be considered as a stochastic process emitting a sequence of parameters $\beta(t)$, and that the speech signal subsequently being submitted to label alignment manifests the same stochastic behaviour as used during the SONN stimulation and calibration. This allows the LAS to be implemented using the Viterbi Search and Level Building technique, known from speech recognition. In the context of label alignment however, the search is constrained by an independently given string of allophones corresponding to the speech signal being labelled. Details are given in [6].

In previous work we have established three independent LAS's for the European languages Danish, English and Italian, and tested them on large speech corpora. The test have shown the following overall tendencies: I) Labelling accuracy was encouragingly high overall, despite the limited amount of training data. II) Performance was no lower when the LAS were used in 'multi-speaker' mode (i.e trained on different speakers from those used for testing). III) Performance for different sound classes varied according to their acoustic segmentability in a manner that parallels human labelling performance: Fricatives and Plosives were labelled most accurately, post-Vocalic Liquids least accurately.

The results, and the general theory of distinctive features as a language-independent communicative framework, prompted the cross-language experiment reported here. We chose the Danish LAS trained on Danish, and ran the following experiments: A) The LAS was trained and calibrated on the complete reference recording by one male Danish speaker (approx. 2.5 minutes of continuous speech taken from the large, manually labelled reference speech database SAM-EUROM0), and the recordings of one English, one Italian, and one other Danish speaker were used to derive distributions for the common features, and to test the alignment accuracy. B) The LAS was trained and calibrated on 3 Danish speakers (approx. 6.5 minutes of speech) and the recordings of one English, one Italian, and one other Danish speaker were used to derive distributions for the common features, and to test the alignment accuracy.

The rationale for the experiments was 1) to investigate the degree to which cross-language variation in the acoustic expression of the common features affects labelling accuracy, compared to the effect of cross-speaker varia-

tion within one language, and 2) to investigate the effect of a larger amount of training data (with accompanying increased variation due to inter-speaker variation) on intra-language and cross-language label alignment.

The implications of the results are both practical and theoretical. In practical terms, cross-language application of a LAS can greatly alleviate the development of recognition systems for other languages by speeding up the annotation of large speech corpora needed for training and research. Theoretically, the results will indicate to what extent distinctive-features can be considered a substance-based system of phonetic distinctiveness that transcends individual language systems, or how far they should be understood as an abstract scheme with no substance outside the individual language.

5. RESULTS

Examination of the feature distributions for the three test speakers (Danish, English, Italian) can be undertaken from two angles. Firstly, the feature distributions for English and Italian derived from the Danish-trained and calibrated SONN can be compared to those found for English and Italian speakers derived from a SONN trained and calibrated on English and Italian, respectively [4,5,6]. Secondly, in preparation for a cross-language alignment test, the distributions of the English and Italian speakers can be compared to the Danish distributions to ascertain the *phonetically* closest sounds.

Lack of space prevents a comprehensive illustration, but in summary, it was found that the English and Italian distributions derived from a SONN trained on the same language are similar to, though somewhat better defined (i.e. with more extreme positive or negative values for the distinctive features) than those derived from a SONN trained on Danish.

As expected, comparison of English and Italian vowels with Danish shows clearly that very acceptable correspondences exist for some vowels, while others deviate in their feature distributions along the dimensions of known phonetic differences between the languages. Transcription conventions (SAMPA symbols used throughout [9]) do not necessarily provide a satisfactory basis for equivalence. Italian (I) and Danish (D) /i/ show good correspondence, while English (E) /i/ is better matched to D /e/ due to the greater degree of closeness of Danish /i/, see Figure 1.

Other good correspondences are : E and I /e/ with D /E/; I /O/ and E /Q/ with D /Q/; E /O/ with D /O/; I /u/ with D /u/; E /V/ with D /A/. In contrast, only weak matches were found for E /u/, which is not well defined by backness, for E /U, I, @/ and I /a/, which all have very weakly defined features.

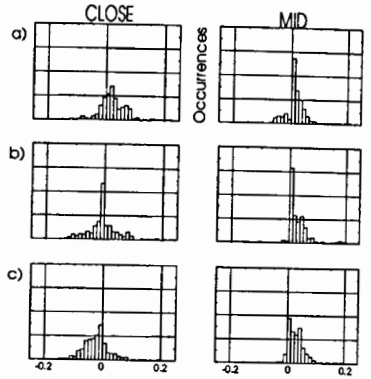


Figure 1. Selected Feature Distributions for a) D /i/, b) E /i/ and c) D /e/

In the consonant systems there are also clear correspondences between Danish and the two other languages (e.g. D, E and I /s/, see Figure 2; D and E /f/; D and E /m/), and cases where only a coarse approximation is possible (e.g. E /S, Z, T/ and I /J, L/).

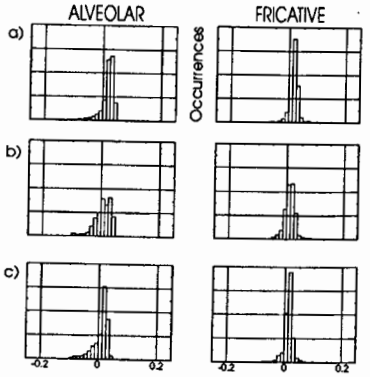


Figure 2. Selected Feature Distributions for a) D /s/, b) I /s/ and c) E /s/

The effect of these approximations is not necessarily manifested in the alignment results. Although overall, alignment accuracy is lower than when English and Italian material is used for training the SONN, greater inaccuracies are not always predictable from the lack of phonetic correspondence. For English, accuracy (to within ±20ms of the manual label) is comparable to the Danish speaker: D 62%, E 60.5%. But this is lower by 15-17% than with a SONN trained on English [4,5]. The most accurate segment transitions are E /s/ and E /S/ together (87%), which are both equated with D /s/, and E /f/ and E /T/ together (79%), which are equated with D /f/. The 47 % overall accuracy rate for Italian at ±20ms is 8% lower than after training with Italian data.

After training and calibration the SONN on three Danish speakers and testing alignment accuracy on the same 3 test speakers as above, accuracy (±20ms) for the Danish speaker rose to 68%, while accuracy for the English and Italian speaker fell to 58% and 46% respectively.

6. CONCLUSIONS

This study illustrates the application of contrastive phonetic principles within a quantitive, speech technology frame. The qualitative assessment of neural-net feature distributions provided a basis for specifying cross-language equivalents in three languages for use in a label-alignment system trained on only one of the languages.

The results of the feature comparison and label-alignment across the three languages indicate that the language-specific manifestation of common features differs enough to make the cross-language application of their distributions less efficient than the language specific application. This is less apparent with the smaller amount of training data (from one speaker), but becomes increasingly evident when the SONN is trained on 3 Danish speakers, given it greater coverage of natural variability, but making it more language-specific.

Given the obvious importance of covering as much of the natural variability as possible, which the effect of the increased training database showed, and given the relative success of some combined categories (E /s/ + E /S/ together, and E /f/ + E /T/ together), a modified approach to multi-lingual labelling appears feasible.

Future research will need to examine ways of covering inter-language variability over approximate phonetic equivalents while broadening the reference categories for alignment purposes. Combining cross-language sound groups into inter-language "broad-categories" [3, 10] for cross-language training is one way in which this might be achieved.

7. REFERENCES

[1] Kai-Fu Lee (1989), "Automatic Speech Recognition, The Development of the SPHINX System", Kluwer Academic Publishers

[2] A. Ljolje, S.E. Levinson (1991), "Development of an Acoustic-Phonetic Hidden Markov Model for Continuous Speech Recognition", IEEE Trans. on Signal Processing, Vol 39, No 1, pp. 29-39.

[3] P. Dalsgaard, A. Baekgaard (1990), "Recognition of Continuous Speech Using Neural Nets and Expert System Processing", International Journal of Speech Communication 9, pp. 509-520.

[4] P. Dalsgaard, W. Barry (1990), "Acoustic-Phonetic Features in the Framework of Neural-Network Multi-Lingual Label Alignment", Proceeding Int. Conf. On Spoken Language Processing ICSLP90, Nov. 18-22, Kobe, Japan.

[5] P. Dalsgaard, O. Andersen, W. Barry (1991), " Multi-Lingual Label Alignment Using Acoustic-Phonetic Features derived by Neural-Network Technique", IEEE Int. Conf. ICASSP91, May 14-17, Toronto, Canada.

[6] P. Dalsgaard (1990), "Phoneme Label Alignment using Acoustic-Phonetic Features", submitted for publication.

[7] T. Kohonen (1990), "The Self-Organizing Map", Proceedings of IEEE, Vol 78, No 9, pp. 1464-1480.

[8] P. Ladefoged (1982), "A Course in Phonetics", Harcourt Brace Jovanovitch, Publishers.

[9] J.C. Wells (1988), "Computer-Coded Phonetic Transcription", J. International Phonetics Association 17, No 2, pp. 94-114.

[10] K. Elenius, G. Takács (1990), "Acoustic-Phonetic Recognition of Continuous Speech by Artificial Neural Networks", STL-QPSR 2-3, Quarterly Report, KTH, Stockholm.

THE USE OF LPC AND FFT IN PHONETIC ANALYSIS

J. Rosenhouse and G. Rosenhouse
Department of General Studies
Faculty of Civil Engineering

The Technion, I.I.T. Haifa, Israel

ABSTRACT

The application of FFT and LPC methods in speech analysis is discussed here. When used side by side,these methods are complementary, which helps clarify various points that may remain if only one method is used.This approach is exemplified here for some Hebrew speech sounds (vowels, consonants) and some general speech features.

1. INTRODUCTION: FFT AND LPC

The Fast Fourier Transform (FFT) yields frequency spectra for given signals of a certain duration.This method is used in speech analysis to represent the speech output in the frequency domain for the given duration, and is a result of both the input signal and the filter (the vocal tract). On the other hand, the Linear Predictive Coding(LPC) method yields the response function of the vocal tract, eliminating as much as possible the effect of the input signal. This method is based on the approximation of the speech signal by a linear combination of past speech samples. Minimizing the sum of square differences over a finite interval, between the actual speech samples and the linearly predicted samples, a unique set of prediction coefficients is defined.

The advantage of the LPC is its accuracy and reliability in defining the basic speech parameters, mainly formants and spectra. It is analytically tractable, easy to implement and suitable for time varying speech signal analysis.

In ordr to define precisely the spectrum of vowels and consonants by the FFT method first a pitch detector is to be used. This enables the selection of signal duration that corresponds to integer numbers of pitch periods. Using inaccurate durations of input signals may yield errors. Another deficiency of the FFT compared to LPC is the large number of terms to be calculated by it, while only a small number of poles is required for the LPC values. Yet the FFT method can help in defining the number of poles to be used in an LPC program.

The LPC method is generally considered a more efficient method.Yet much experimental evidence from our work(using samples of 35ms duration)has shown a good correspondence between FFT and LPC in formant frequency definition (see Figures 1,2), though large differences have been found for amplitudes due to the input effect.In addition the FFT method yields the Fo and certain effects that are difficult to identify by LPC only. Hence, the use of both LPC and FFT in speech analysis to complement each other seems to be favorable.

In the literature comparison of FFT, LPC or other methods is to be found (see, e.g., |3|,|6|). Woods (|6|), for example, compares the spectrograph output with the LPC method. It should be noted, that FFT can be used so that the results would be easier to read. (This can be found in e.g. improved spectrographs, and the MATLAB(c) package which applies the FFT function may be used by users to write programs according to needs).

In the sequel, we note some examples based on our experiments performed at the Lab of Medical Electronics, the Faculty of Electrical Engineering at the Technion. Some experiments were also done in the framework of a Technion D.Sc. thesis (|1|). Our recorded natural speech material was analyzed by programs written at the Technion, by both LPC and FFT methods.

2. EXAMPLES

2.1. Pitch Detection

The LPC method is normally not intended for pitch analysis(Fo).As the FFT program gives "raw" harmonics (without "smoothing") of both source and filter, it is easy to find the Fo and the other formant values and distinguish between them visually. Applied natural speech analysis(for linguistic or even medical purposes of voice quality measurements) often needs to define or find the speaker's pitch (i.e., Fo) as well as the formants, and in such a case combining both these methods for the analysis seems important. (See Figure 1.)

2.2. Formant Frequencies, Band-Width Variations

It is well-known that speech is not stationary.Therefore, no speech segment is the same as any other segment, even if they are adjacent. There is thus always also a "movement" of formant bandwidths along the frequency axis of the spectrum. It is hard to decide just by an FFT program output what the really important formant areas are (besides formant peaks), as the final output of a speech signal analyzed by an FFT analysis program is a series of harmonics along the spectrum which are effected by the input signal In this case, then, the LPC program may be more suitable because it presents formants including the full bandwidth covered by each formant. Thus, even if there are some local peaks within this formant band area,it can become clear that they are not individual formants but part of a specific frequency domain. This presentation is advantageous over spectrographic outputs, where formant limits are not clear and formant centers (their peaks) are not accurate. An LPC program may also provide the precisely calculated point of a formant peak frequency.

2.3. Amplitude Features

As mentioned,the FFT program calculates signals including both their source and filter while the LPC program calculates only filter features, namely the formants. Thus, formant amplitudes are more accurate in the LPC program, although formant amplitudes of an FFT program output seem to be more conspicuous, due to some energy gain values of the voice source. As a matter of fact, for the hearing system the whole

formant range is important rather than a single peak-frequency, which even more justifies the use of LPC for speech sound analysis.

2.4. Sex-Dependent Phonetic Features of Native Speakers of Hebrew

It is likewise well-known that for the same phonemes there are different formant values, which depend on the speaker's sex(and the physical structure of the vocal tract). Such differences may occur also in F2 and F3 (which in usual spectrograms are hard to see) and in relative amplitudes of each formant. Sex-related differences were found, for instance, in the pronunciation of the vowels /o,u,a/ by some speakers (|2|) and /ḥ,x/ (|4|).

2.5. Fricative Features

In Hebrew as in many languages, there are fricative phonemes.Some are more common than others, as in other languages (e.g., /f, s, sh/), and some are less common (the laryngeal and velar /h,ḥ,x/). These sounds are hard to analyze accurately because of the large amount of noise involved in their articluation and the lack of voicing, and traditional sound spectrograms yield rather blurred printouts of such phonemes. In this case, then, the FFT analysis seems again to be of less value than the LPC program which yields well defined formant domains (|4|, |5|).

3. CONCLUSIONS

Many speech analysis techniques exist now, relying on various theoretical approaches and algorithms. It seems useful to find the merit of each method in order to extract the best results of all of them in order to fully understand speech structure. Combining various analysis methods, more insight may be gained as to many problems that still exist in this field. The few examples shown here represent clearly this viewpoint concerning language-specific and general(universal) phonetic issues.

4. REFERENCES

(1)Aronson,L.(1990) Electrical Parameters of Cochlear Prosthesis Adapted to Deaf Hebrew Speakers,D.Sc.thesis, School of Medicine, Technion (supervisers: Prof. G.Rosenhouse and Prof. L. Podoshin)

(2)Aronson,L., J.Rosenhouse, G.Rosenhouse and L.Podoshin, "An analysis of Modern Hebrew Vowels and Voiced consonants" (submitted for publication)

(3)Iivonnen,A. (1987) "A Set of German Stressed Monophthongs Analyzed by RTA, FFT and LPC," in: Channon,R. and L. Shockey (eds.) In Honour of Ilse Lehiste, Dordrecht, Holland/Providence, U.S.A., pp. 125-138.

(4) Rosenhouse, J. (1989) "Issues in Hebrew Phonetics: the Articulation of Ḥ/X and Tsere",Proceedings of the 10 -th World Congress of Jewish Studies, Division S, vol. I, pp. 125-132 (in Hebrew)

(5) Rosenhouse, J. "Two Unstable Phonemes in Israeli Hebrew and Colloquial Arabic :'Aleph and 'Ayin"(to appear in the volume In Honour of Prof. W. Leslau, 1991).

(6) Woods, S. (1987) "The Precision of Formant Frequency Measurement from Spectrograms and by Linear Prediction", Abstracts of Papers, The 3rd Swedish Phonetics Symposium,Fonetik-89, in: Speech Transmission Lab. Quarterly Progress & Status Report,April 1989,Stockholm, R.I.T. pp. 91-93.

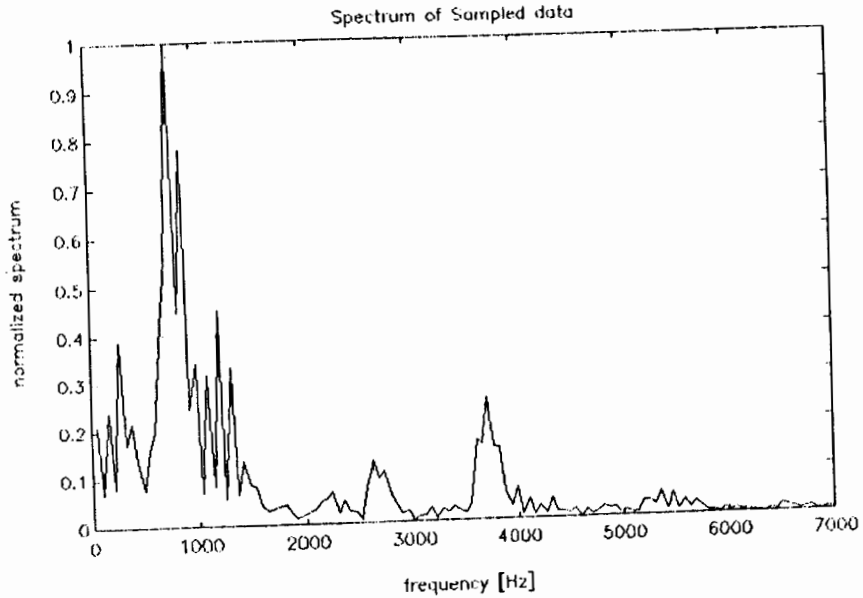


Figure 1. FFT analysis of /ʿ/ as uttered by a male native speaker of Hebrew (cf. |5|)

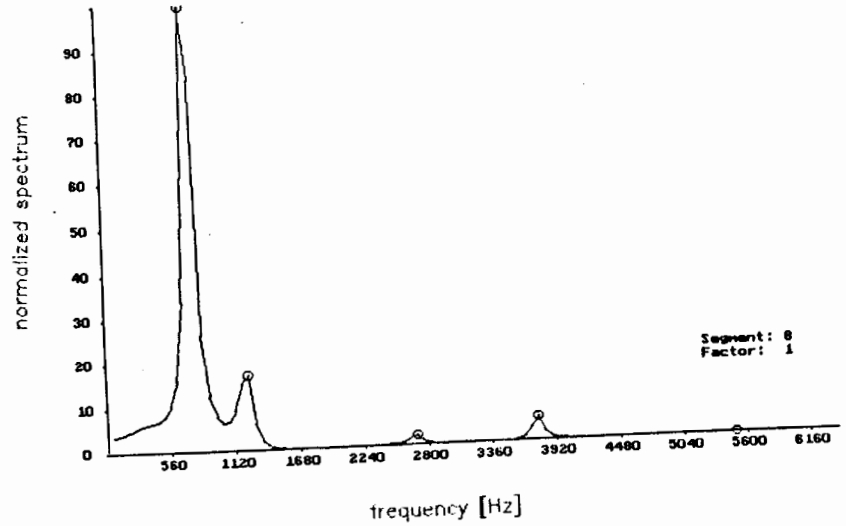


Figure 2: LPC analysis of /ʿ/ (same segment as Figure 1) as uttered by a male native speaker of Hebrew (cf. |5|)

LA VITESSE D'ARTICULATION ET LES UNITÉS SONORES DANS LA CHAINE PARLÉE

M. Dohalská-Zichová

Institut de Phonétique, Université Charles, Prague.

ABSTRACT

The carefully arranged material represents a set of spontaneous annoucements in natural working communication uttered continuously and with ever increasing speed until the articulation maximum is reached.It is used to examine the phenomena affecting intelligibility. A preliminary aural analysis is followed by a spectrographic study with special focus on the problem of linear and non-linear shortening of sounds, groups and possibly the influence of stress. The results of our research will be used to further improve the synthetic speech signal in Czech.

Les résultats de nos recherches précedentes concernant l'intelligibilité de la parole spontanée sur différents lieux de travail ainsi que la perfection de la qualité de la parole synthétique du tchèque ont inspiré la réalisation de l'expérience suivante.

Le point de départ de cette recherche représente un choix de 50 phrases tirées de la communication spontanée des réseaux de communication sur les différents lieux de travail. Ces 50 phrases ont été enregistrées par 8 speakers à qui on a demandé d'articuler assez distinctement pour que les phrases soient compréhensibles dans un débit accéléré, mais de garder en même temps la prononciation naturelle d'une communication quotidienne courante. Chaque phrase a été prononcée d'abord dans un débit"ordinaire"et, après une petite pause, le locuteur a répété plusieurs fois la phrase en augmentant continuellement le débit de l'articulation jusqu'au point où il ne peut plus répéter nettement la phrase. Ce type de réalisation sonore, tout a fait naturelle et différente de chaque accélération "mécanique",nous a permis de nous poser plusieurs questions concernant, d'une part, les problèmes de perception et, de l'autre, les changements dans l'image spectrographique.

Dans la première phase de nos investigations, nous nous sommes concentrés d'abord à l'analyse auditive (collective et individuelle). Il nous intéressait de savoir quel type de changements les auditeurs perçoivent au cours de l'augmentation de la vitesse de la parole.

Sur la base de sondages préliminaires,nous avons fait un enregistrements spectrographique représentatif (le matériel sonore dure plus de 80 min., les enregistrements spectrographiques représent plus de 10 000 données). Ici, je voudrais remercier, M. M. Ptáček et Mlle P.Žáčková pour leur fructueuse collaboration. Nous avons voulu confronter d'abord les matériaux spectrographiques avec les acquis tirés de l'analyse auditive et nous nous sommes posés encore d'autres questions, comme p. e.:

- quelle est l'influence de l'accent sur la durée des voyelles?
- est-ce qu'il y a un parallèle dans l'abréviation des voyelles brèves et longues? (La longueur des voyelles en tchèque est fonctionnelle!)
- comment se reflète l'augmentation de la vitesse du débit dans les petits groupes rythmiques de la phrase?
- en quoi la forme sonore de la phrase prononcée le plus lentement est différente de celle qui est prononcée le plus vite?

Comme nous connaissons les travaux remarquables de Klatt, Umeda, Huggins et d'autres, comme les analyses fondamentales ont été réalisées sur le matériel du tchèque soutenu par M.Maláč et ses collaborateurs et comme, nous-mêmes, nous nous ocupons de l'analyse spectrographique liée à la perception depuis longtemps, nous avons espéré que les sonogrammes seraient analysables facilement aussi bien par les méthodes statistiques ce qui nous permettrait de confronter, non seulement, les résultats de l'analyse auditive et spéctrographique, mais d'obtenir aussi de nouveau résultats contribuant au perfectionnement de la qualité de la synthèse de la parole tchèque.

Au cours de notre travail,nous avons vu qu'on ne pouvait y appliquer les critères utilisés pour l'élaboration des matériaux "de laboratoire", car il s'agit plus ou moins de la parole naturelle ainsi que de l'accélération naturelle, ce qui pose de problèmes nouveaux.

S'il faut considérer les paramètres influençant la durée des sons, nous devons constater que même quand il s'agit du même locuteur, la voyelle ou consonne donnée ne se comporte pas de la même façon dans le même entourage, dans le même contexte, si on augmente le débit de la parole. Chez les différents locuteurs, on peut constater des changements similaires,différents, mais aussi contradictoires - et ceci dans les circonstances adéquates. P.e. /a/ accentué dans la syllabe initiale se montre ainsi (5 locuteurs, 3 réalisations accélérées continuellement; durée en "ms"):

55	40	40	45	65
55	40	**45**	40	45
40	40	**45**	40	40

Dans la syllabe accentueé au milieu de la phrase les changements dûs au débit sont les suivants:

45	50	40	40	50
45	40	35	30	**60**
45	40	30	**55**	40

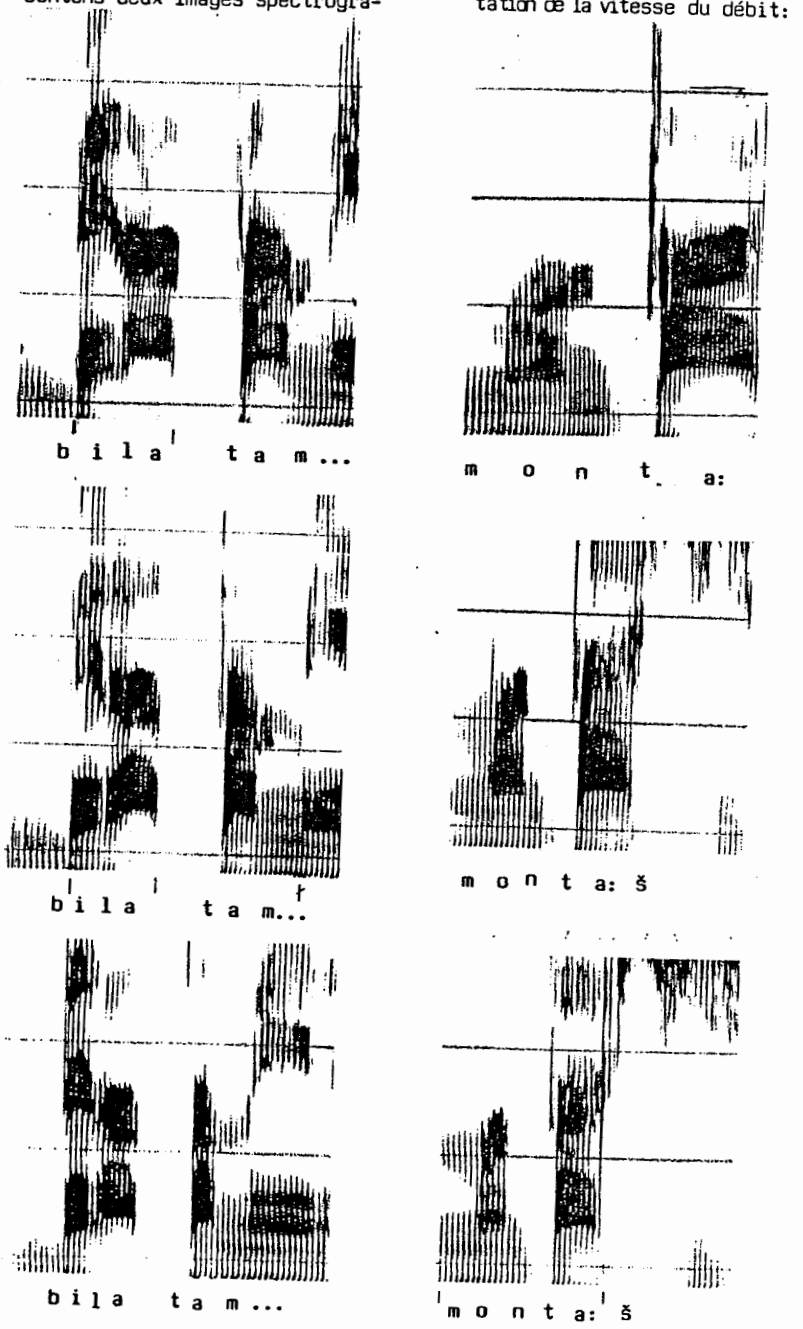
La relation du /i/ accentué et du /i/ inaccentué dans les syllabes successives se montre assez intéressante: dansla première réalisation non-accélérée le /i/ accentué est plus long ou semblable au /i/inaccentué: 60:45, 35:35, 30:30, 45:40, 40:40 -dans les deux réalisations accélérées suivantes, se manifeste, au contraire,l'abréviation du /i/ accentué et l'allongement du /i/ inaccentué; la relation du /i/ inaccentué par rapport au /i/ accentué est la suivante:

213% 125% 133% 225% 175%

En ce qui concerne la constatation suivante citée dans des travaux différents, par ex. (2). " les consonnes se comportent tout-à-fait régulièrement dans l'entourage vocalique et les lois compliquées commencent à partir de groupes de consonnes",nous ne pouvons confirmer que la deuxième partie de la constatation mentionnant les "lois compliquées" (si l'on peut en effet parler des "lois"?)

On fait pour l'instant une observation assez importante dans le domaine des pauses pré-explosives. Le petit changement des pauses pré-explosives (une légère abréviation et même un léger allongement) se montre sur les images spectrographiques, par rapport aux autres éléments relativement stables. Les chifres suivants présent (en %) la relation entre la première réalisation et les deux réalisations accélérées suivantes(résumé des réalisations

Comme illustration, nous présentons deux images spectrographiques reflétant l'augmentation de la vitesse du débit:



de 5 locuteurs) - possibilité de comparer avec la durrée des voyelles dans la même phrase:

a)Durée de la pause pré-explosive (en %).

/b/ 100,0	/t/ 100,0	/t/ 100,0
85,8	**100,6**	**100,8**
90,5	87,6	**105,8**

b) Durée de la voyelle /a/ (en%).

/a/ 100,0	/a/ 100,0	/a/ 100,0
85,8	78,0	87,0
80,0	71,0	73,0

c) Dans un autre phrase, les pauses pré-explosives sont les suivantes(/t/par rapport à /a:/).

/t/	/a:/	/t/	/a:/
100,0	100,0	100,0	100,0
84,5	68,2	75,8	73,3
78,6	57,3	70,6	54,7

Après une analyse plus en profondeur des matériaux, nous en arrivons à la constatation qu'il est préférable d'étudier toujours attentivement de petits ensembles en y examinant les phénomenes similaires ou contradictoires, attendus ou inattendus. Il est souvent très difficile de quantifier les mesures et les faits donnés, car de nombreuses compensations se manifestent ici, mais différemment chez plusieurs locuteurs, comme chez le même locuteur (dans les mêmes conditions).

Un travail statistique global effacerait les reliefs ainsi que les petits écarts contradictoires mais importants.Il est nécessaire de respecter une "microsegmentation"qui englobe aussi les phénomènes suivants: même dans la plus grande accélération on peut fixer sur les images spectrographiques (et ceci même en accord avec la perception) certaines unités stables, tandis que les autres unités vocaliques et consonantiques se groupent dans une "unité" nouvelle,où elle perdent petit-à-petit leur stabilité et leurs contours.

Outre les observations concernant la stabilité relative des pauses pré-explosives, nous pouvons constater que cette recherche en cours apporte de nouvelles remarques surtout dans le domaine de la description quantitative des changements non linéaires pendant l'accélération du débit (de la parole).Les faits suivants peuvent encore servir d'illustration intéressante. Dans plusieurs réalisations (les exemples ci-dessus) se manifeste une augmentation de durée de certains éléments, même si la phrase au total (ainsi que ses segments)est continuellement abrégée. Par ex.: la réalisation la plus rapide de la 4^{e} phrase représente chez le locuteur n^{o}1 une accélération de 46%, tandis que la pause pré-explosive devant le /t/ reste la même. Chez le locuteur n^{o} 2 l'accélération est de 36% et la pause s'allonge au contraire de 11%.Chez le locuteur n^{o}3, l'abrégement est de 23% et l'allongement de la pause pré-explosive de 12,3%.

En conclusion, nous aimerions encore mentionner que les matériaux choisis ici apportent souvent toute une série de suprises, surtout en ce qui concerne les découvertes de nouvelles relations et connexions dont nous avons choisi au moins quelques exemples typiques. Ces problèmes peuvent être traités plus largement au cours de prochaines discussions.

REFERENCES

/1/ DOHALSKÁ-ZICHOVÁ,M. (1989), "To the testing of the dependence of transients on the speaking tempo", 28. akustická konference, Bratislava, 186-189.
/2/ HUGGINS, A.W.P., (1972), "On the Perception of Temporal Phenomena in Speech",JASA,51,1271-1290.
/3/ KLATT, D.H.(1976),"Lingvistic uses of Segmental Duration in English", JASA,59, 1208-1221.
/4/ MALÁČ, V. (dir.) (1978), "Analýza časového členění věty", Tesla-VUST,, Praha.
/5/ UMEDA, N. (1977), "Consonant duration in American English", JASA, 61, 846-858.

AUTOMATIC EXTRACTION OF PHONETIC FEATURES IN SPEECH, USING NEURAL NETWORKS.

F. BIMBOT, G. CHOLLET, J.P. TUBACH.

Télécom Paris - Département SIGNAL, C.N.R.S. - URA 820.

ABSTRACT

We report in this paper a series of experiments aiming at automatically extracting phonetic features in speech, using a specific family of neural networks, namely TDNNs. The results show the interest that exists in using phonetic knowledge to guide speech recognition. The visualisation of connection weights inside the optimised networks illustrates the various strategies used by TDNNs for classifying sounds into features, after their acoustic content.

RESUME

Cet article présente un ensemble d'expériences visant à l'extraction de traits phonétiques à partir de signal de parole, à l'aide d'une famille particulière de réseaux neuro-mimétiques : les TDNNs. Les résultats montrent l'intérêt d'utiliser des connaissances phonétiques pour orienter la tâche des réseaux, en reconnaissance de parole. La visualisation des poids des connexions dans les TDNNs après optimisation illustre la façon dont ceux-ci utilisent les caractéristiques acoustiques des sons pour déterminer les différents traits.

INTRODUCTION

Connectionnist networks are one of the possible tools for performing classification tasks, such as those required in particular for speech recognition. Multi-layer neural networks are made of several layers of cells (or nodes), each of them delivering, as an output, a (usually) non-linear transform of the input; for instance a sigmoid. The input itself is obtained as the weighted sum of the activations of the nodes in the previous layer that are connected to the current node. A multi-layer neural network can be viewed as a black box made of a large number of elementary units with a rather simple individual function, the global behaviour of which makes it possible to model quite complicated non-linear transfer functions.

The number of nodes in each layer and the connection structure (i.e. what is called the architecture) is usually fixed a priori, whereas the values of each weight (what could be called the "furniture") are task specific and classically estimated by the back-propagation algorithm, given a set of training examples. In other words, it is possible, with neural networks, to automatically learn some non-linear discriminations between families of patterns, without any careful human time-consuming specification of classification rules. However, the computing time required is rather high.

THE TDNN

Waibel et al introduced TDNNs (Time-Delay Neural Networks) as a specific neural network architecture that can take into account the "dynamic nature of speech" [1]. Indeed, such a network is able to represent temporal relationships between successive acoustic time-slice (frames), which is a property of major importance, since static characteristics of the speech signal are certainly unsufficient for a proper identification of sounds or sound features. Moreover, TDNNs provide some invariance under time-translation, which lessens the sensitivity of the processing in front of unavoidable segmentation inaccuracies. Figure 1 illustrates the TDNN structure.

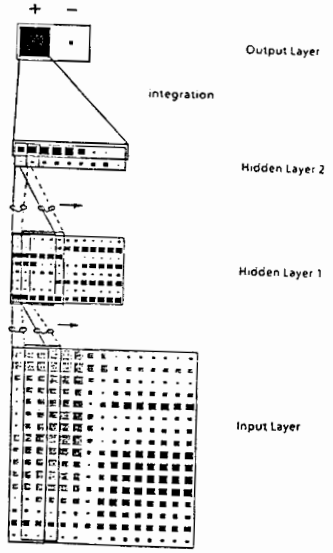


Figure 1 : A TDNN (adapted from Waibel).

TDNNs have 4 layers (an input layer, 2 hidden layers and an output layer). Successive layers are sparsely connected (in practice, some connections weights are set to zero). Moreover, there exist groups of weights that are constrained to be identical to one another, in order to warrant the time-shift invariance property. More detailed descriptions of TDNNs can be found in several papers [1] [2] [3].

In recent publications, TDNNs proved to be very efficient for tasks such as classifying [b] vs [d] vs [g] among a set of observation containing phonetic realisations of these 3 phonemes, in japanese [1], or in french [4]. Waibel et al. also showed that a modular all-phoneme classification network could be successfully designed on the basis of the collaboration of a collection of such elementary networks (one for [bdg], one for [ptk], one for [fsʃ], ...), provided an other network makes concurrently a rough classification of each sound, in order to decide which of these elementary networks is to be resorted to. In parallel to these developments, Haffner et al. proposed fast learning methods for TDNNs that bring down the training time to reasonable figures [2].

All these properties make of TDNNs appealing tools for classifying speech sounds, with the ultimate goal of speech recognition. Our approach of phonetic features extraction is slightly different from the modular all-phoneme recognition process proposed by Waibel et al. For the latter, a hierarchical decision is needed, whereas, for the former, several parallel classifications are made and then integrated in a final phoneme identification.

PHONETIC FEATURES

The phonetic description of speech events is classically based on distinctive features. Features are usually binary, since they indicate the presence or absence of a specific characteristic for a given sound (or family of sounds).

Features can be defined according to acoustical, articulatory, or even linguistic properties. For instance, the *grave / acute* opposition is based on acoustical considerations, while *front / back* is articulatory and *vowel / consonant* based on higher-level linguistic concepts.

Each phoneme (or, more precisely in our case, each phone) results from the simultaneous realisation of several elementary binary features, which Trubetzkoï describes as a "Korrelazions-bündel" (i.e. a bundel of correlations), and which Jakobson qualifies as a "bundle of concurrent binary distinctive features" [5].

Several systems of features have been proposed, among which must be mentioned those of Malmberg [6] and Rossi [7] for the french language, those of Jakobson [5] and Chomsky [8], in english. Jakobson's classification was the first one to rely on acoustical criteria only.

A few speech recognition systems have been calling for feature extraction (or macro-classes identification) strategies, most of them using expert systems that track cues on the acoustic signal, with the help of rules.

The approach using neural networks can te understood as a mean of expressing and modeling, with a mathematical non-linear tool, the (sometimes) complex relationship existing between the abstract notion of phonetic feature and its physical manifestation through acoustic cues.

PROTOCOL AND CORPUS

We have been evaluating the use of TDNNs for different feature extraction tasks, for the french language.

A set of 13 binary features was thus designed, relying on the most classical phonetic oppositions. This set allows a total description of the french phonetic system. It is however not minimal, but our goal was to investigate what type of oppositions TDNNs are best adapted to.

An other set of 6 discriminative and minimal "random" features was artificially built, so that it opposes 34 phonemes with one another. This set has, of course, no phonetic background, since phonemes that have "nothing to do" with each other are member of the same random class and opposed to an other (complementary) class of phonemes that have "nothing to do" with each other either ! These artificial features however serve as a point of comparison for judging the relevance and the usefulness of classical phonetic oppositions versus arbitrary ones.

Other sets or "ternary" features have also been used [3], but results are not reported in this article.

The corpus for the experiments contains 200 french phonetically balanced sentences [9], uttered by one male speaker in excellent recording conditions and sampled at 16 kHz. Each phoneme realisation was coarsely labeled at the center by hand, and represented by 16 time frames (8 on each side of the label) of 16 Mel-scaled filter-bank coefficients; in other words, a quite low resolution (256 pixels) spectrographic representation, with non-linear frequency scale. The corpus contains 5270 tokens (unbalancedly dispatched between the phonetic classes), which were halfed between a learning set and a test set.

The corpus was transcribed prior to its production and labeled according to this normative transcription, using 34 symbols :

i	y				u							#						
	e	ø			o													
		ɛ	œ	ɔ								ɛ̃		ɔ̃				
			a											ɑ̃				
p	t	k		b	d	g					j	ɥ		w				
f	s	ʃ		v	z	ʒ												
				m	n	ɲ							l	ʁ				

The phonetic system used in our experiments. [#] denotes "silence". [ə] = [œ].

Automatic phonological rules were applied to modify the features of some phonemes, to account for basic contamination effects.

Evaluations were done on the whole corpus, on the sub-corpus of vowels only (2952 items) and on the sub-corpus of consonants only (2318 items). Sub-corpuses were also halfed in training and test sets.

The cross-validation strategy was used to decide when to stop the learning phase [3].

RESULTS

Performances on the whole corpus, on the vowels-only corpus and on the consonants-only corpus, for each feature, are shown in Table 1 below. Two scores are given for each experiment : one (in bold) is the score of correct feature identification on the test set (i.e. for the the data that were not used to train the network), while the second score (in standard characters) corresponds to the score obtained on the training set itself (self-consistency). Naturally, the second score is usually higher than the first one, and the difference between the two gives a hint of the generalisation capability of the network; that is its ability to solve a similar problem to the one it was trained for, with data that it has not seen yet.

On the whole corpus, all scores (but one) range between 90 % and 99 %, with more than half of them over 95 %. Moreover, all scores for phonetic features (but one) are higher (by approximately 10 %, in the average) than the average score for arbitrary random features. This clearly evidences the contribution of phonetic knowledge for determining what kind of task the TDNN is most likely to work with. Beside this, while features related to manner of articulation are very efficiently detected (features III, V, XI, XII, XIII), those linked to the more abstract notion of place of articulation provide less satisfactory results.

Not all phonetic features that were tested on the whole corpus were also experimented for the vowel and the consonant sub-corpuses, but only some of them that allow a discriminative non-redundant system. The improvement of feature extraction when moving from the whole corpus to the consonants-only corpus is rather disappointing, but must be owed to a change of the repartitions. Conversely, scores for vowels-only improve significantly, in general. In both cases, the self-consistency tends to increase, since the number of learning examples is smaller and makes it possible for the TDNN to memorise the particularities of the training data (which is clearly undesirable).

VISUAL EXPLORATION

It can be shown that, under certain constraints, the matrices of weights within the first layer of TDNNs can be viewed as typical patterns that are searched for in the spectral picture of the input token to classify [10]. In other words, TDNNs develop in some ways their own expertise for classifying speech sounds, a little bit like a human expert would do, from experience.

In figure 2 (last page), we have visualised 3 sets of weight matrices, for 3 features : *voiced / unvoiced*, *nasal / non-nasal*, and *vowel / consonant*. A full comment is given with the figure : the cues used by the network are most of the time in accordance with the classical acoustic descriptions of phonetics, and thus directly interpretable, which was not at all a priori warranted.

CONCLUSION

Phonetic knowledge can thus be used to help TDNNs in their task : not only to a priori choose the kind of task that is the most likely to be successfull, but perhaps also to initialise the weights of the network using human expertise on the problem to be solved. This last point is a challenging topic for further research.

Conversely, TDNNs can learn automatically from a set of typical examples (like other neural networks); but because their architecture is speech dedicated, they certainly represent a new tool for phoneticians' investigations.

REFERENCES

[1] WAIBEL et al : Phoneme recognition using Time-Delay Neural Networks, *IEEE-ASSP, vol 37, n° 3, 1989.*
[2] HAFFNER et al : Fast back-propagation learning methods for large phonemic neural nets, *Eurospeech 89.*
[3] BIMBOT : Phonetic features extraction using Time-Delay Neural Networks, *ICSLP 90.*
[4] DEVILLERS et al : Reconnaissance monolocuteur des phonèmes du français au moyen de réseaux à masques temporels, *XVIIIèmes JEP, 1990.*
[5] JAKOBSON et al : Preliminaries to speech analysis, *M.I.T. Press, 1951.*
[6] MALMBERG : Structural linguistics and human communication, *Springer, 1967.*
[7] ROSSI : Les traits acoustiques, *La Linguistique, vol 13, fasc 1, 1977.*
[8] CHOMSKY et al : The sound pattern of english, *Harper & Row, 1968.*
[9] COMBESCURE : 20 listes de 10 phrases phonétiquement équilibrées, *Rev. d'Acoustique, n° 56, 1981.*
[10] BIMBOT et al : TDNNs for phonetic features extraction : a visual exploration, *ICASSP 91.*

phonetic feature	whole corpus	vowels only	consonants only
I vowel vs non-vowel	**95.8 %** (96.2 %)	-	-
II vocalic vs non-vocalic	**96.5 %** (97.0 %)	-	-
III voiced vs unvoiced	**98.9 %** (99.4 %)	-	**98.8 %** (100 %)
IV sonant vs non-sonant	**96.9 %** (98.4 %)	-	-
V nasal vs non-nasal	**97.7 %** (99.5 %)	**97.7 %** (99.6 %)	**98.2 %** (99.7 %)
VI grave vs acute	**90.6 %** (96.5 %)	**95.7 %** (99.8 %)	-
VII extreme vs central	**84.4 %** (92.1 %)	**87.5 %** (96.6 %)	-
VIII compact vs diffuse	**91.7 %** (96.0 %)	**94.0 %** (97.8 %)	-
IX rounded vs unrounded	**94.8 %** (97.5 %)	**93.2 %** (94.5 %)	-
X bemol vs non-bemol.	**90.1 %** (98.8 %)	-	-
XI delayed vs non-delayed	**97.1 %** (98.7 %)	-	**89.4 %** (99.1 %)
XII discontinuous vs cont.	**97.9 %** (99.1 %)	-	**93.9 %** (97.6 %)
XIII fricative. vs non-fric.	**97.3 %** (99.9 %)	-	**95.1 %** (99.9 %)
Average for 6 random features	**85.3 %** (92.1 %)	-	-

Table 1 : Scores for feature extraction on the 3 corpuses : scores on test set, (on training set).

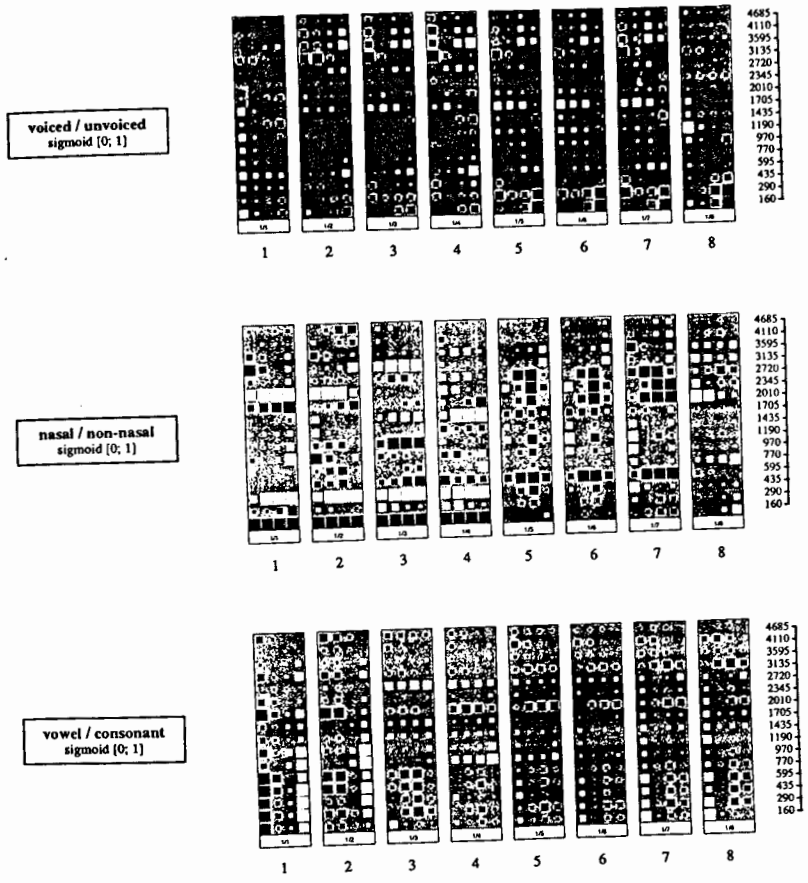


Figure 2 : Visual display of weight matrices (masks) in TDNNs for phonetic features extraction. (top : voiced / unvoiced, center : nasal / non-nasal, bottom : vowel / consonant).

The side of each square is proportional to the magnitude of the weight. Black corresponds to positive values, white to negative. Only the first layer of weights is represented, as 8 masks (4 x 16). Time is horizontal, frequency vertical (Mel-scale).

[+/- voiced] : All 8 masks are quite similar to each other, which evidences that the network is over-dimensioned for the task. The presence of energy in the first 2 bands (160 Hz and 290 Hz) is clearly used as a cue for identifying voiced patterns, especially for masks 5, 6 and 7. But a temporal decrease in high frequencies (3135 Hz and 4685 Hz bands), jointly with an increase in low frequencies is also an indication of [+ voiced] (masks 2, 3, 4, 7 and perhaps 5 and 6). This is certainly owed to phonetic combinations such as unvoiced plosive / fricative + voiced sound, for which the sudden change of the spectral tilt evidences the beginning of voicing.

[+/- nasal] : Masks show here again some redundancy. Masks 1, 2, 3 and 4 underline the significant role played by the joint presence of low frequencies in the 160 Hz band (all nasals are voiced...) and the absence (white weights) of "medium-low" frequencies (around 435 Hz). This is fully consistent with the observations concerning spectral zeros around the medium-low frequencies, for nasal sounds. These 4 masks differ mainly by the location of an other spectral hollow (2345 Hz for masks 1 and 2, 3135 Hz for mask 3, 1705 Hz for mask 4). Mask 5, 6 and 7 are very similar with one another. They search for joint energy around 595 Hz and between 2010 and 2720 Hz. This may correspond specifically to nasal vowels that usually have a first formant in [500 Hz - 700 Hz] and a high third formant in [2300 Hz - 2800 Hz]. The role of mask 8 is not clear yet (no energy around 770 Hz nor 2010 Hz).

[+/- vowel] : Here, masks 1 to 8 are ordered according to a noticeable progression : from left to right, a pattern of important energy somewhere between 290 Hz and 770 Hz (region covering roughly all possible first formants) can be approximately retrieved in all masks, with a different time-shift. Note that the lowest band is not really used, because it is ambiguous with voicing. Note also that an absence of energy around 2720 Hz (just beyond the theoretical maximum second formant for [i]), and the presence of energy around 3135 Hz (region for several third formants) are both considered as in favour of [+ vowel].

ACOUSTIC PROPERTIES AT FRICATIVE-VOWEL BOUNDARIES IN AMERICAN ENGLISH

L. F. Wilde and C. B. Huang[1]

Research Laboratory of Electronics and Department of Electrical Engineering and Computer Science,
Massachusetts Institute of Technology, Cambridge, MA, USA

ABSTRACT

Previous work has shown that acoustic properties which signal place of articulation and voicing for fricative consonants can be located at the fricative-vowel boundary. Therefore, we focused on the region within 30 ms of the boundary in our search for acoustic regularities.

Our goal is to better understand the perceptual salience of these acoustic cues. In this paper, results will be presented from perceptual tests with natural and synthetic CVs. As expected, the non-strident fricatives were most often confused and the performance for synthesized fricatives was poorer than that for natural speech. Acoustic evidence for these results is examined.

1. INTRODUCTION

Previous acoustic analysis of natural speech has shown that acoustic properties in the vicinity of fricative-vowel boundaries can be associated with cues for consonant perception. The following acoustic attributes are associated with fricative production: (1) an interval of frication noise with a spectrum that is shaped by the location of the constriction in the vocal tract; (2) formant transitions into adjacent vowels that provide additional place of articulation information; and (3) details in the transition from noise production to voicing onset which signal the distinction between voiced and voiceless fricatives. [4].

Our goal is to better understand the acoustic properties of the fricative-vowel boundary and, particularly, their perceptual relevance for place. Describing fricative-vowel boundaries is an especially interesting problem because these occur between continuous sounds produced by different source mechanisms: the supraglottal source, frication noise, which is generated as air flows through a narrow contstriction in the vocal tract, and the two glottal sources, voicing and aspiration.

Fricative synthesis provides a means for systematically examining the relative timing of the different sound sources. By comparing perceptual and acoustic measures for natural and synthetic stimuli, we can evaluate the adequacy of existing rules for modelling fricatives. We performed a series of listening tests to provide a baseline for intelligibility of natural fricatives and fricatives produced by a high-quality speech synthesizer. We also examined the acoustic properties of these stimuli to determine which differences could account for observed deficiencies in intelligibility.

2. PERCEPTUAL TESTS

2.1 Objective

Identification tests were run with natural CV speech tokens as stimuli to provide a baseline measure of intelligibility of fricatives. We used one of the best existing rule-based, text-to-speech synthesizers available to obtain corresponding synthetic stimuli. Nevertheless, we expected that the identification of the synthesized speech would be more difficult.

[1]Presently affiliated with Dragon Systems, Inc., Newton, MA, USA.

2.2 Method

The natural stimuli were CV tokens excised from C'VCVC'VC nonsense utterances spoken by one male speaker, Dennis Klatt [4]. The bandwidth of these utterances, which were digitized at a 10 kHz sampling rate, corresponds to the bandwidth used for synthesis of male speech. The C was one of the eight English fricatives, which can be classed according to place: labiodental (/f/, /v/), dental (/th/, /dh/), alveolar (/s/, /z/), and palatal (/sh/, /zh/). The V was one of four American English vowels (/iy/, /eh/, /aa/, /uw/), chosen to be representative of front, back and rounded vowels. The vowels were truncated 40 ms after vowel onset, which was defined as the beginning of the first identifiable pitch pulse for voiceless fricatives and the point where voicing amplitude increases abruptly for voiced fricatives.

Corresponding CV stimuli were synthesized using the phoneme input mode of KLATTALK, a research version of Klatt's text-to-speech system. The KLATTALK algorithm for formant transitions begins by looking at the segment following the fricative. Values from previous trial and error matching of natural frication spectra were used to optimize table values for synthesis of frication [4]. We used parameter values that Klatt chose to model his own voice.

The voiceless (/f/, /th/, /s/, /sh/) and voiced (/v/, /dh/, /z/, /zh/) fricatives were presented in separate identification tests in each of the natural and synthetic conditions. Five repetitions of the 20 distinct stimuli in each test were presented in random order over headphones in a sound-treated room. Five phonetically trained listeners acted as subjects. Subjects were asked to identify the fricatives by making a forced choice among the four possibilities. No responses regarding the vowel identities were required and voiced-voiceless distinctions were not explicitly examined.

2.3 Results

All of the natural and synthetic strident fricative tokens were identified correctly (0 errors out of 800 total responses). The non-strident fricatives were most often confused (175 errors out of 800 responses). Figure 1 compares the percentage of errors made on the natural and synthetic non-strident fricative CVs in the voiceless and voiced tests.

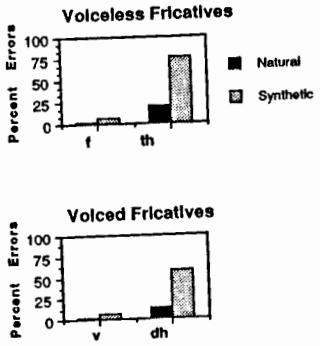


Figure 1. Error distribution, combined across vowel contexts, as a percentage of 100 responses (5 repetitions X 4 vowels X 5 listeners)

As expected, the performance for synthesized fricatives was consistently poorer than natural speech. These results also show that the predominance of errors was for the synthetic /th/ and /dh/ tokens.

3. ACOUSTIC ANALYSIS

3.1 Objective

We examined the acoustic properties of the natural and synthetic stimuli to determine which differences could account for the observed deficiencies in intelligibility. Following Klatt[4], we focused on the following attributes involved in moving from a frictive to a vowel: the evidence of the changing sound sources (voicing, frication, aspiration) and the onset frequency of formants (F1, F2, F3).

3.2 Method

All measurements were performed with the set of tools for speech analysis available on the MIT Speech Vax cluster. The formant onset frequencies were measured at the first identifiable pitch pulse. Discrete Fourier transforms were calculated with a 6.4 ms Hamming window that was carefully placed in order to maximize inclusion of the closed portion of waveform.

The noise spectra were calculated over a 30 ms Hamming window. Spectra were computed at 10 ms intervals from consonant onset to the fricative-vowel boundary. We compared relative amplitudes of noise in different frequency regions to vowel formant amplitudes.

3.3 Results

In view of perceptual test results, the acoustic findings for the non-strident fricatives only are presented here.

A close correspondence was found between natural and synthetic tokens for the first three formant onset values. As previously seen in the high front vowel context, F2 onset formant frequencies contradict the general rule that formant frequency is always lowest for labial place of articulation [4]. Figure 2 illustrates the F2 formant onset frequencies for voiceless natural and synthetic fricatives.

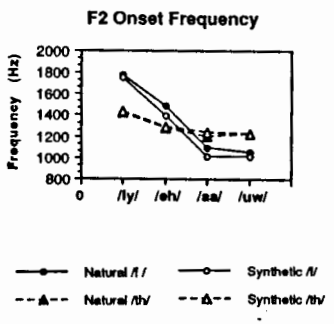


Figure 2. F2 onset frequencies for voiceless fricatives /f/ and /th/ in each vowel context.

Evidence of aspiration, if present, was usually found within 20 ms before the vowel onset. Aspiration may be distinguished from frication by an F2 prominence in the noise spectrum continuous with the formant at the fricative-vowel boundary. Aspiration was indicated, according to this criterion, for all the natural /f/ stimuli. Aspiration was not seen for the natural /th/ stimuli, except for /thuw/. The synthesized stimuli did not include any aspiration in the parameter specifications.

In examining the noise spectra prior to the fricative-vowel boundary, we concentrated on the non-strident token pairs whose difference in intelligibility between the natural and synthesized stimuli was largest (/feh/-/theh/ and /veh/-dheh/). The natural /f/ and /th/ differed most noticeably in their average amplitudes, as compared between the approximate noise amplitude in the 2000-3000 Hz region 70 ms before vowel onset relative to the amplitude of F3 in the vowel (2500 Hz). The natural /f/ was 20 dB lower, whereas /th/ was 30 dB lower than the vowel. The synthetic /f/ was 25 dB weaker while the /th/ was only 20 dB weaker. Time-varying spectral amplitude was observed for the natural /f/, which increased by approximately 10 dB from its beginning to the vowel onset, whereas the spectral amplitude for the natural /th/ appeared constant throughout its duration.

The spectral characteristics of the natural /f/ and /th/ tokens were found to be otherwise similar to each other, consistent with findings in previous work with a larger number of tokens and speakers [2]. The synthesizer models the spectrum of the non-strident fricatives as Gaussian noise with no formant structure. The spectra of the synthesized /f/ and /th/ were tilted, emphasizing the low frequencies with a 6 dB roll-off, whereas the natural stimuli were flat.

The spectra for the natural voiced non-strident fricatives (/v/ and /dh/) were very different from those of the synthesized stimuli. In the natural stimuli, the formant structure extended far into the fricative and noise excitation coexisted with essentially vocalic-looking formant structure. In constrast, there was no region with strong formant structure in the synthetic voiced fricative; instead there was an abrupt change between noise excitation and prevoicing by the glottal source. For /dh/, the onset of the vowel from the prevoiced region was abrupt enough to appear stop-like.

4. DISCUSSION

All of the strident fricatives, which are characterized by a relatively high spectrum amplitude as compared to the adjacent vowel, were identified correctly.

Harris [3] found that the frication noise provides the dominant cue for discriminating /s/ and /sh/, but that formant transition cues dominate the differences in noise spectra for /f/ and /th/.

The need to further investigate distinctions between the labiodental and dental fricatives is highlighted by the current perceptual and acoustic results. The F2 formant frequency onsets in front, back and rounded vowel contexts in the present database illustrate that listeners may adapt to regularities in formant onset frequencies, even if these present unexpected patterns. One possible explanation for the lower F2 formants for /th/ as compared to /f/ before high front vowels is that for labiodentals, the tongue is freer to move in anticipation of the following vowel.

The current findings suggest that even if the formant transitions are reproduced accurately, as in the Klattalk stimuli, deficiencies in intelligibility for synthetic stimuli remain. This implies that further consideration of noise amplitude and shape is needed. While the natural /th/ was 5-10 dB weaker than /f/ relative to the vowel, the synthetic /th/ was too strong and the /f/ too weak to maintain this distinction. This difference could partially explain confusions between synthetic non-strident tokens.

Acoustic variations can be interpreted with respect to existing production models [1] and predictions regarding the interaction and relative amplitudes of frication noise, aspiration, and voicing as constriction sizes vary in time [6]. Our analysis results for the natural and synthetic stimuli suggest the need to better model these source changes between the vowel and the fricative.

In some voiceless fricatives, aspiration can lead to a smoother transition at the fricative-vowel boundary. The role of aspiration in analysis, synthesis and perception of voice are described in Klatt and Klatt[5]. We intend to use the KLSYN88 formant they describe, which provides more flexible control over the glottal source, to continue to model the acoustic characteristics we observed. We can then test if the extra formant structure present in aspiration may provide place-of-articulation information and thus enhance both intelligibility and naturalness.

After synthesizing new tokens with our modified rules, we plan to evaluate how the addition of these rules affects the intelligibility and naturalness of synthetic fricatives. We already observed that subjects can easily classify the CVs used in the present study as natural or synthetic, even with very short vowels. Finally, we must investigate additional speakers and consider higher frequency cut-offs to determine whether the phenomena we observed are typical.

5. ACKNOWLEDGEMENTS

This work was supported in part by a grant from the National Institute of Neurological and Communicative Disorders and Stroke and the National Science Foundation, no. DC0075.

6. REFERENCES

[1] Badin, P. and Fant, G. (1989), "Fricative production modelling: aerodynamic and acoustic data. *Eurospeech 89 Conference, Paris. Vol. 2.*

[2] Behrens, S. J. & Blumstein, S. E. (1988), "Acoustic characteristics of English voiceless fricatives: A descriptive analysis", *J. Phonetics , 16,* 295-298.

[3] Harris, K. S., (1958), "Cues for the discrimination of American English fricatives in spoken syllables. *Lang. Speech 1, Part 1,* 1-7.

[4] Klatt, D. H. (Unpublished manuscript), "Fricative consonants"

[5] Klatt , D. H. and Klatt, L. C. (1990), "Analysis, synthesis, and perception of voice quality variations among female and male talkers", *J. Acoust. Soc. Am, 87(2,* 820-857.

[6] Stevens, K. N. (1987), "Interaction between acoustic sources and vocal-tract configurations for consonants", *In Proceedings of 11th International Conference of Phonetic Sciences, Estonia, Vol. 3,* 385-389.

JUSTIFICATION PERCEPTIVE DU SPECTROGRAPHE AUDITIF

Christophe d'Alessandro & Denis Beautemps

LIMSI-CNRS BP133-91403 Orsay Cédex, France.
ICP-INPG 46, avenue Félix-Viallet 38031 Grenoble Cédex, France.

ABSTRACT

An auditory spectrograph is presented and discussed, which is quite different from the initial proposal of [4]. A set of descriptive acoustic parameters are derived from speech signals analysed according to the time and frequency resolution caracteristics of the spectrograph: sinusoids in the area of F0 and F1, dominant frequencies and envelope modulation above about 1 kHz, for voiced speech. These parameters are used in an analysis/synthesis system which delivers a synthetic signal perceptively equivalent to the original signal. This preliminary work demonstrates the possibility of using alternative auditory-based acoustic parameters for speech synthesis and analysis instead of production-based acoustic parameters.

1 Introduction

L'avènement du spectrographe a permis un développement considérable de l'étude phonético-acoustique descriptive de la parole. Deux raisons conceptuellement distinctes ont contribué à ce succès: 1. le spectrographe permet l'observation de corrélats acoustiques importants du mécanisme de production de la parole, comme les formants, la vibration des cordes vocales; 2. il existe une analogie entre analyse spectrale à court terme et analyse du signal par le système auditif périphérique: les objets visibles sur un spectrogramme correspondent à des caractéristiques acoustiques perceptivement pertinentes. Les relations entre ces deux aspects ne sont pas toujours clairement considérées, et des notions qui relèvent du modèle acoustique de production, comme les formants ou l'onde de débit glottique s'avèrent d'une importance variable, voire contestable, du point de vue perceptif [3]. Nous pensons ainsi que le spectrographe a contribué à exagérer l'importance de paramètres acoustiques qui ne sont pas perceptivement pertinents.

La modélisation de l'analyse du signal par le système auditif périphérique a conduit à proposer de nouvelles représentations assimilables à la représentation spectrographique, sous forme de "spectrogrammes auditifs" [4], ou le "cochléogrammes" [5]. Le but de cet article est de discuter d'une forme de spectrographe auditif, et de justifier par une évaluation perceptive, en utilisant la resynthèse, les paramètres acoustiques auditivement pertinents qui apparaissent sur cette représentation.

2 Spectrographe auditif

Le spectrographe utilisé est présenté dans cette section, et confronté aux formes de spectrographes auditifs proposées précédemment. Les "spectrogrammes auditifs" de [4] combinent deux types d'informations spectrales: 1. l'énergie spectrale à la sortie d'un banc de filtres passe-bandes s'appuyant sur des échelles fréquencielles (Bark) et d'intensité (phon) auditives; 2. les fréquences dominantes dans chaque canal d'analyse, issues du modèle d'analyse temporelle DOMIN. D'autres auteurs [5], proposent comme représentation spectrographique la visualisation des signaux en sortie d'un banc de filtres auditifs. C'est une variante de cette seconde solution que nous avons adopté pour les figures 1 et 3. Ce tracé, équivalent au redressement simple alternance de signaux filtrés par un banc de filtres auditifs, est obtenu en portant le produit de l'amplitude par la phase principale (entre $-\pi$ et π), pour les phases positives seulement, d'une analyse par ondelettes sur une échelle auditive [1]. 250 filtres sont régulièrement répartis en échelle Bark, de largeur de bande constante 1 bark. Une échelle logarithmique est employée pour porter les amplitudes. La différence de lisibilité visuelle avec une échelle auditive Phon apparaît tout à fait négligeable. Nous avons préféré ce type de tracé à celui proposé dans [4] pour deux raisons:

1. Une erreur d'interprétation de la dualité temps/fréquence semble à la base du procédé de calcul utilisé pour les figures de [4]. Les spectrogrammes semblent (le procédé de calcul n'est pas explicite dans l'article) calculés par transformée de Fourier à court terme et visualisés en utilisant les échelles Bark et Phon. Alors seule la résolution spectrale est celle d'une analyse en échelle Bark, et augmente en raison de la fréquence centrale d'analyse. Par contre la résolution temporelle est manifestement fixe sur les spectrogrammes publiés, égale à celle d'une analyse spectrographique classique en bande large. 2. Il n'apparaît pas nécessaire d'introduire un modèle supplémentaire, comme DOMIN, pour rendre compte qualitativement (visuellement) de l'analyse temporelle: la visualisation de la phase d'analyse dans chaque bande permet de distinguer les fréquences dominantes, temporellement dans le grave du spectre, et au delà d'environ 1-1.5 kHz grâce à l'amplitude. Il faut ajouter que la resynthèse ou la modification du signal à partir de notre représentation est directe, ce qui est montré dans [1], mais ce qui est hors du propos de cette communication.

3 Justification perceptive

La lecture des spectrogrammes auditifs suggèrent un ensemble de paramètres acoustiques descriptifs, liés à la fois à l'appareil phonatoire et à la résolution spectro-temporelle du dispositif d'analyse utilisé. Les relations entre la résolution d'analyse, c'est-à-dire la largeur de bande effective des filtres auditifs, et les grandeurs fréquencielles types produites par l'appareil phonatoire sont résumées figure 4. L'abcisse représente les fréquence centrales d'analyse (en Bark), l'ordonnée les bandes passantes des filtres auditifs (en Hz) correspondant. Les deux courbes partagent le plan en deux zones: dans la zone intérieure aux deux courbes, deux fréquences pures présentes pour une fréquence centrale donnée ne sont pas distinguées par le filtre d'analyse; dans la zone extérieure deux fréquences pures sont distinguées par le filtre d'analyse. Si l'on applique un spectre harmonique, comme pour de la parole voisée, comportant un ensemble de raies spectrales équidistantes la courbe de résolution prédit le nombre de raies séparées, en fonction de la fréquence d'analyse. Deux situation types se manifestent: dans le cas A de la figure 4 (F0=100 Hz, fréquence d'analyse 200 Hz) le filtre auditif isole une composante spectrale et des sinusoïdes redressées en simple alternance sont représentées sur le spectrogramme; dans le cas B (F0=100 Hz, fréquence d'analyse 2000 hz), plusieurs raies sont intégrées par un même filtre. Dans ce cas, des battements sont visibles sur le spectrogramme. La période de la modulation d'amplitude du signal filtré, des battements, est l'inverse de la différence de fréquence entre les composantes, soit la période fondamentale 1/F0. Les battements donnent naissance à un signal possédant une fréquence dominante, obtenue par la moyenne les fréquences des raies spectrales pondérées par leurs amplitudes. Ainsi, lorsqu'un pic spectral (ou formant) est présent dans la bande d'analyse du filtre, la fréquence dominante due aux battements est approximativement égale à la fréquence centrale de ce formant. Une première dimension fréquencielle relève de la fréquence fondamentale. Une seconde dimension est donnée par l'espacement des formants. Lorsque la résolution des filtres d'analyse est plus fine que l'espacement entre deux formants, la séparation des harmoniques ou les battements formantiques se produisent: c'est la situation rencontrée respectivement pour F1, cas C et F2, cas D de la figure 4. Lorsque cette résolution diminue, ou lorsque plusieurs formants sont proches, les battement deviennent plus complexe, et une masse spectrale apparaît sur le spectrogramme. La modulation d'amplitude est plus rapide que la fréquence fondamentale, et la fréquence dominante peu saillante. Pour illustrer ce propos, les figures 1 et 3 présentent des spectrogrammes auditifs d'une voix féminine prononçant /wiski/ et d'une voix masculine prononçant /lopotiʃa/. Si l'on considère les parties voisées, pour les deux exemples, le grave du spectre est décomposé en harmoniques. Les amplitudes et phases de ces harmoniques dépendent d'une part de la source de voisement et d'autre part de l'influence du premier formant. Dans la région du second formant, l'intégration de plusieurs harmoniques dans un même filtre auditif provoque l'apparition de battements, avec une fréquence dominante et une modulation d'amplitude à la période fondamentale. Au delà d'environ 3 kHz pour la figure 1, deux formants s'agglomèrent et la fréquence dominante comme la période de modulation devient plus difficile à définir. Il est probable que cette masse spectrale est perçue par son centre de gravité, et que sa contribution se limite à des aspects non-linguistiques du signal, comme la brillance, le degré de souffle etc. La validité de ces observations peut être testée par un procédé d'analyse/synthèse. Un signal naturel est décomposé en utilisant les paramètres acoustiques précédents:

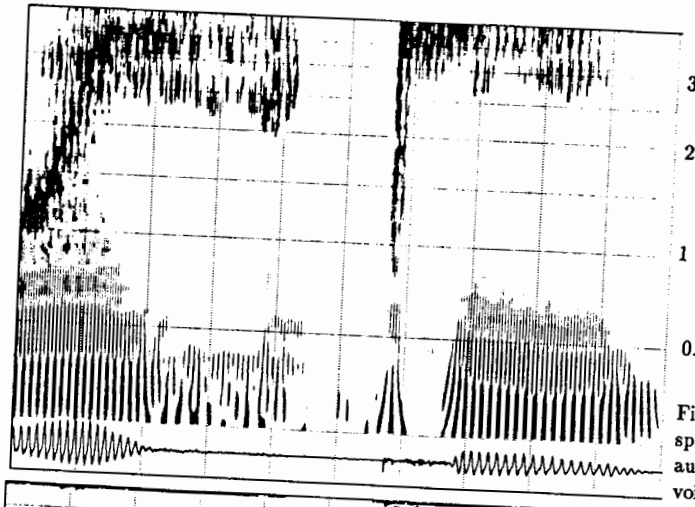


Figure 1: spectrogramme auditif /wiski/ voix féminine

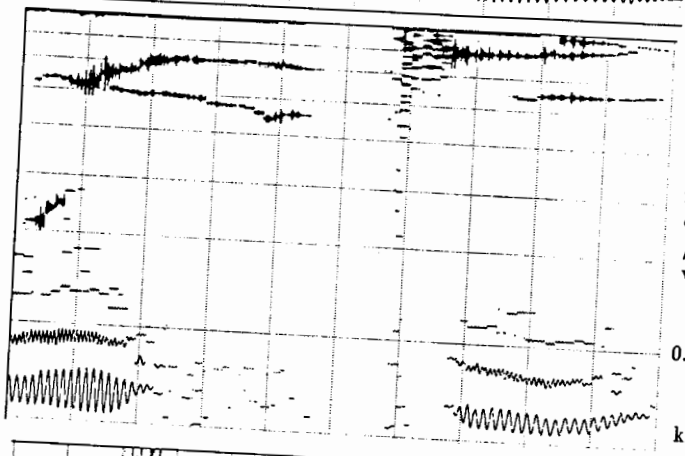


Figure 2: formes d'ondes de synthèse /wiski/ voix féminine

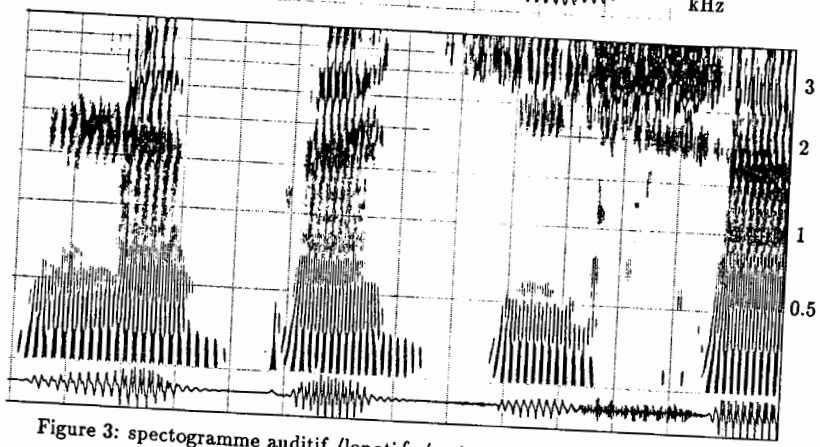


Figure 3: spectogramme auditif /lopotiʃa/ voix masculine

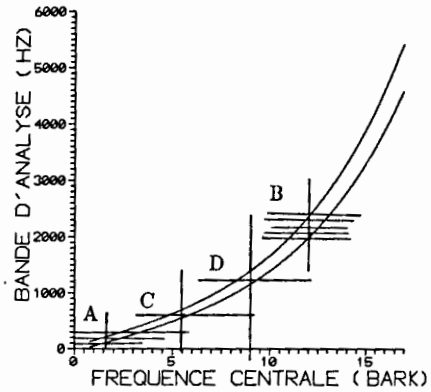


Figure 4: résolution des filtres auditifs

des sinusoïdes (amplitudes et phases) pour le grave du spectre, dans la région du premier formant et en dessous; les fréquences dominantes et la modulation d'amplitude au dessus du premier formant, par recherche des fréquences formantiques et par modélisation de la modulation d'enveloppe temporelle dans chaque filtre. Le système complet, conçu pour un but différent, est décrit dans [2]. Les signaux synthétiques obtenus sont perceptivement équivalents aux signaux originaux: seule une écoute attentive permet de les distinguer. La figure 2 montre le signal synthétique correspondant à la figures 1: les formes d'ondes utilisées pour la synthèse sont portées dans le plan temps-fréquence. Cette image montre les paramètres acoustiques sinusoïdaux et formantiques déduit de l'analyse et eessemble à un squelette de la figure 1.

4 conclusion

Ce papier présente une justification perceptive des paramètres acoustiques apparents sur une représentation spectrographique auditive. Après une discussion sur le type de spectrographe utilisé, les relations entre la résolution du spectrographe auditif et les grandeurs acoustiques du signal de parole sont examinées. Des paramètres acoustiques descriptifs sont proposés pour une représentation acoustique auditive de la parole: pour de la parole voisée, le grave du spectre (en dessous d'environ 1 kHz) peut se décomposer comme une somme d'harmoniques, qui subissent l'influence du premier formant et de la source de voisement; dans une région spectrale moyenne (entre environ 1 et 3 kHz), région des seconds et troisièmes formants les signaux filtrés apparaissent comme des signaux de fréquences dominantes approximativement égales aux fréquences formantiques, modulés en amplitudes à la fréquence du fondamental; au delà d'environ 3 kHz, les formants supérieurs perdent de leur individualité et se regroupent en masses spectrales dont les fréquences dominantes sont peu saillantes et dont la modulation d'amplitude est complexe. Un système d'analyse/synthèse par formes d'ondes élémentaires a permis de montrer l'équivalence perceptive entre le signal naturel et un signal synthétique obtenu en utilisant ces paramètres. Le spectrographe auditif propose une représentation acoustique descriptive qui se démarque de celle basée sur un modèle de production, mais qui paraît perceptivement justifiée. L'avenir permettra de juger de l'efficacité de cette représentation, grâce à un synthétiseur de parole utilisant ce type de paramètre qui est actuellement à l'étude pour la synthèse à partir du texte.

Références

[1] d'ALESSANDRO, C. et BEAUTEMPS, D. (1990), "Representation et modification du signal de parole par transformée en ondelettes utilisant des contraintes auditives", Notes et Documents LIMSI 90-10.

[2] d'ALESSANDRO, C. (1990), "Time-frequency speech transformation based on an elementary waveform representation", Speech Comm. Vol. 9. Nos 5/6, pp. 419-431.

[3] BLADON, A. (1982), "Arguments against formants in the auditory representation of speech", in The Representation of Speech in the Peripheral Auditory System, R. Carlson and B. Granström eds, Elsevier Biomedical Press (North-Holland). pp. 95-102.

[4] CARLSON, R. and GRANSTROM, B. (1982), "Towards an auditory spectrograph", in The Representation of Speech in the Peripheral Auditory System, R. Carlson and B. Granström eds, Elsevier Biomedical Press (North-Holland). pp. 95-102.

[5] COOKE, M.P. (1986), "A computer model of peripheral auditory processing uncorporating phase-locking, suppression and adaptation effects", Speech Comm., Vol. 5. Nos 3/4, pp. 261-281.

MICROWAVE SPEECH SYNTHESIS FROM TEXT

B.LOBANOV AND E. KARNEVSKAYA

Institute of Engineering Cybernatics
Institute of Foreign Languages,Minsk,BSSR

ABSTRACT

The present work suggests a microwave method as a way to synthetic speech quality perfection. The paper deals with the basic principles of the adopted method and some problems arising in the course of its applicati on in a multilanguage progr am: choice of an invariant framework and specific types of microwaves ; the mechanism of microwave conca tenation and their modifi cations in accordance with linguistically significant prosodic changes.

INTRODUCTION

Systems of speech synthesis from text today are general ly based on a formant sig nal method,as it permits a wide range of combinatory and positional modifica tions of the acoustic inva riants representing the pho nemes of the language . It thus meets the requirements of the given type of speech synthesis , namely those of unrestricted vocabulary and sentence structure.Although modern formant synthesisers are capable of producing speech of a fairly high intelligibility and quality [1,2],much is left to be de sired.However,there is hard ly any possibility of a radical improvement at the present time . The reasons for it lie in the inherent deficiencies of the speech formation model being used and in particular, the latter's inability to reflect voice individuality.This is largely because formant synthesisers neglect the interaction of the excitation source and the vocal tract (coupling effect) . Nor do they take account of the de pendence of the excitation pulse shape on the properties of the vocal tract modifications . As a result there still exist problems with the synthesis of a female voice as well as imitation of any definite voice.A way towards the solution of these difficulties, as it seems,is the use of speech segments as the basic elements of synthesis. The minimal units of synthesis in the present work are microwaves (henceforth, MW). They are elements of a natural speech signal coextensive with a F0 period.Actually,the use of microwaves for synthesis programmes was first proposed in [3]. In [4] this idea was successfully approbated in the system of diphone synthesis from text for male and female voices.Yet,there are quite a number of problems in MW synthesis that have not been solved so far (see Abstract). The present work being a multilanguage

programme lays special emphasis on finding language-invariant strategies and compiling language-specific MW sets.The number of micro waves in a set is ultimately determined by the phoneme inventory of the given language,phonetic distances between phonemes belonging to the same class and the difference in the degree of coarticulation between various types of sounds both within one language and across languages . The exact number and types of MW in each set , however,can only be defined experimentally.

2. GENERAL PRESENTATION

2.1.Microwave Phoneme Representation

Like in the formant synthesis,the basic principle of MW method is allophonic representation of the phonemes of the language , but unlike the former , there's further disintegration of allophones into linear segments.Thus MW synthesis consists essentially in obtaining adequate linear models of phoneme combinatory and positional realisations.

Clearly there can be various degrees of discretisation both as regards the relevant list of allophones and their internal structure. The main argument for the validity of the MW sets selected for the present work is that they provide all significant variation of sounds in connected speech.For the Russian vowels, e.g., it is necessary first of all to distinguish between the soft and hard vowels:{A,O,E,U,I},on the one hand , and {'A,'O,'E,'I},on the other.It means that the target units are not phonemes in the strict sense of the word , but allophones viewed as sound types grouped on the grounds of nonfunctional,phonetic,identity . Each allophone of this kind,a higher-rank allophone,so to say,is represented horisontally by three successive segments : initial, mid and final.The segments, like boxes,are to be filled with appropiate microwaves, according to the modifications the given allophone (soundtype) undergoes under the influence of various ajacent sounds . In view of the accepted two-lewel allophonic representation the mid segment,i.e. the vowel stationary,was regarded in this paper as constant for all possible CV and VC combinations of a concrete higher-rank allophone . The choice of the MW type for the initial and final segments,as could be predicted from the results of formant analysis and synthethis , does not follow this principle:transitional microwaves vary in accordance with the ajacent consonant articulation place . The number of MW types then should correspond to the number of consonant classes opposed by this feature.For Russian consonants, e,g.,we distinguish labial, dental,alveolar, velar and lateral places of articulation.

Allophonic variation of con conant phonemes in Russian and in English (as in other languages) is caused both by the impact of the neighbouring consonants and vowels.The former may lead to noticeable qualitative chan ges ,e.g. the emergence of higher-rank allophones,such as the voiceless[r] in English. The latter is mainly confined to variations on the transitional segments. Thus,e.g.,we take three types of microwaves for initi al segments of hard conso-

nants: before [i,e], before [a] and before [o,u].Clearly the three linear segments -are to be determined for each consonant allophone.

2.2. MW concatenation in the Speech Flow

MW concatenation at the stationary segment comes simply to their successive reading-outs.This procedure could be suitable for the transitional segments, too, if readings of several MWs for every type of transition had been preliminarily made.This can hardly be put into practice because of the amount of work needed for the preparation of the speech material and an excessive increase of the required memory volumes as well as the number of rules for the synthesis of transi tional segments.There is an interesting possibility of avoiding these difficulties which is based on the use of the inertial properties of auditory perception.

Let us recall in this connection that the visual imp ression of a smooth replace ment of slides can be achie ved through a smooth decrea se of the brightness (down to zero) of one image and a simultaneous increase of the brightness(from zero up to the required degree) of the other image , projected onto the same screen.Our re search has shown that a similar effect of replacement is observed in sound percep tion.The auditory effect of smooth replacement is achieved by making an overlap interval between the contac ting sounds during which a gradual amplitude decrease of sound 1 and a simultaneous amplitude increase of sound 2 takes place.The amplitude summing up in the field of the overlap leads to the appearance of a complex sound, perceived as a smooth transition from sound 1 to sound 2.

2.3.FO-Parameter Control. The simplest method of controlling the fundamental frequency in the MW synthesis system is the following Let the initial MW have the duration TO which is chosen from the range of variations determined by prosodic rules:TOmin<TO<TOmax.As for a concrete TO value, it can be defined as a statistic mean value of the speaker's FO period used for the formation of the MW set.If the current TO = TO′,the speech signal is formed by a simple repetition of the given microwave.When TO > TO′,the MW repeated reading-out begins after the time interval TO-TO′ and the interval itself is filled by zeroes. If TO < TO′,the reading-out process stops at the moment = TO and a repeated MW read-out resumes.Experimental investigations of this control method have shown that it provides a sufficiently high quality of the synthesized sound,in particular,when TO>TO′,with the interval TO - TO′ not exceeding 30% max. of a period duration TO.When TO<TO there are no percertible distortions only if the end of the read-out falls at a MW value close to zero (10-20%% of MW amplitude).Other wise, there's a clear sound distortion resembling nasalization. This unwanted effect can be removed by smoo thing away the abrupt reading-out cessation process. It can be achieved by switching on an order 2 filter with the time constant equ. 0.25*TO at the moment the repeated read-out begins or before this moment (0.25TO) by multiplying the MW by a

smooth single function of the type y=exp(-t). The use of either of these methods for the case TO < TO′yields fairly good results . If TO > TO′, the method of period zeroing can be applied provided TO~=0.7*TOmax.

3.IMPLEMENTATION

The algorithm implementing the above model consists of 10 blocks. The written text intended for synthesis is produced by the PC main pro gram. This text,sentence by sentence, gets into block 1, in which sentences are segmented into intonation-groups and marked for stresses and melody.These procedures are performed in acccordance with definite rules vary ing from language to language. For every phoneme then in blocks 4,5 are calculated : the rhythmic (sound duration), the dynamic (sound intensity) and the melo dic (pitch) characteristics according to the rules specified for the languages . Futher on,in block 5, allophonic identification of the phonemes is carried out followed by the division of the allophones into linear segments.To each elementary segment corresponds a definite MW which is selected by block 6 out of the MW - set,determined for each lan guage and type of voice. In block 7 modifications of the MW duration take place (i.e. of the FO period) in accordance with the informa tion coming from block 4 , and in this way the tonal pattern of the synthetic speech is produced. Control led by block 3,the duration of phoneme segments is defi ned in block 8 by means of a step-by-step reading of the required number of MWs. Finally in block 9 MW ampli tude (intensity)is set out, while block 10 serves for smoothing the abrupt changes at the transitions from one MW type to another in the process of generating a continuous speech signal. Changing the voice type in MW synthesis is achieved by replacing or modifying the MW set . Passing over to another language inplies the replacement of the phonetic base rules and MW set

REFERENCES:

1.KLAT D.The Klattalk text-to-speech conversation system.Proc.IEEE ICASSP,Paris, 1982.

2.LOBANOV B. The Phonemaphone text-to-speech system. Proc. ICPhS,Tallinn,1987.

3.MOREL M. Synthese vokale par recordment de segment d'ooscillogrames Revue d'Ac oustique,vol.14,no 56,1981.

4.HOMON C. Synthese par con catenation de formes d'ondes.

5.KARNEVSKAYA E. The linguistic principles of multilanguage synthesis . Proc. ICPhS, Tallinn,1987.

VOICE OPERATED MULTILINGUAL INFORMATION DISPLAY/RETRIEVAL SYSTEM

Abdul Mobin, Anil Kumar, and S.S.Agrawal

CENTRAL ELECTRONICS ENGINEERING RESEARCH INSTITUTE CENTRE, NEW DELHI, INDIA.

ABSTRACT

This paper describes the design and development of a voice operated multilingual information display and retrieval system. The words or commands spoken by a speaker can be displayed on the multilingual terminal for varification and data entry applications. Previously stored information can also be accessed through a spoken word command and displayed on a CRT terminal.

The system consists of an Isolated Word Recognition Unit, an interface unit and multilingual CRT terminal. The design aspect and the functioning of the system are described in detail The code generated by the recognition unit corresponding to the recognized word is accepted by the interface card through the Z-80PIO Having compared this code with the code stored in the interface card,the information to be displayedis sent to the CRT terminal using serial communication. The system is based on the Z-80 microprocessor and utilises 64 KB of RAM/EPROM for storage of information and software programmes.ASCII data of the information to be displayed corresponding to the recognized word is sent by the interface card and displayed on the terminal. The terminal use a software called GIST and has the capability of displaying some Indian and Roman scripts. The information can also be displayed in different font and graphics forms.

1. INTRODUCTION

Speech is the most natural and efficient means for humans to communicate with each other as well as with machines. Among the human-machine interaction system speech recognition is an attempt to have machines responding to spoken commands. Attempts are being made to have automatic speech recognition systemof various complexities.However, the success so far is limited to handle small vocabulary and to great extent they are speaker dependent systems. This paper describes the development of a voice operated multilingual information display/retrieval system and the flexibilities ofit's practical applications. Both the word recognition system as well as the interface card for displaying the information are Z-80 A microprocessor based hardware modules.

Interesting applications are to be anticipated with automatic speech recognition systems in the areas of industrial controls, transport,communication and military technologies. As one of the interesting applications, the automatic spoken word recognition system can be successfully used for the display and retrieval of stored informations on a CRT by voice commands. The CRT used for this purpose is equipped with a GIST card which accepts serially the ASCII codes of the information to be displayed. To do this a controlling unit has been designed and developed at CEERI. The unit is capable of supplying the serial ASC II codes of information desired to be displayed on the CRT.

With GIST card incorporated in the above system it is possible to display the information in any of the several Indian scripts and Roman script for English either simultaneously or in one of the selected scripts on the screen. Since India is a multilingual country, the objective of developing the Voice Operated Information Display System is to have interaction with computers using multilingual scripts.

2. DESCRIPTION OF THE SYSTEM

The system consists of two major parts, one is a word recognition unit and the other is an interface unit for informtion display. These two units are suitably interfaced as shown in the figure 1.

2.1 Spoken Word Recognizer

The recognition processor is a single board microcomputer based on a Z-80 A CPU. It consists of 64 Kbytes of RAM/EPROM, Parallel interface adapter(PIA), Counter Timer Circuit(CTC) and two Asynchronuous Adapters (ACIA). One ACIA interfaces to the video terminal controller unit.

The system is speaker dependent but it can be easily trained for a new speaker.The speaker has to speak through a close talking microphone. This speech signal is fed to signal conditioning circuitry which consists of pre-amplifier, equalizer, automatic volume control etc. The preprocessed signal is then passed through an audio spectrum analyzer chip ASA-16.It has 16 contiguous band pass filters covering the frequency range of 200Hz to 7KHz. Each bandpass filter is follwed by a rectifier and a low pass filter with cut off frequency of 25 Hz. This arrangement gives the slowly varying signal proportional to sound pressure level in that channel. The ASA-16 also has 16-channel multiplexer which provides spectral data for each of the 16 frequency bands. The multiplexed signal is allowed to pass through the 8-bit analog-to-digital conveter and this digitised data is stored in the memory.

The underlying principle behind the isolated word recognizer is pattern matching. In the first instance, patterns of all the words corresponding to a vocabulary are obtained and their templates stored in the system memory. Next, the pattern of the test word is generated and coma

pared with the templates recorded earlier. In case of a correct match there is a display of the correctly recognized word on the front pannel of the machine. Corresponding to the matched word the system generates a specific code. The code is accepted by the interface unit through I/O device (Z-80 PIO). This code is further processed for another comparison with prestored codes. Having compared with the correct code this interface unit sends the ASCII data of the stored information to CRT terminal through serial communication.

2.2 Interface Unit

This unit is also based on Z-80A CPU which is driven by 4.0 MHz clock. It uses a special purpose integrated circuit designed to ease the problems associated with serial communication i.e. Universal Synchronuous/Asynchronuous Receiver and Transmitter (USART).

The unit provides a three line serial communication port. One line transmits character to the terminal, second line receives character entered at the terminal keyboard and the third line simply provides a common system ground connection. Information is put on these wires one bit at a time and in this way it takes eight seperate transmissions to send a whole byte. The baud rate i.e. speed at which the data is transmited is chosen to be 9600 in our case. However, this rate depends upon the specific application and the equipment involved. Since there is no handshaking in this I/O process, it is the microprocessor's responsibility to send characters not faster than speed at which the termial can receive. Similar method is used to accept the incoming characters at the frequent intervals to avoid missing data being sent by the terminal.

To communicate with the terminal, the transfers to and from interface unit are conveyed by using the American standard EIA-RS232 C, because the data are no longer activated by current intensity but by voltage level i.e. +12V and -12V. Also such levels are not TTL compatible and can not be provided by the microprocessor. Thus a physical interface is interposed using a driver and a receiver circuit. The functional diagram of the complete system is shown in Fig.2

3. Testing/Features of the System

The system is tested with 10 words vocabulary spoken by both male and female speakers. Though it can accomodate 40 words vocabulary. The words are spoken once into a microphone to generate the reference pattern. Corresponding to 10 words vocabulary 10 pages of stored information can be retreived. Transcript provides emulation for the DEC VT-52 and Ansi compatible terminals for display of English as well as Indian scripts. Another mode displays high resolution, print quality characters for Indian scripts which can be used for text processing requirements.

ACKNOWLEDGEMENT

The authors are grateful to Dr.W.S.Khokhle, Director CEERI for kind support and encouragement for the above work. The authors are thankful to Mr.M.Ganesan for various discussions and help. They wish to acknowledge the financial support of Deptt. of Electronics, Govt. of India.

REFERENCES

1. Abdul Mobin, S.S.Agrawal and K.D.Pavate "An Online Spoken Word Recognition System using a microprocessor", Proc. European conf. on Speech Technology, Edinburgh, U.K. Vol. 2, pp.296-299, Sept 2-4, 1987.

2. C.K.Chaung and S.W.chan, "Speech Recognition using Variable Frame Rate Coding", IEEE Proc. of Int CASSP, pp.1033-1036, 1983.

3. R.Pieraccini & R.Billi, "Experimental Comparison among Data Compression Techniques in Isolated Word Recognition" IEEE Proc. of ICASSP, pp. 1025-1028, 1983.

4. B.A. Dautrich, L.R.Rabiner and T.B.Martin, "On the use of Filter Bank Features for Isolated Word Recognition" IEEE Conf. ASSP, pp. 1061-1064, 1983.

5. W.Daxer, Speech Communication, Vol.1, No.1, 21, 1982.

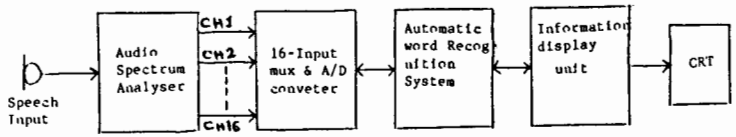


Fig.1. Block Diagram Of the System.

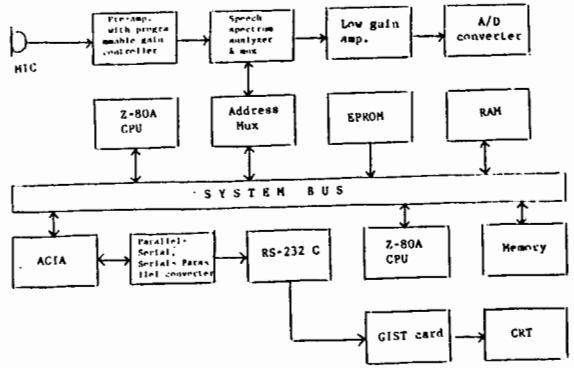


Fig.2. Functional Diagram Of the System.

THE EXTRACTION AND INTEGRATION OF SELECTED CUES FOR VOICING INTO A CONTINUOUS-WORD AUTOMATIC SPEECH RECOGNITION SYSTEM

Dariusz A. Zwierzyński and Claude Lefèbvre

Neil Squire Foundation and National Research Council
Speech Research Centre, Building U-61
Montreal Road, Ottawa, Ontario, Canada K1A 0R6

ABSTRACT

This paper deals with the enhancement of automatic recognition of stops in isolated word-initial stressed syllables of minimal word-pairs in normal and degraded speech. It was found that the error rate for stops could be decreased by incorporating the cues to voicelessness/voicing extracted from a short-time spectral content of the speech signal. The acoustic cues comprised: F_0 perturbations in post-occlusive vowels, voice onset time (VOT), and the presence/absence of low-frequency at voicing onset. Results confirm that additional input features improve performance.

1. INTRODUCTION

Automatic recognition of stops in word-initial position in minimal word-pairs poses problems for recognition systems. Our studies indicated that many errors were caused by the system not distinguishing between the voicing status of the initial stops. Apparently, the spectral information from the filter-bank analysis was insufficient or too weak for the system to make an efficient distinction.

We propose a method of reducing the error rate, by extracting new information on the voicing status from the subphonemic level of a stop, and integrating it with other spectral information derived by the filter-bank analysis. This concept represents a phonetically-based approach to recognition of stops, which enhances the successful performance of algorithms employing linear discriminant analysis (LDA) developed in Canada [1] and implemented in our recognition system.

2. SPEECH RECOGNITION SYSTEM

The recognition system used in the experiments is a speaker-dependent, robust, small-vocabulary, continuous speech recogniser. The system performs very well in quiet and in distorted speech [2].

Spectral representations from the filter-bank analysis are processed by a linear discriminant network. Our system can use static and dynamic spectral representations as input to the linear discriminant network [Fig. 1].

The linear discriminant analysis generates a set of discriminant functions which are applied to the acoustic features extracted from the speech signal. As the linear discriminant network combines various mel-scale spectral representations into a single set of discriminant functions, the transformations have been called IMELDA for integrated mel-scale LDA.

3. EXTRACTION OF NEW FEATURES

3.1. F_0 perturbations

The speech database used in our experiments included 5 examples of 6 CVC minimal word-pairs comprising the 6 stops and the vowel /i/ recorded by 7 male speakers.

F_0 at voicing onset of vowels after voiceless stops starts higher and falls deeper than it does after voiced stops where it either slightly falls or remains level [5]. The deep F_0 fall is one of the stronger cues to a voiceless stop.

Fundamental frequency perturbations are extracted by a modified custom-designed AMDF pitch detector, operating on the raw audio waveform and outputting pitch data from the initial frames at voicing onset. [4]. The pitch information is then integrated with the spectral information.

Experimental results demonstrate that the benchmark IMELDA-2 representation was outperformed when supplemented with pitch information (IMELDA-P) extracted from voicing onset (Tab. 1). The improvement is evident across the 4 acoustic experimental conditions.

3.2 Voice Onset Time

In English VOT is longer for voiceless stops than for voiced ones. The average time separating the two voicing categories is approx. 44 msec. Two different methods were used to derive the VOT spectral representations; in both of them the burst and VO points had been manually labelled.

In the first method, a VOT representation was obtained from a second derivative of the static spectral features (LCE). The linear regression analysis applied twice on 7 frames of 6.4 msec of LCE gives an effective separation of the two VOT groups

Table 1 illustrates that in comparison with IMELDA-2, the addition of VOT resulted in improved recognition in all 4 conditions. Compared with IMELDA-P, IMELDA-VOT is worse only in Tilt.

In another experiment IMELDA-2 was combined with pitch and VOT (Fig. 2). A comparison of IMELDA-PVOT with IMELDA-VOT (Tab. 1) reveals improvement in Noise-1, but the error rate is slightly higher for Quiet and Tilt. Yet, here again, IMELDA-PVOT outperformed IMELDA-2 in all 4 conditions.

The conclusion derived from this series of experiments is that IMELDA-VOT constitutes the optimal representation for the four acoustic conditions. The addition of pitch and VOT to IMELDA-2 indicates that improvement is possible, however it is not so uniform as with IMELDA-VOT and IMELDA-P, and better than IMELDA-VOT only in one experimental acoustic condition. With white noise added (15dB SNR) there is an improvement with IMELDA-PVOT over the other representations, which may suggest that in this condition the glottal excitation pulses are more resistant to noise than the burst pulse. Indeed, better recognition in noise was obtained with all representations incorporating the extracted cues. We suspect that tilting the spectral balance resulted in higher error rates with IMELDA-VOT and IMELDA-PVOT, as by changing the gain in each channel, the perceptual salience of the voicing cues may thereby have been changed.

Deriving a VOT representation from a 2nd derivative is effective but too slow for real-time applications. Hence, a faster method was developed.

4. TIME-DOMAIN FILTERS

Linear discriminant time-domain filters were derived from the use of linear discriminant analysis on the LCE representation of quiet speech. (Fig. 3). The VO instant was the timing reference point. Between- and within-class onset periods were computed for 4 and 13 frames before VO and 4 and 3 frames after VO, giving a time-domain filter with 8 or 16 coefficients.

The filter was then applied to all frames of the analysed word. Also, experiments were conducted with a different number of discriminant filters. Using 8 frames enabled us to study the contribution of cues detectable in the time domain in the VO area. By using 16 frames, the discriminant filter could be derived from running LDA on the segment spanning an entire VOT area, including both the burst and the voicing onset.

In the 8 frame analysis, we wanted to extract the cues related to the F1 transition, and specifically the absence or presence of the low-frequency energy after VO which cues the perception of a voiceless/voiced stop respectively [3]. Table 1 illustrates that 3 discriminant filters (IMELDA-3f8c) produce the best results for this condition, outperforming IMELDA-2 in the 4 acoustic conditions.

In the other analysis, discriminant filters were computed over 16 frames to maximise the range by deriving filters from a VOT area, capturing the possible burst, VOT, and F1 cues in one type of analysis. The results (Tab. 1) reveal that 4 time-domain discriminant filters (IMELDA-4f16c) perform much better in quiet and with 15 dB SNR than IMELDA-2, IMELDA-VOT, and IMELDA-PVOT, and slightly worse with 9 dB SNR and in tilted speech. On the other hand, it turns out that IMELDA-3f8c outperforms IMELDA-4f16c in degraded speech, and is slightly worse in Quiet. It may be here that degraded speech obliterates the burst, and thus the VOT cue, much more than it happens when the filters are derived from the voicing onset area only.

Complete results from tests with adding pitch data to IMELDA-3f8c/4f16c were not available at the time of writing. Initial data suggest that adding pitch can improve the performance in degraded speech.

5. DISCUSSION AND CONCLUSIONS

The incorporation of additional cues to voicelessness/voicing into our recognition system has decreased the error rate for stop consonants. Despite using a small speech database, we believe that the benefits from incorporating the new input features have been demonstrated. It has to be emphasised that IMELDA and the time-domain filters were derived from a quiet speech representation. Previous findings [2] indicate, however, that deriving an IMELDA transform from integrated quiet, noisy and tilted speech representations produces better results in degraded speech, and only negligibly worse in quiet, than when the transform is derived from quiet speech only. We shall publish corresponding results obtained with IMELDA representations computed on degraded speech as they become available. We expect that computing the IMELDA transform and the time-domain filters on degraded speech may result in an improved performance of the more versatile IMELDA-4f16c compared to IMELDA-3f8c.

The usage of time-domain discriminant filters provides for focussing the analysis on different areas of interest, as shown in our experiments. Similarly, different time-domain filters could be used for processing intervocalic stops, where the VOT cue may not be reliable, or word-final stops, where again the presence and salience of the relevant cues is different. In addition, time-domain filters are suitable for real-time implementation, and will be incorporated into the future versions of the hardware recogniser jointly developed by the Neil Squire Foundation, the Canadian Marconi Company, and the National Research Council of Canada. Current work in this area also includes the refinement of a neural network (MLP) for real-time pitch and voicing onset/offset detection for the automatic calculation of VOT representations [6].

6. ACKNOWLEDGEMENTS

Part of this work was conducted under a contract (041ST.W7714-0-3529) with the DND Canada. We are grateful to Gary Birch and David Starks for discussions on various topics related to research reported in this paper.

7. REFERENCES

[1] HUNT, M. J. and LEFEBVRE, C. (1989), "Distance measures for speech recognition", *Aeronautical Note, NAE-AN-57*, Ottawa, March.

[2] HUNT, M. J. and LEFEBVRE, C. (1989), "A comparison of several acoustic representations for speech recognition with degraded and undegraded speech", *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Proc., ICASSP-89*, Glasgow, Scotland.

[3] KLATT, D. H. (1975), "Voice onset time, frication, and aspiration in word-initial consonant clusters", *J. of Speech and Hear. Res., 18*, 686-706.

[4] LEFEBVRE, C. and ZWIERZYNSKI, D. A. (1990), "On the use of F_0 variations in automatic speech recognition", *J. Acoust. Soc. Am. 87, Supl. 1*, S105.

[5] SILVERMAN, K. (1986), "F_0 segmental cues depend on intonation: The case of the rise after voiced stops", *Phonetica, 43*, 76-91.

[6] ZWIERZYNSKI D. A. and LEFEBVRE C. (1990), "Improvement of the NRC automatic speech recognition system", *Proc. of the Canadian Conf. on Electr. and Comp. Eng., 2*, 5.3.1.-5.3.4, Ottawa, Canada.

Table 1. Isolated CVC recognition results for 5 examples of 6 minimal-word pairs spoken by 7 male speakers. Test material presented in 4 conditions: undegraded (Quiet), with white noise added to give a 15dB SNR (Noise-1) and 9dB SNR (Noise-2), and with a 6dB/octave tilt applied (Tilt).

SPEAKER-DEPENDENT ISOLATED CVC RECOGNITION ERRORS (%)				
Representation	Quiet	Noise-1	Noise-2	Tilt
IMELDA-2	4.3	11.4	23.4	29.6
IMELDA-P	3.8	11.0	20.5	22.4
IMELDA-VOT	2.4	10.0	20.5	25.7
IMELDA-PVOT	2.9	7.6	20.5	27.6
IMELDA-3f8c	1.4	3.3	16.6	23.3
IMELDA-4f16c	0.0	6.2	24.8	32.4

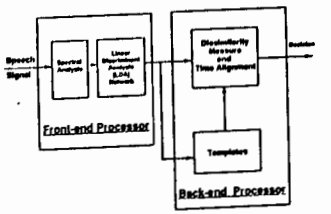


Fig. 1. Block diagram of the speech recognition system

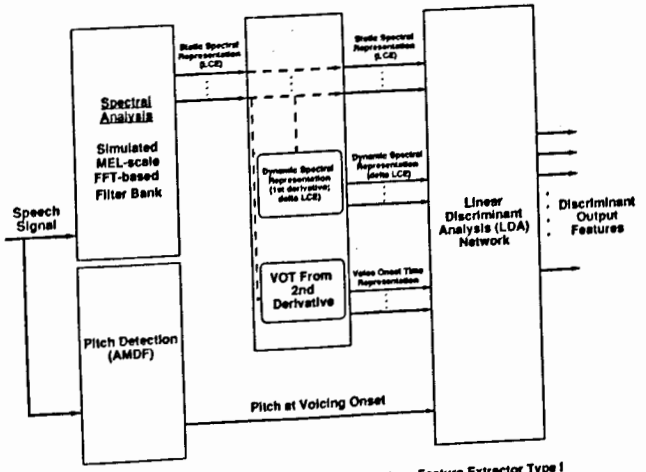


Fig. 2. Block diagram of the automatic speech recognition system: Feature Extractor Type I

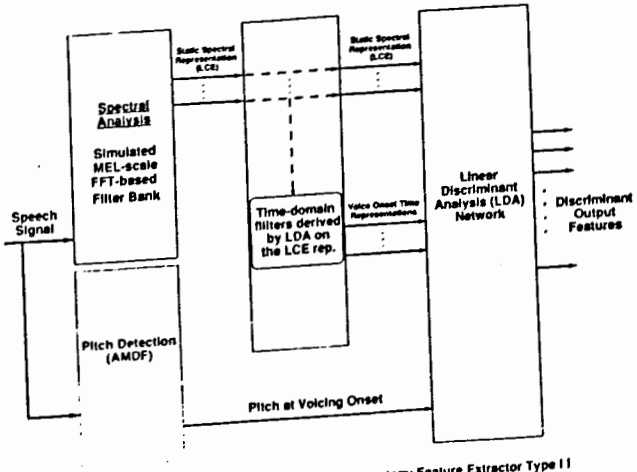


Fig. 3 Block diagram of the automatic speech recognition system: Feature Extractor Type II

PERCEIVED SPECTRAL ENERGY DISTRIBUTIONS FOR EUROM.0 SPEECH AND FOR SOME SYNTHETIC SPEECH

Chaslav V. Pavlovic, Mario Rossi, and Robert Espesser

Institut de Phonetique, LA 261 CNRS, Université de Provence,
Aix en Provence, France
& (1st. author only) University of Iowa, Iowa City, Iowa, USA.

ABSTRACT

Normative data on speech energy variations over time has been obtained for five different languages (French, Dutch, English, Italian, and Danish) available on the EUROM.0 CD-ROM produced by the SAM partners.

1. INTRODUCTION AND METHOD

The purpose of this research was to obtain normative data on speech energy variations over time for five different languages (French, Dutch, English, Italian, and Danish) available on the EUROM.0 CD-ROM produced by the SAM partners [1]. Continuous passages spoken by four talkers (two of each sex) were analyzed for each language. The variable whose distribution is analyzed is termed "perceived spectral energy." It represents the energy of speech contained within a band 1-Hz-wide at the output of an exponential time window. The exponential weighting is performed to simulate the auditory filter. Five time constants (τ) of the window are used in the analysis: 13 ms, 30 ms, 80 ms, 125 ms, and 200 ms. For a given τ the perceived spectral energy is calculated as

$$E_\tau(t) = \int_{-\infty}^{t} p_f^2(\xi) e^{-(t-\xi)/\tau} d\xi$$

where p_f is the sound pressure density in pascals. The distributions were obtained for each critical band.

2. RESULTS

The results indicate that there are no significant main effects of sex and language. Therefore, the normative values are reported as the mean values across all talkers. Fig. 1 gives results for the τ of 13 ms. Each of the family of curves corresponds to the sound pressure level below which the perceived spectral energy occurs for the percentage of time indicated at the right of the curve. All the values are in reference to the long-term spectrum level specified in Table 1. No data below the 25% contour are depicted in Fig. 1 because they were contaminated by noise. Fig. 2 refers to the measurement with the τ equal to 200 ms. The apparent dynamic range of speech with this τ appears much reduced in comparison with the results obtained with the τ of 13 ms. The results for the time constants of 13, 30, 80, and 125 ms are given numerically in Tables 2 to 5, respectively.

3. DISCUSSION

These data will be useful in future studies trying to develop physical measures of speech quality. As an illustration of the possible utility of these measures Fig. 3 gives "range" values for speech synthesized by a commercially produced synthesizer (solid lines) and EUROM.0 speech (dashed lines). The variable "range" essentially measures the dynamic range of speech. In Fig. 3 the top

curve refers to the differences in the sound pressure levels below which the speech is 95% and 25% of time. The three curves below it refer, respectively, to the differences of 75% - 25%, 65% - 35%, and 55% - 45%. It would appear that the synthesizer somewhat compresses the dynamic range of speech. This analysis may provide a means to quantify this compression in various frequency bands and relate it speech quality. A more complete version of Tables 2 - 5, as well as the table for the τ of 200 ms are given in [2].

4. ACKNOWLEDGMENTS

This research is made possible by a grant from the EEC Esprit SAM project (Grant no. 2589).

5. REFERENCES

[1] Grice, M., and Barry, B. (1988). "EUROM.0 technical description," Doc.no.:SAM-UC-135. (SAM - ESPRIT Project 2589, Wolfson House, 4 Stephenson Way, London NW1 2HE).

[2] Pavlovic, C.V., Rossi, M., and Espesser, R. (1991). "Perceived spectral energy distributions for EUROM.0 speech and for some synthetic speech. Doc.no.: CP_03_91.AIX. (SAM - ESPRIT Project 2589, Wolfson House, 4 Stephenson Way, London NW1 2HE).

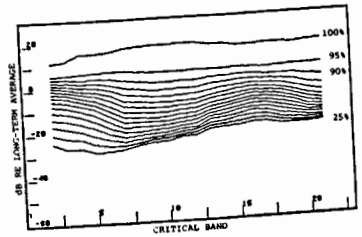


Fig.1 Perceived energy distribution for τ = 13 ms.

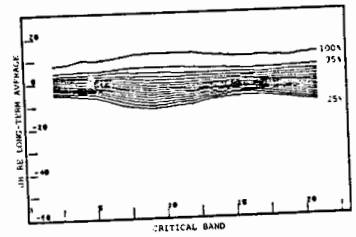


Fig.2 As Fig.1 but for τ = 200 ms.

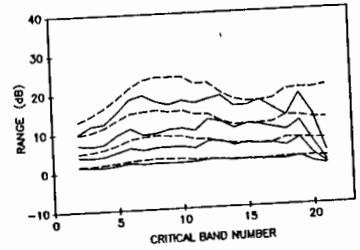


Fig.3 Range for a synthesized speech (solid lines) and EUROM.0 speech (dashed lines).

Table 1 The long-term rms of EUROM speech; 60 dB SLP overall

CRIT. BAND	CRIT. BAND C.F.(Hz)	RMS (dB SPL)
2	150	30.4
3	250	34.8
4	350	29.3
5	450	31.4
6	570	29.5
7	700	26.3
8	840	24.8
9	1000	22.7
10	1170	19.8
11	1370	18.2
12	1600	17.3
13	1850	15.0
14	2150	11.6
15	2500	12.5
16	2900	11.8
17	3400	10.6
18	4000	8.9
19	4800	7.8
20	5800	7.6
21	7000	5.9

Table 2 EUROM speech distribution for τ = 13 ms

CRITICAL	% OF SPEECH BELOW THE LEVEL IN THE TABLE									
BAND	25	30	40	50	60	70	80	90	95	100
2	-25.1	-21.9	-15.5	-9.6	-4.8	-1.2	2.0	4.9	6.5	12.9
3	-27.4	-22.7	-14.7	-8.8	-4.4	-1.3	1.6	4.7	6.7	12.4
4	-25.7	-22.2	-15.4	-9.8	-6.0	-2.7	0.3	4.1	6.7	15.5
5	-27.8	-24.4	-17.1	-10.9	-6.0	-2.5	0.8	4.3	7.0	15.5
6	-27.5	-25.2	-19.9	-14.4	-9.4	-4.7	-0.5	4.1	7.1	15.9
7	-27.7	-26.3	-22.5	-17.9	-13.0	-8.1	-2.8	3.6	7.6	17.1
8	-27.3	-26.0	-22.6	-18.5	-14.1	-9.2	-4.0	2.6	7.0	18.0
9	-25.5	-24.3	-21.5	-18.3	-14.5	-10.1	-5.0	1.7	6.6	18.0
10	-23.8	-22.8	-20.3	-17.3	-13.8	-9.7	-4.6	1.9	6.6	18.6
11	-22.6	-21.7	-19.5	-16.7	-13.3	-9.3	-4.7	1.6	6.1	19.3
12	-21.9	-20.8	-18.5	-15.7	-12.4	-8.6	-4.1	1.9	6.5	19.7
13	-19.4	-18.2	-15.6	-12.8	-9.9	-6.5	-2.5	2.9	6.8	18.3
14	-19.4	-18.2	-15.4	-12.4	-9.3	-5.9	-1.8	3.4	7.1	17.6
15	-18.5	-17.0	-13.9	-11.0	-8.1	-4.8	-1.0	3.6	7.0	16.9
16	-17.1	-15.5	-12.1	-9.1	-6.4	-3.6	-0.2	3.9	6.9	15.6
17	-16.6	-14.9	-11.6	-8.7	-6.2	-3.6	-0.8	3.0	6.3	17.2
18	-17.7	-16.1	-12.7	-9.5	-6.7	-3.9	-1.0	3.0	6.5	17.8
19	-18.7	-17.6	-14.8	-11.8	-8.7	-5.5	-2.0	2.9	7.0	17.1
20	-19.6	-18.7	-16.7	-14.3	-11.8	-8.7	-4.5	2.0	7.1	17.7
21	-19.1	-18.5	-16.7	-14.5	-12.1	-9.3	-5.3	1.8	7.3	18.2

Table 3 EUROM speech distribution for τ = 30 ms

CRITICAL	% OF SPEECH BELOW THE LEVEL IN THE TABLE									
BAND	25	30	40	50	60	70	80	90	95	100
2	-18.9	-15.7	-10.2	-6.4	-3.3	-0.5	2.1	4.7	6.2	11.8
3	-18.6	-14.7	-9.5	-6.0	-3.1	-0.7	1.8	4.6	6.4	12.0
4	-18.1	-14.9	-9.8	-6.7	-4.1	-1.8	0.8	4.1	6.5	14.8
5	-20.0	-16.0	-10.5	-6.8	-4.0	-1.4	1.2	4.3	6.7	14.2
6	-22.0	-19.0	-13.8	-9.7	-6.0	-2.7	0.5	4.3	6.8	14.8
7	-24.2	-21.9	-17.3	-12.9	-9.1	-5.3	-0.9	4.2	7.5	15.7
8	-24.4	-22.4	-18.1	-14.2	-10.4	-6.5	-2.3	3.3	7.1	16.9
9	-23.1	-21.6	-18.2	-14.7	-11.2	-7.5	-3.2	2.6	6.8	17.0
10	-21.9	-20.4	-17.4	-14.1	-10.8	-7.1	-2.9	2.8	6.7	17.4
11	-20.9	-19.7	-16.7	-13.6	-10.3	-7.0	-2.9	2.4	6.3	18.1
12	-20.0	-18.6	-15.7	-12.7	-9.5	-6.2	-2.4	2.8	6.5	18.2
13	-17.2	-15.7	-12.8	-10.1	-7.5	-4.6	-1.2	3.4	6.7	16.8
14	-17.2	-15.6	-12.6	-9.8	-7.1	-4.1	-0.6	3.8	6.9	16.1
15	-15.9	-14.2	-11.3	-8.6	-6.0	-3.1	-0.1	3.9	6.9	15.3
16	-14.4	-12.6	-9.5	-7.0	-4.7	-2.3	0.5	4.1	6.7	14.3
17	-13.7	-12.0	-9.0	-6.8	-4.7	-2.6	-0.1	3.4	6.3	15.7
18	-15.0	-13.1	-9.9	-7.3	-5.0	-2.7	-0.1	3.5	6.5	16.1
19	-16.7	-15.1	-12.0	-9.2	-6.7	-4.0	-0.7	3.8	7.0	15.7
20	-18.0	-16.9	-14.4	-12.0	-9.5	-6.4	-2.3	3.4	7.2	16.3
21	-17.9	-16.9	-14.7	-12.5	-10.1	-7.2	-2.9	3.3	7.5	16.7

Table 4 EUROM speech distribution for τ = 80 ms

CRITICAL	% OF SPEECH BELOW THE LEVEL IN THE TABLE									
BAND	25	30	40	50	60	70	80	90	95	100
2	-10.3	-8.4	-5.6	-3.5	-1.6	0.4	2.1	4.2	5.6	10.5
3	-10.5	-8.4	-5.6	-3.5	-1.6	0.1	1.9	4.2	5.8	10.7
4	-10.0	-8.3	-5.9	-4.1	-2.5	-0.7	1.3	3.9	6.0	12.7
5	-10.7	-8.5	-5.7	-3.6	-1.9	-0.2	1.6	4.1	6.0	11.9
6	-13.2	-11.0	-7.7	-5.1	-2.9	-0.8	1.4	4.2	6.2	13.1
7	-16.2	-13.9	-10.3	-7.4	-4.7	-1.8	1.0	4.7	7.0	13.5
8	-17.7	-15.4	-11.7	-8.8	-6.1	-3.3	0.0	4.1	7.0	15.0
9	-17.9	-15.9	-12.7	-9.8	-7.0	-4.1	-0.7	3.7	7.0	14.9
10	-17.2	-15.4	-12.3	-9.5	-6.6	-3.7	-0.5	3.6	6.8	15.4
11	-16.6	-15.0	-12.0	-9.3	-6.6	-3.8	-0.7	3.4	6.4	15.8
12	-15.6	-14.1	-11.0	-8.4	-5.8	-3.2	-0.5	3.5	6.6	15.8
13	-12.7	-11.3	-8.8	-6.6	-4.5	-2.3	0.4	3.9	6.4	14.4
14	-12.6	-11.1	-8.6	-6.3	-4.1	-1.7	0.8	4.0	6.6	14.0
15	-11.4	-10.0	-7.5	-5.3	-3.3	-1.2	1.1	4.0	6.4	13.1
16	-10.0	-8.6	-6.3	-4.4	-2.5	-0.7	1.4	4.1	6.1	12.1
17	-9.6	-8.3	-6.2	-4.5	-2.9	-1.1	0.8	3.6	6.0	13.4
18	-10.5	-8.9	-6.6	-4.8	-3.0	-1.1	0.9	3.9	6.3	13.5
19	-12.2	-10.7	-8.1	-6.0	-3.8	-1.6	1.0	4.3	6.5	13.3
20	-14.4	-12.9	-10.5	-8.0	-5.5	-2.7	0.6	4.3	6.8	14.1
21	-14.7	-13.4	-11.0	-8.8	-6.3	-3.2	0.2	4.5	7.0	14.5

Table 5 EUROM speech distribution for τ = 125 ms

CRITICAL	% OF SPEECH BELOW THE LEVEL IN THE TABLE									
BAND	25	30	40	50	60	70	80	90	95	100
2	-7.7	-6.3	-4.2	-2.5	-0.9	0.6	2.1	3.9	5.2	9.5
3	-8.0	-6.4	-4.3	-2.6	-1.0	0.3	2.0	4.0	5.5	9.8
4	-7.7	-6.5	-4.7	-3.2	-1.7	-0.2	1.5	3.8	5.7	11.7
5	-7.9	-6.4	-4.3	-2.6	-1.2	0.2	1.7	3.9	5.6	10.8
6	-9.9	-8.3	-5.7	-3.7	-1.9	-0.2	1.7	4.0	5.9	12.1
7	-12.6	-10.7	-7.8	-5.3	-3.1	-0.9	1.6	4.6	6.6	12.2
8	-14.2	-12.2	-9.2	-6.8	-4.5	-2.1	0.8	4.3	6.9	13.6
9	-14.9	-13.2	-10.3	-7.8	-5.3	-2.8	0.1	4.1	7.0	13.8
10	-14.4	-12.8	-10.0	-7.4	-5.0	-2.5	0.2	3.9	6.7	13.9
11	-14.0	-12.5	-9.9	-7.4	-5.0	-2.6	0.0	3.7	6.5	14.6
12	-13.1	-11.5	-8.9	-6.6	-4.4	-2.2	0.3	3.7	6.4	14.7
13	-10.5	-9.2	-7.0	-5.2	-3.4	-1.4	0.9	3.9	6.3	13.3
14	-10.4	-9.1	-6.9	-4.9	-2.9	-0.8	1.2	4.1	6.3	12.8
15	-9.2	-7.9	-5.9	-4.1	-2.4	-0.6	1.4	4.0	6.1	12.0
16	-8.1	-6.9	-5.0	-3.3	-1.8	-0.1	1.7	4.0	5.7	10.8
17	-7.8	-6.7	-5.1	-3.5	-2.1	-0.6	1.1	3.6	5.7	12.1
18	-8.4	-7.3	-5.4	-3.8	-2.2	-0.5	1.4	3.9	5.9	12.3
19	-10.0	-8.7	-6.6	-4.5	-2.7	-0.6	1.6	4.2	6.1	12.0
20	-12.1	-10.9	-8.3	-5.9	-3.7	-1.3	1.3	4.3	6.3	12.8
21	-12.5	-11.3	-9.0	-6.8	-4.3	-1.7	1.3	4.5	6.6	13.2

EUR-ACCOR:TIIE DESIGN OF A MULTICHANNEL DATABASE*

ACCOR: A.Marchal, W.Hardcastle, P.Hoole, E.Farnetani, A.Ni Chasaide, O.Schmidbauer, I.Galiana-Ronda, O.Engstrand, D.Recasens.

CNRS, Aix-en-Provence, University of Reading, Ludwig Maximilians Universität, München, CNR, Padova, University of Dublin, Siemens AG, München, Universidad Politecnica de Valencia, Stockholm Universitet, Universitat Autònoma de Barcelona.

ABSTRACT

The EUR-ACCOR database has been designed tó allow access and analysis of articulatory, aerodynamic and acoustic characteristics of seven European languages: French, English, German, Italian, Catalan, Swedish and Irish. The data are acquired using a PC-based multi-channel system developed in Reading (UK) and Aix (France), which enables simultaneous recording of up to 16 input channels. The design of the corpus, consisting of (1) structured VCV nonsense items, (2) real words matching the structure of (1), and (3) short sentences, illustrating the main connected speech processes in the different languages, is discussed with reference to the main scientific goals of the ACCOR project.

1.INTRODUCTION

The ACCOR project aims to provide a detailed description of the main articulatory and acoustic correlations in coarticulatory processes. A cross-language approach is adopted as a means of identifying, on the one hand, the major language-independent universal regularities of these processes (due to factors such as the mass, inertia and elasticity of the speech organs, the mechanical linkage between them and the neuromuscular complexities of the cranial nerve system) and also how these interact with language specific factors such as the phonological rules of the language.

In carrying out such an investigation, activities of the main physiological systems underlying speech production are investigated: the respiratory system (producing a flow of air), the laryngeal system (modifying the airflow by the valving mechanism of the vocal folds) and the complex system of supraglottal structures in the mouth and nose, such as the tongue, lips, jaw and soft palate which shape the vocal tract into different resonating cavities The ACCOR project aims at a detailed description of the complex coordination between these different structures and the resulting acoustic output. Specific articulatory processes in the seven languages will be examined with a view to determining how such processes differ according to the different phonological systems. Also, it will be possible to examine the functions of the different motor subsystems in the same speaker.

The cross-language nature of the project has meant considerable attention has had to be paid to the design of a suitable database. Three main scientific goals of the project were considered in the design:

[i] to allow the testing of several hypotheses regarding articulatory/phonological interactions in the different languages.

[ii] to investigate fundamental issues in speech production theory particularly those relating to coarticulatory processes.

[iii] to provide a unique resource for researchers in speech science and related fields.

2. DESIGN CONSIDERATIONS FOR DATABASE

2.1. Testing Articulatory/Phonological Interactions

One of the main design considerations for the database has been to provide a rich enough corpus to allow the testing of various different hypotheses regarding interactions between articulatory characteristics and phonological patterns in the different languages.

The seven languages under investigation differ considerably in their vowel and consonant inventories, and it seems reasonable to expect that the structure of such inventories may constrain the type and extent of coarticulatory influences. For example, comparable consonants in the different languages may differ in their coarticulatory patterns because of the presence of other phonemes in the language. Italian, for example, contrasts alveolar with palatal nasals, and alveolar with palatal laterals; the two alveolars may be more resistant to the coarticulatory effects of a "palatal" vowel such as /i/, than another language where the possibility of perceptual confusion does not exist. Similarly, French, with the presence of nasal vowels in its inventory may exhibit less coarticulatory nasality than other languages without nasal vowels, such as German. It may be possible to formulate a general "density" hypothesis; coarticulatory processes vary inversely with the number of elements in the phonological class.

The database should allow the testing of more general hypotheses also such as :

- do some languages show greater carry-over/anticipatory coarticulation than others?
- does the same subject show different anticipatory/carryover effects for different motor subsystems (lip, jaw, tongue, velum)?
- do all languages exhibit a tendency towards "instability" of alveolar stops in connected speech?
- do all sounds show the same degree of contextually induced variability, or is there an increasing scale, of the order, say, of /s,t,d,n,l/, valid for all languages?

The VCV context would seem most appropriate for examining coarticulatory effects, as it is possible to study a variety of transitional phenomena in such a sequence: vowel-to-vowel effects, VC and CV effects. This context therefore forms the basis of the corpus. Another requirement for testing the above hypotheses is the inclusion of a range of sounds involving the different motor subsystems under investigation as well as voiced/voiceless distinctions. Thus, labial sounds such as /u/, /p/, /b/ were included as well as oral lingual consonants differing in both place and manner of articulation(e.g. /t,d,k, ʃ ,s,z,tʃ,/) and nasals such as /n/.

2.2 Speech production theory

The database has been designed to allow the investigation of fundamental issues in speech production theory. One area of considerable theoretical significance, for example, is the question of inter-gestural timing for consonants and the corpus has been designed to allow for the investigation of this aspect. The /kl/ sequence seems a particulary useful item to use for studying timing relationships, such as for example, between the tongue body gesture for the /k/ and apical contact for the /l/. There is preliminary evidence that the timing between the body and tip of the tongue in this sequence is language specific and may be related to prosodic aspects such as rhythm, syllabic structure and patterns of stop aspiration and voicing. Another area of significance is a specification of both acoustic/articulatory relationships in complex sounds such as fricatives. This

area is currently under investigation as part of Workpackage 2 of the project.

2.3 Unique resource for speech science

Recently a large number of digitized acoustic databases have been compiled for various purposes, (e.g. ESPRIT/SAM EUROM.O [1]; DARPA/TIMIT [2]). The EUR-ACCOR database is unique in that it includes articulatory and aerodynamic (nasal and oral air flow) as well as acoustic data in a number of different languages. The requirements of the database are such that it is necessary to simultaneously investigate the activity of different motor subsystems of speech, the resulting movement trajectories, aerodynamic effects and the acoustic output. To achieve this aim, the multi-channel data acquisition systems, EDIT and PHYSIOLOGIA have been developed at Reading [3] and Aix [4] respectively.

3. CORPUS

The study of coarticulation and its possible implications for the linguistic and neuromotor programming processes would seem a priori to exclude the use of nonsense items. On the other hand, although undoubtedly desirable, the analysis of spontaneous speech is problematic in that it is difficult to control all the parameters which may be responsible for coarticulatory effects. The emphasis on cross-language comparison makes this problem even more acute. In the ACCOR project we decided the best compromise was to work in parallel on three different types of speech material; structured VCV nonsense items, real words matching the nonsense items in their VCV structure as closely as possible, and connected speech in the form of a series of short sentences.

In choosing the items for inclusion in the corpus we were constrained by a number of criteria. Firstly, all items, including the nonsense items, had to comply with the phonological rules of the language. For vowel sounds, it was desirable to cover a range of different tongue postures and different perceptual characteristics. The three "extreme" vowels /i/, /a/ and /u/ were chosen as these vowels exist as phonemes in each language, and their phonetic manifestations in the different languages are similar. For the consonants, a range of different places and manners of articulation were chosen. It was decided to focus on VCV sequences for the reasons stated above. This raised the problem of stress placement. Obviously, it would have been preferable if the stress placement were uniform for each language. Unfortunately, however, language specific stress placement rules precluded this possibility if the items were to be produced "naturally". In all cases, except French and Swedish, stress was placed on the first syllable. In French the more natural placement is on the second syllable and in Swedish there is equal stress on each syllable. Details of the three corpora are as follows:

3.1. Nonsense items

Vowels /i,a,u/ in isolation. VCV sequences, where C = /p,b,t,d,k,s,z,n,l, ʃ tʃ / and the sequences /kl,st/; V = /i/, /a/ (/ ə / when unstressed in English) and /u/. These items are phonotactically permissable in all the languages under investigation, with one or two exceptions, (e.g. /z/ and / ʃ / do not occur in Swedish).

3.2. Real words

These match the VCV nonsense sequences in section 3.1. as closely as possible. Thus nonsense item /iti/ is matched by English "meaty", /uti/ in Italian is matched by "muti" etc. It was not possible to obtain a complete set of matching real words for all the languages.

3.3. Sentences

A set of short sentences was constructed in each language to illustrate the main connected speech processes in that language. Thus in English, sentences such as "Fred can go, Susan can't go, and Linda is uncertain" have numerous examples of word final alveolar stops and nasals which many speakers would assimilate into the place of articulation of following velars when spoken naturally.

Systematic comparison of data from the three types of recordings seems to us indispensable to take account of the specific nature of the processes involved and to resolve the inherent theoretical problems in establishing a multi-lingual corpus such as this. Moreover it should enable us either to show clearly the limitations of a single corpus approach or to vindicate the exclusive use of nonsense items. We are finding, in the initial results, that there are interesting and predictable differences between nonsense and real words, at least on some articulatory measures.

The items in 3.1, 3.2 and 3.3 constitute a core corpus which is read 10 times by each of the 10 speakers in the seven languages. Of the 10 repetitions, five are carried out with the Electropalatograph (EPG), Laryngograph and audio recording (sampled at 20 KHz), and five with EPG, nasal volume velocity of air, oral volume velocity, (using a Rothenberg mask and pneumotach), Laryngograph and audio recording. The core corpus will constitute approximately 300 Mbyte per speaker. In addition to the core corpus, a number of additional items are added for each language to illustrate particular language-specific features, e.g. geminates in Italian, front rounded vowels in German, palatals in Catalan, palatalisation in Irish Gaelic etc. The EUR-ACCOR database is well on schedule; approximately 5 speakers in each language have already been recorded and the first phases of data processing is nearing completion.

REFERENCES

[1] A.J.Fourcin, G.Harland, W.Barry and V.Hazan (Eds), "Speech input-output assessment; multi-lingual methods and standards", Chichester: Ellis Horwood, 1989.

[2] DARPA/TIMIT "Acoustic phonetic continuous speech database". CD-ROM Produced in the USA by PDO, 1988.

[3] A.Trudgeon, C.Knight, W.Hardcastle, G.Calder, and F.Gibbon, "A multi-channel physiological data acquisition system based on an IBM PC, and its application to speech therapy", Proceedings of Speech 1988 (7th Fase Symposium, Edinburgh: Institute of Acoustics, pp 1093-1100, 1988.

[4] B.Teston and B.Galiando "Design and development of a workstation for speech production analysis", Proceedings of VERBA Conference, Rome, January, 1990.

ACKNOWLEDGEMENT

We acknowledge financial support from The European Economic Commission DGXIII under the auspices of ESPRIT II Basic Research Action.

*The paper is presented on behalf of the ACCOR consortium

CRITICAL PARAMETERS IN THE DEFINITION OF SPEECH RECOGNISER PERFORMANCE

W. J. Barry

Department of Phonetics and Linguistics, London, UK

ABSTRACT

Six of seven "Critical Parameters" used in speech recognition assessment in the aproach known as "Recogniser Sensitivity Analysis" (RSA) are examined. An experimental study of phonetically controlled material confirms the vocabulary dependence of two of the parameters as currently defined. Basic principles of the parameter definitions necessary to assure maximum vocabulary independence are identified, and proposals for the redefinition of the Critical Parameters are presented.

1. INTRODUCTION

There have been several reports recently [4,6,7] which address the question of speech recognition assessment from the perspectives of 1) reducing the amount of speech data needed in the test database, and 2) predicting the performance of a recogniser in the field for given speaker and operational characteristics. Both these goals should be achievable if a relatively small database can be precisely defined along a number of critical parameters which specify important dimensions of speaker variability and which are not sensitive to variation in vocabulary. Situational characteristics can be simulated by post-processing "dry" recordings. This approach is known as "Recogniser Sensitivity Analysis" (RSA) [2]. Given information on the intended speakers, a controlled test of selected items from the assessment database should allow field performance to be predicted. The vocabulary independence of the parameters is crucial to the approach, otherwise every operational lexicon would have to be covered in the database.

Two projects concerned with speech recognition assessment (UK Alvey project MMI/132 - STA, and ESPRIT project 2589 - SAM) have been examining the relationship between recognition performance and the values of the test vocabulary along the critical parameters (CP), with some confirmation that variation in recogniser performance can be explained by variance in several of the parameters [7]. However, the evidence is not unequivocal [3]. This may well be a consequence of the recognisers selected for the trials, but the results are such that a scrutiny of the parameters and their definitions is warranted. This paper firstly undertakes this scrutiny, secondly presents results of an experiment which illustrates the inherent vocabulary dependency of two of the Critical Parameters as defined at present, and thirdly suggests some modified parameter definitions which are more appropriate to the underlying rationale of speech recognition.

2. CP DEFINITION

The present CP definitions are as follows [5]:

1. *Speaking rate:* Duration of one utterance relative to the overall average of all utterances of that word.
2. *Vocal tract area:* Average VTA over the frames of one utterance in relation to the overall average VTA for all utterances of that word.
3. *Temporal congruence:* Average DTW distance between all pairs of utterences of the same word by one speaker. The distance value applies to all utterances of a given word by that speaker; it is a measure of speaker consistency.
4. *Vocal effort:* Ratio of peak and average energy. Calculated for each utterance independently.
5. *Spectral definition:* Average (and variance of) ratio of energy < 2kHz and total energy, computed over each utterance independnetly.
6. *Fundamental frequency:* Mean F0 over each utterance independently.

A seventh parameter, vocabulary complexity, lies outside the focus of this paper.

3. TEST OF VOCABULARY INDEPENDENCE

The above definitions make it most likely, *a priori*, that "spectral definition" and "vocal effort" are strongly vocabulary dependent. Using the SAM_SPEX CP-extraction software, developed for use on the SAM PC-based workstation (SESAM) by Jutland Telephone according to the Logica algorithms [5], two sets of phonetically controlled words were analysed with respect to these two CPs:

1) 18 /hVd/ words spoken 3 times with different intonation contour by 4 speakers (2F, 2M). These were selected at random from the 10F and 8M speakers recorded in the Normative Reference database of the UK Alvey project MMI/132 (STA) [1]. These words were examined to test the effect of vowel variation on parameter values.

2) Five phonetically varied words selected from Logica's RSA recogniser test vocabulary [7] ("Aberdeen, Darlington, Manchester, Ipswich, sixteen") These were spoken 5 times each with 4 different voice qualities: modal, breathy, creaky and falsetto). With the combination of open and close vowels, and the presence or absence of consonants with fricative elements, these words tested the combined effect of changes in vowels and consonants on the parameter values.

Speaker variation was simulated in the second set by using the same, experienced speaker to record the words with the four radically different voice qualities.

4. RESULTS

The results may be summarised as follows:

1) Values in the /hVd/ words were critically dependent on the threshold setting chosen for endpointing. The 10% (of mean energy) threshold chosen as default for the tests was clearly too high to capture the tri-phonemic structure of these words. The parameters were, in effect, being calculated over the vowel portion alone. None of the words in the second set were affected by this threshold setting; it is solely a problem for words beginning with acoustically "weak" consonants (e.g. /h, f, v, T, D/- SAMPA symbols [8] are used throughout) and/or ending with such consonants, including weakly exploded stops. Reducing the threshold to 5% and 3% in two further analyses allowed the full /hVd/ word to be captured, but it is clear that this also increases the risk of including breath-noise and lip-smacks in the parameter estimation (and the recognition process) if they occur immediately prior to the actual utterance.

2) Two of the parameters, as defined at present: "spectral definition" and "energy" are dependent on the phonetic structure of the word.

A two-way ANOVA was performed on both sets of data for each parameter. For the /hVd/ words, "spectral definition" varied significantly with both speaker (F = 15.05, DF 2) and word (F= 15.06, DF 17). Energy" did not vary significantly across the words, inter-vowel intensity variation not being great or regular enough. For the second set of words, "spectral definition" again varied highly significantly across the words (F = 51.51, DF 4) and also reached significance for voice quality (F =

3.09, DF 3). With the "energy" parameter, there was significant interaction between voice quality and word (F = 11.54, DF 79).

The phonetic factors underlying these differences can be summarised as follows:

1. *Spectral definition* (Voice quality) (Energy < 2kHz/total energy): In the /hVd/ words, the main difference was between words containing mid to high front vowels and the others. This is due to the presence of strong higher formants, and especially the high second formant (> 2kHz) reducing the energy quotient. However, this variation was small compared to the value shifts across the words in the second set, where the presence or absence of fricatives (high frequency noise) changed values by up to 25%. 2. *Energy* (peak energy/mean energy) maximises the vocabulary effect and thus undermines speaker differences by relating peak energy (= peak energy of stressed vowel) to mean energy (which is proportionally lower, the more unvoiced consonants a word contains).

In the /hVd/ words, the values are most dependent on the personal strength of production, the intrinsic intensity of the vowels (open > close vowels) not contributing to any appreciable extent; in the second word-set, the greater complexity of the word structure contributes to a very strong interaction between differences in energy resulting from the different voice qualities and the word. For example, the consonants in "Ipswich" reduce the mean energy, and increase the quotient considerably despite the vowel /I/, which has relatively low intrinsic intensity. So, despite consistently lower quotient values for each individual word for the "breathy" than for the "modal" voice quality, the values for breathy "Ipswich" are higher than the "modal" values for the other words.

5. DISCUSSION OF DEFINITIONS

In the light of these results, it is important to consider how possible improvement in the word independence of these measures can be achieved. Firstly, for "spectral definition" and "energy" separating the main source of word-dependent variability, the voiced and voiceless portions, is an important prerequisite. For "spectral definition", only the voiced portions would be used in the calculation, reflecting the rationale behind the parameter, namely of capturing some aspect of voice quality. For "energy", mean voiceless intensity should be related to mean voiced intensity.

Secondly, for all measures, it is essential to avoid any utterance-dependent expression of peak value, mean value or variability. Speaker-dependent aspects of the parameters can only be calculated by relating the value for an individual utterance to the mean of all utterances of a particular word. In speaker-dependent recognisers, it may be sufficient to relate the parameter value for the individual utterance to the mean for all utterances of the same word/expression by the same speaker. In general, however, for speaker-independent systems the individual value has to be related to the mean of all utterances of the word/expression for all the other speakers (temporal congruence is the one exception being a speaker-dependent measure). This is already done with "speech rate" and "vocal-tract area" where the relativity of the measure was clear from the outset. Such normalisation should, however, be extended to "voice quality", "intensity" and "F0".

Thirdly, the "fundamental frequency" parameter in the form of *mean* F0 is dispensible, because it only separates male from female at present. This is information that is available independent of analysis. F0 variance, which is currently also being calculated, may be a more useful measure, since wide fluctuations in F0 could correlate with variation in recogniser performance; in Logica's recogniser tests [7], variance appears to be a more important factor than mean F0. The measure should, however, be related to the mean to avoid confounding Hz variance and the male-female distinction. The coefficient of variance (the quotient of mean and standard deviation) is therefore suggested. However, this measure still requires normalisation by relating it to the mean of the other speakers.

The principle of taking variance rather than mean values should also be considered for "vocal tract area". Although, in its present definition, it is already normalised, it is the mean of the frame-based area coefficients which are being normalised. The use of variance would relate the parameter more closely to the recognision process. It is, after all, *variation* of signal properties during the course of a word which makes it more or less distinctive to a recogniser not its *mean* sig-nal properties. Variance in vocal tract area for a given utterance related to all speakers' variance for utterances of the same word or expression would differentiate speakers with clear and less clear articulation.

6. CONCLUSIONS

The results from the CP analysis of phonetically controlled vocabulary confirmed the vocabulary dependence of two of the Critical Parameters as currently defined for RSA. It was concluded that parameter normalisation across all utterances of a word is a fundamental pre-requisite for the vocabulary independence of Critical Parameters. Further consideration of the basic rationale underlying automatic speech recognition points to the need to redefine all parameters except "temporal congruence" to contain an expression of normalised variance.

7. REFERENCES

[1] FULLER, H., FOURCIN, A.J., GOLDSMITH, M.J. and KEENE, M. (1990): A Database of Normative Speech Recordings. *Proceedings Institute of Acoustics* 12, part 10, 1-6

[2] KNIGHT, J.A. and PECKHAM, J.B. (1984): *A Generic Model for the Assessment of Speech Input Applications*. Logica Report for RSRE.

[3] KORDI, K. (1990): *Field Trial Report*. Alvey Project MMI/132, Logica, November 1990

[4] PECKHAM, J., THOMAS, T. and FRANGOULIS, E. (1989): Recogniser Sensitivity Analysis: Trial Results and Future Directions. *Proc. ESCA Workshop, Speech Input/Output Assessment and Speech Databases*, 4.2.1-7, Noodwijkerhout.

[5] THOMAS, T.J. (1988): *Algorithms Used for Parameter Extraction for Recogniser Sensitivity Analysis*. Project Report. Alvey Project MMI/132, Logica, July 1988

[6] THOMAS, T.J. (1989): A Determination of the Sensitivity of Speech Recognisers to Speaker Variability. *Proc. ICASSP*, S10b.8, 544-547

[7] THOMAS, T.J. (1990): *The Sensitivity of a Speech Recogniser to Speaker Variability*. Project Report. Alvey Project MMI/132, Logica, Cambridge, September 1990

[8] WELLS, J.C. (1988): Computer Coded Phonetic Transcription. *J. International Phonetic Association* 17 no. 2, 94-114.

SPEECH KNOWLEDGE,STANDARDS AND ASSESSMENT

presented on behalf of the SAM consortium by

Adrian Fourcin and Jean-Marc Dolmazon
University College London & Institut de Communication Parlee

INTRODUCTION

This brief overview is designed to provide background information for the poster related to the work of the SAM 'Speech Assessment Methods' Project (2589) - concerned with the design and application of multi-language EC standards. At present the project is based on the collaboration of twenty-six laboratories in eight countries, six within the EC and two from EFTA. The project is now at the start of its third year in ESPRIT II. This follows a preliminary 'Definition Phase' (ESPRIT 1541) in which the status of work in the area, and the requirements in Europe and the rest of the world were investigated, and a 'bridging' 'Extension Phase', in which preparatory work for the Main Phase was undertaken.

Current work is in progress in three inter-connected working areas:

I Speech Recognition Assessment (Input)

II Speech Synthesis Assessment (Output)

III Enabling Technology and Research (ETR)

At the beginning of the SAM Project, the need to ensure a practical basis for ready collaboration between so many different laboratories in different countries was met by the definition of a reference, standard, workstation - SESAM. The minimum hardware requirements for SESAM are an IBM pc-at or compatible computer, an analogue interface board, OROS-AU21 or AU22, 1 Mbyte of extended memory, and means for accessing speech data eg CD-ROM reader. C is used as the common programming language. Each one of three workgroups, above, has made use of this simple reference standard so that software, data and assessment results can be interchanged. This has proved to be very successful both between project members and in the provision of data and support for other laboratories across Europe - all the work of the SAM Project is designed to be readily available within the European Community.

I INPUT ASSESSMENT

In recognition assessment, the simple reference standard workstation has been implemented and tested in multi-lingual, multi-laboratory trials.

In order to provide a flexible tool for recogniser assessment, the component software packages are designed as separate modules which can be independently developed by different laboratory groupings within the project. The first package, PAOSAM, is designed to be capable of managing the information associated with the standard SAM format speech databases. The second package, EURPAC, primarily controls the interaction between the assessment system and the recogniser itself and the third package. This last package, SAM_SCOR, provides a series of performance measures. All three software modules are interconnected via ASCII files, and all programs are in C using the microsoft 5.1 compiler, and executable on the SESAM workstation running MS-DOS as the operating system

Database Management

The RISE program has been developed to cater for the major needs of data retrieval and data archiving for all languages and all speakers in the SAM project. A commercially available DataBase Management System (ORACLE) is used as the basic building block. The management structure has been designed to allow the integration of all of the characteristics of both present and future SAM speech databases. Effectively, RISE enables the user to specify the characteristic assessment aspects to be targetted in terms, for example, of language, speaker and speech types, and for an automatic procedure to be utilised for the composition of training and testfiles in the assessment of a defined recogniser.

Control Module

The EURPAC program is designed to operate from this basis in controlling the assessment of isolated or connected word recognisers. The assessment session can be controlled by information given in a separate control file, defined by the user, and giving details of the unique serial number of the test run, the identification of the recogniser, and the names of the configuration-, training-, test- and response-files. An important aspect of the design of this particular software module is that it uses resident drivers to control individual recognisers. In this way, the greater part of the software is quite independent of the analogue interface board which is utilised, and it is easier to develop new recogniser drivers which can have separate communication protocols.

Scoring

The SAM_SCOR program provides a range of recognition performance measures - hit; miss; substitution; correct rejection; false alarm. In addition, at the isolated word level, confusion matrices, confidence analyses, and the application of the McNemar test are standard facilities. For connected word and continuous speech recognisers string matching at the orthographic level is available employing NIST scoring routines which have been made executable on the SESAM workstation. The output of this scoring software is designed to provide uniform presentations of the assessment results that are easy to understand and cross compare. SAM_SCOR generates a file which can subsequently be fed back into the DBMS to make it possible to relate speech material characteristics to recogniser performance measures.

Applications

More than 10 EC laboratories in the Project have been involved in the application of recogniser assessments so far for six commercially-available or in-house recognisers. Considerable use has been made of the first SAM CD-ROM speech database - EUROM 0 - which gives 5 hours from 20 speakers in five languages. This cross laboratory single and multi-language testing of equivalent recognisers has provided the foundation for the setting up of a basic calibration procedure for the SESAM input assessment workstation. Work is currently in progress to define a common method for standard reference calibration and hardware setting up protocols.

In collaboration with the ETR Group, a new multi-lingual speech database has been designed and is in the process of being recorded. The contents of the database have been defined to meet the present and near future need for the development of diagnostic and predictive assessment methodologies. The database is divided into two sets: a 'Many Speaker set' and a 'Few Speaker set'. The vocabulary of the 'Many Speaker set' contains a list of selected numbers between zero to nine thousand nine hundred and ninety nine covering all the phonotactic possibilities of the languages' number systems, and blocks of five sentences giving continuous speech with paragraph prosody rather than individual sentences. The vocabulary of the 'Few Speaker set' is expanded with a CVC list and more repetitions per item.

II OUTPUT ASSESSMENT

Standard word-level and sentence-level segmental multi-lingual intelligibility tests have already been defined. They can be automatically generated on the SESAM workstation in the languages of the project using phonotactic and word frequency constraints. Compatible software provides for response collection, collation and scoring.

Segmental Structures

The SAM segmental test contains guidelines for the automatic generation of nonsense-word lists for all eight partner languages, using a set of fixed word structures and phoneme lists. The test material is language specific in that phoneme combinations respect phonotactic constraints for the languages in which they are prepared. The SAM group has chosen to use nonsense words in its definition of this standard with an open response set in order to get an intelligibility score which is not influenced by contextual information or semantically restricted answer choices. This type of material is the most relevant when an analysis of phoneme confusions is required, and in application, for instance, to synthesis material where error patterns may be quite device-specific. The SAM Segmental test consists of two parts: a first "core test" containing structures common to all languages of the consortium, and which cover consonants in initial, medial and final positions: VCV, VC (+ fixed final V for Italian) and CV. In all cases, the full inventory of consonants is used with only a sub-set of vowels. This sub-test cannot be considered as a *full* diagnostic test, but it is substantially diagnostic for consonants. The "full test" will include more extensive language-specific and even synthesiser-specific sub-tests with complex structures such as CVC, CVVC, VCCV and possibly CnVC, CVCn. Phonemes are presented in equal numbers per list so that an equal probability score will be obtained. This score can then be weighted according to phoneme-frequency-of-occurrence counts to obtain scores which reflect phonemic balance.

Segmental Assessment

A system to support the automatic segmental assessment of synthesisers has been implemented on the SESAM workstation. The subject responds using the keyboard, results are then automatically scored, to produce percentage scores, confusion matrices, analysis in terms of certain types of phonetic feature, eg place of articulation and voicing, and the effect of vowel environment.

Assessment of words in context

A test of word intelligibility in sentence context has also been developed for SESAM, using semantically unpredictable sentences (SUS). Grammatical structures and word lists are defined for all the languages of the consortium to permit the generation of an unlimited number of test sentences. Work on prosodic assessment is also in progress.

III ENABLING TECHNOLOGY AND RESEARCH

The core SAM workstation, SESAM, has been specified and implemented for data collection, following standard protocols, database management, and speech signal labelling. A phonemic notational system for all European languages, SAMPA, has been developed and is in use both for manual labelling and, currently, for semi-automatic label alignment. Phonemic level structural constraints across the languages of the project have been compiled and are used in corpus definition. Broader descriptors are being investigated for multi-lingual application. Other, physical, levels of description are being quantified as a contribution to analytic methods of assessment. Information on cross-language lexica is being compiled.

SESAM

Hardware (see the INTRODUCTION above) and software specifications are now well established and widely applied in regard, for example, to: the structure and code normalisation of software; the formatting of data and organisation of databases; and the provision of interfaces.

Two, key, software packages are central to the use of the workstation within the project. The first is EUROPEC, which is designed to provide for the realisation of large speech databases. Two-channel acquisition (eg for microphone and laryngograph signals) and monitoring is now possible with visual prompting for the speaker which may be manually controlled or automatically triggered as a function of signal level. Automatic end-point detection facilitates the handling and recording of large organised corpora. This is also substantially assisted by the automatic

inclusion in the database of description text files in standard form with header and body, so that the orthographic prompt can be routinely incorporated together with complete sessional and recording item and condition information.

A complementary package, VERIPEC, is designed to give ready access to these standard data and text files, making it possible to display the orthographic prompts, access and monitor recorded items and show the label files.

The second important package, PTS, is designed fo operate from the data acquired via EUROPEC. Its primary function is to enable the labelling of speech data files with either SAMPA (see below) or IPA notations, using window based displays of waveform and spectrograms of the signal.

Data

The first SAM database, EUROM 0, was distributed on a single CD-ROM and contained five hours of speech material recorded using a condenser microphone in anechoic rooms from four single accent speakers in each of five languages (with 16 kHz sampling). NATO single and triple digit sequences were obtained with only the speech signal, and a continuous speech passage, with a common numeric theme across languages, was recorded using two channels - with both speech and laryngographic inputs.

A new database is now in preparation using the same standard format with sixty speakers in each of eight languages. Nonsense words, number sequences up to 9999, (both phonotactically balanced) and situationally linked sentence blocks are being recorded. Anechoic condenser microphone recordings will permit the subsequent imposition of post-production effects. A small subset of data will have two channel representation, as above. Two CD-ROMs are planned for each language - using 20 kHz sampling.

SAMPA

The SAM Phonetic Alphabet (SAMPA), which defines a standard keyboard based notation (ASCII) corresponding to the relevant International Phonetic Association symbols for each of the languages represented in the project, was agreed very early in the project, and has now been extended to cover all the major European languages. It has also been adopted by a number of ESPRIT projects and both the British and German national speech databases. This consensus for the representation of phonemic contrasts in all the languages of the group provides a common labelling basis for cross-comparison and for a structured multi-lingual approach to database specification in the development of standard methods of assessment. The basic SAM transcription system was originally intended to evolve as a multi-tier labelling tool and work is currently directed towards the introduction of prosodic and acoustic element levels of description.

Labelling

Multi-lingual labelling, in which phoneme categories are assigned to successive regions of the speech signal, has always been an important part of the SAM group's activity. This is because overall assessment, detailed evaluation and the processes of training themselves ultimately depend on an accurate definition of speech which can be given in phonetic and orthographic terms. So, although the precise assignment of discrete categories, for different sound classes, to the continuous speech signal is an impossible task - since the subjective level of labelling is not compatible with any physical set of exact temporal stretches of the signal - the consistent correlation is of real value. The SESAM workstation is designed to support this work, and manual labelling in all the languages of the project has provided essential reference material.

A further development of this work currently involves a semi-automatic approach, using label alignment. In this way the larger quantities of speech material generated by current database gathering, and which are in need of labelling, can be accommodated without imposing an impossibly large manual task.

In Conclusion - The project has provided an opportunity for multi-language work in Phonetic Sciences,which would not otherwise have been possible and we are glad to stress the collaboration and goodwill which have made the work, across Europe, so effective.

SCALING OF SPEECH INTELLIGIBILITY USING MAGNITUDE AND CATEGORY ESTIMATION AND PAIRED COMPARISONS

S.C. Purdy and C.V. Pavlovic

University of Iowa, Iowa City, U.S.A. &
University of Provence, France

ABSTRACT

Subjects judged speech intelligibility using either magnitude estimation, category estimation, or paired comparisons. Speech scores for CID Sentences and NU-6 words were also obtained. The speech was bandpass filtered so that a monotonic increase in intelligibility was predicted by articulation index (AI) theory. The reliability and sensitivity of intelligibility judgments and speech scores were compared. The validity of intelligibility judgments was investigated by comparing judged intelligibility to AI predictions.

1. INTRODUCTION

Psychophysical scaling procedures have been used to investigate the intelligibility and quality of hearing-aid transduced speech. These include the methods of paired comparisons (PC) in which subjects judge which of two stimuli has more or less of the attribute being investigated [4, 8], magnitude estimation (ME) in which subjects choose any positive number to represent the subjective magnitude of the attribute present in the stimulus [9, 10], and category estimation (CE) in which subjects rate their impressions by choosing numbers or adjectives from a fixed range of scale values [1, 5]. Subjects can make judgments of speech intelligibility or quality that differentiate reliably among hearing aids. Subjective judgments may be more sensitive to hearing aid differences and more reliable than speech recognition scores [4, 8, 10].

We investigated the reliability, sensitivity, and validity of speech intelligibility judgments obtained using ME, PC, and CE. Validity was tested by comparing intelligibility judgments to AI predictions. According to AI theory, intelligibility is monotonically related to the articulation index: $A = \sum_{i=1}^{n} I_i W_i$, where I_i describes the importance of each frequency band for speech intelligibility, and W_i is a weighting function representing the speech dynamic range contributing to intelligibility [6]. Since the exact form of I_i was not known for our speech, we chose filter settings that produce a monotonic increase in AI for speech materials spanning the probable range of importance functions [6]. Sensitivity was investigated by determining how well procedures differentiated filter conditions. To assess reliability subjects were tested twice

2. PROCEDURE

Thirty subjects aged 60-87 years with hearing better than 30 dB HL at 500-2000 Hz were tested. The 60+ age group was chosen as representative of the majority of hearing aid wearers. Subjects were divided into three groups using ME, CE, and PC, respectively.

The speech was filtered using eight bandpass filter settings: 510-920 Hz, 510-1000 Hz, 510-1100 Hz, 630-1500 Hz, 770-2000 Hz, 700-2000 Hz, 630-2000, and 570-2000 Hz, producing a monotonic increase in the AI for nonsense syllables, easy speech, and average speech [6]. Spectrum density levels were computed for the sentences used for intelligibility judgments, NU-6 words, and CID Sentences. Assuming a +12 to -18 dB dynamic range, the entire signal was audible for all subjects.

Subjects judged the intelligibility of sentences from Grade 8 English texts. A commercial recording of the NU-6 lists was used. Sentences for intelligibility judgments and CID sentences were recorded using an male talker who spoke standard American English. The speech was presented via a TDH-50P earphone. Half the subjects doing ME and half doing CE were given the CID Sentence test. The remaining subjects in these groups were given the NU-6 test. On each visit subjects make 16 practice and 56 test judgments.

For ME subjects assigned a number to match the intelligibility of the sentence. No modulus was used. For CE subjects rated intelligibility using a 20-point scale. Number 1 was marked "very very unintelligible" and number 20 was marked "very very intelligible". For PC each stimulus was paired twice with every other stimulus, with order randomized. Subject decided which sentence was more intelligible. Intelligibility was defined as "..how well you understand the sentence".

3. RESULTS

Visit 1 intelligibility judgments are shown in Figs. 1-3. PC preference scores are the number of times that a filter condition was judged more intelligible than the other in the pair. NU-6 and CID word scores are shown in Figs. 5-6. NU-6 phoneme scores were also analysed.

To compare test-retest reliability Pearson correlation coefficients were calculated between visit 1 and 2 data for each subject. R values (0.51-0.99) were transformed [2] and the means were compared using a Bonferroni-adjusted significance level. Mean test-retest correlations for ME and CE judgments and NU-6 word and phoneme scores did not differ significantly. ME, CE, and NU-6 correlations were higher than correlations for PC and CID Sentences. Test-retest reliability was also investigated by calculating intraclass correlations between visit 1 and 2 data Intraclass correlation coefficients show the absolute similarity between pairs of values. There were no differences in intraclass reliability between ME, CE, and PC. CE and PC judgments and NU-6 scores were significantly more reliable than CID scores.

Relative sensitivity was investigated by

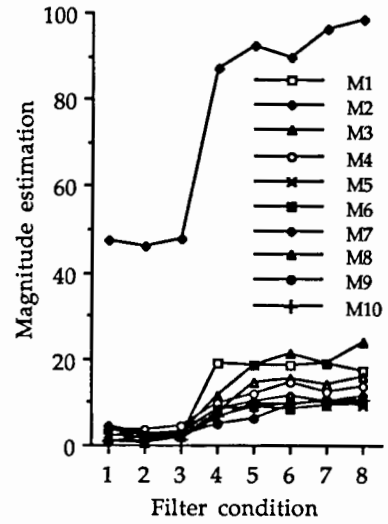


Fig. 1. Individual visit 1 magnitude estimations of speech intelligibility. Each magnitude estimation is the geometric mean of seven judgments.

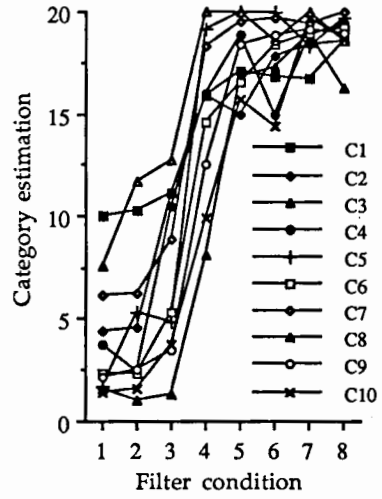


Fig. 2. Individual visit 1 category estimations of speech intelligibility. Each category estimation is the arithmetic mean of seven judgments.

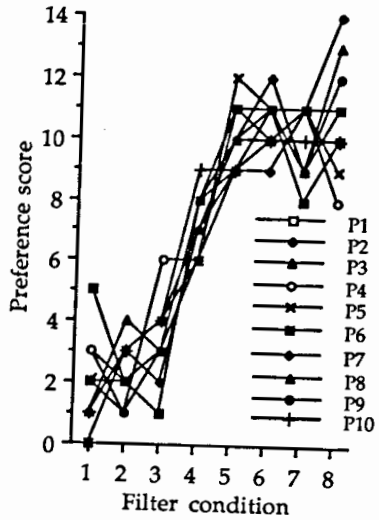


Fig. 3. Individual visit 1 paired comparison judgments of speech intelligibility. Preference scores are the number of times that a filter condition was judged more intelligible.

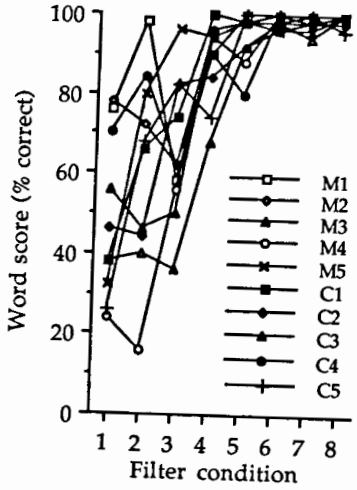


Fig. 4. Individual visit 1 word recognition scores for the CID Everyday Sentence test. Each point is the % of keywords correctly identified in a list.

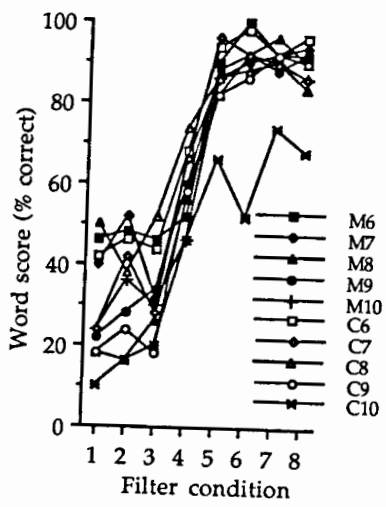


Fig. 5. Individual visit 1 NU-6 word scores. Each point is the % of words correctly identified in a list.

calculating Spearman rank order correlations between intelligibility judgments and AI rankings of intelligibility for the eight filter conditions, and between speech scores and AI ranks. T-tests performed on the transformed r_s values showed no significant differences in sensitivity among the procedures. Because of concern that the lack of statistical power of rank order methods may have disguised any real differences in the data, similar tests of sensitivity were conducted using Pearson's r. Speech scores and intelligibility judgments were correlated with the AI values for nonsense syllables, easy speech, and average speech. For ME, CE, PC, and NU-6 scores, mean correlations were 0.82-0.95. Correlations were generally lower for CID Sentences (0.75-0.86). For all three speech materials there were no significant differences between the correlations for ME, CE, PC, and NU-6 scores. However, intelligibility judgments and NU-6 scores correlated better with the AIs than did CID scores. That is, the CID Sentence test was least sensitive to differences among conditions.

AI theory predicts that, on average, intelligibility should increase monotonically across conditions. Accordingly, if subjective judgments are

valid, intelligibility judgments should be highly correlated with AI rankings of intelligibility for the filter conditions. Even if AI theory is disregarded, an increase in intelligibility is predicted as bandwidth increases across conditions 1 to 3 and 5 to 8. Rank order correlations between the mean intelligibility judgments and the AI rankings were 0.98-1.0. Thus, although individuals confused similar filter conditions, on average there was a monotonic or near-monotonic increase in judged intelligibility across filter conditions, as predicted by AI theory.

4. DISCUSSION

We compared the reliability, sensitivity, and validity of three scaling procedures and two clinical speech recognition tests. Both analyses of test-retest reliability showed that reliability was poorest for CID Sentences. Pearson r values showed that the test-retest reliability of PC judgments was poorer than ME or CE.

There were no differences in sensitivity among the scaling or speech recognition measures based on the analysis of rank order correlations. However, the analysis of Pearson correlation coefficients showed that intelligibility judgments and NU-6 scores were more sensitive to differences between conditions than CID Sentence scores, presumably because CID scores showed the most severe ceiling effect. This is consistent with earlier studies in which speech tests were less sensitive to differences between hearing aids than intelligibility judgments [4, 10].

There were no differences in reliability or sensitivity between NU-6 scores and ME and CE judgments, suggesting that ME or CE could be used instead of speech tests for hearing aid selection. A major advantage of ME and CE procedures over the NU-6 test (and other speech recognition tests) is their efficiency.

The high correlations between mean intelligibility judgments and AI rankings of intelligibility and between mean intelligibility judgments and speech scores suggest that subjects were judging intelligibility and not some other aspect of the speech signal. The good agreement between the data and AI predictions indicate that judgments were primarily based on intelligibility. This is consistent with previous evidence for the validity of intelligibility judgments [1, 3, 5, 7, 8].

5. REFERENCES

[1] COX, R. M., and McDANIEL, D. M. (1984), "Intelligibility ratings of continuous discourse: Application to hearing aid selection", *Journal of the Acoustical Society of America, 76*, 758-766.

[2] FISHER, R. A. (1921), "On the 'probable error' of a coefficient of correlation deduced from a small sample", *Metron, 1*, 1-32.

[3] GRAY, T. F., and SPEAKS, C. E. (1978), "Ability of hearing impaired listeners to understand connected discourse", *Journal of the American Auditory Society, 3*, 159-166.

[4] LEVITT, H., SULLIVAN, J. A., NEUMAN, A. C., and RUBIN-SPITZ, J. A. (1987), "Experiments with a programmable master hearing aid", *Journal of Rehabilitation Research and Development, 24*, 29-54.

[5] NAKATANI, L. H., and DUKES, K. D. (1973), "A sensitive test of speech communication quality", *Journal of the Acoustical Society of America, 53*, 1083-1092.

[6] PAVLOVIC, C. V. (1987), "Derivation of primary parameters and procedures for use in speech intelligibility predictions", *Journal of the Acoustical Society of America, 82*, 413-422.

[7] SPEAKS, C., PARKER, B., HARRIS, C, and KUHL, P. (1972), "Intelligibility of connected discourse", *Journal of Speech and Hearing Research, 15*, 590-602.

[8] STUDEBAKER, G. A., BISSET, J. D., VAN ORT, D. M., and HOFFNUNG, S. (1982), "Paired comparison judgments of relative intelligibility in noise", *Journal of the Acoustical Society of America, 72*, 80-92.

[9] STUDEBAKER, G. A., and SHERBECOE, R. L. (1988), "Magnitude estimations of the intelligibility and quality of speech in noise", *Ear and Hearing, 9*, 259-267.

[10] TECCA, J. E., and GOLDSTEIN, D. P. (1984), "Effect of low-frequency hearing aid response on four measures of speech perception", *Ear and Hearing, 5*, 22-29.

This work has arisen as a result of collaboration between members of the SAM project (Esprit Project 2589) and the University of Iowa.

KNOWLEDGE-BASED ACOUSTIC-PHONETIC DECODING OF SPEECH : A CASE-STUDY WITH THE APHODEX PROJECT

J.P. Haton

CRIN-CNRS / INRIA, Nancy

ABSTRACT

The APHODEX project aims at investigating the role of Artifical Intelligence knowledge-based reasoning techniques in the acoustic-phonetic decoding (APD) of continuous speech. This paper constitutes an evaluation of this project. It briefly presents the present state of the APHODEX system and concentrates of some issues of APD that were raised during the project regarding several aspects : segmentation, amount of knowledge necessary for APD, choice of a proper unit decoding strategy.

1. INTRODUCTION

The acoustic-phonetic decoding of speech consists in automatically mapping the semi-continuous acoustic speech wave into a set of predefined discrete linguistic units such as phones, phonemes, syllables, etc. This is a very difficult operation which constitutes a major level in automatic speech recognition, especially for continuous speech, multi-speaker operation [1]. Despite substantial advances the problem has not yet received a totally satisfactory solution. One reason for that might be that APD makes it necessary to take into account a large body of information : data, facts, knowledge, contexts, etc. Following this idea, we launched in 1984 the APHODEX project with the aim of investigating to which extent the knowledge-based methodology developed in Artificial Intelligence may be helpful in solving the problem of APD.

After a brief recall of the different approaches to APD proposed so far we summarize the main features of APHODEX. We then present the major issues in APD in light of the experience we gained during the project.

2. APPROACHES TO ACOUSTIC-PHONETIC DECODING

The task of APD is of crucial importance in analytic speech recognition since the overall performances of any sentence recognizer depends heavily upon the quality of the phonetic decoding. The role of the acoustic-phonetic decoding level in speech recognition is threefold :

- extraction of pertinent features and cues from the acoustic signal,
- segmentation of the speech continuum into phonetically meaningful units such as phones, phonemes, syllables, etc.,
- labeling of the segments with fine phonetic labels and/or gross phonetic classes.

These three different tasks are highly interrelated and must moreover interact with the other linguistic processing levels in order to come out with the best possible phonetic lattice.

APD was initially considered as a simple pattern recognition problem. But the actual size and difficulty of the task were then clearly identified, particularly in relation with the major importance of coarticulation and context effects and of speech variability.

Present approaches to APD belong to three main categories :

- *stochastic modelling* [2] : the problem of optimally matching an input utterance against every possible concatenation of phonetic units can be expressed in terms of stochastic processing, especially in the framework of hidden Markov models. Such models make it possible to capture in a statistical way the variability of speech by processing huge amount of data. They provide one of most efficient framework for multi-speaker APD ;
- *connectionist neural-like modelling* [3], [4] : neural networks are experiencing a new growth of interest in different fields of Artificial Intelligence, including APD. Basic models (multi-layer perceptrons, Boltzmann machines, etc.) have been adapted to speech requirements, especiallyfor taking into account the inherent temporal nature of the speech phenomenon. New models more closely related to neuro-biological data have also been proposed, e.g. phonotopic maps or cortical columns [5]. Neural networks have achieved good performances in APD and represent a promising approach both for phonetic labeling and for preprocessing of speech data ;
- *knowledg-based reasoning* [6], [7] : the use of knowledge-based reasoning techniques is an alternative to the two previous statistically based approaches to APD. The major difficulty lies in the elicitation and formalization of a proper body of knowledge. Such techniques are used in the APHODEX project that will now be described in more details.

3. OVERVIEW OF APHODEX [8]

3.1. Basic ideas and motivations

Phonetic decoding by reading speech spectrograms is typically a knowledge intensive activity during which an experienced phonetician conducts an explicit and contextual reasoning based on the knowledge and expertise he gained by experience [9]. It seems therefore fruitful to elicitate and formalize this knowledge through a close interaction with a phonetician and by using the methodology developed in Artificial Intelligence for the design of knowledge-based systems. This idea was the basis of the APHODEX project when we started it in 1984. We considered at that time that the conjunction of the knowledge of an experienced spectrogram reader, François Lonchamp, on one hand and of our know-how in automatic speech recognition and knowledge engineering might help progressing in APD. Our main motivation was to gain a better understanding of the process of phonetic decoding and underlying processes. Another motivation was to implement an experimental knowledge-based system capable of carrying out the phonetic decoding of continuous speech in a multi-speaker way. The present state of this system together with its performances will now be briefly described.

3.2. Knowledge and architecture

Thanks to an in-depth study of the activity of spectrogram reading by our phonetician (cf. section 4.2.) we gathered a large body of procedural and declarative knowledge. This knowledge was then implemented in the APHODEX system into two forms :

- *procedures* coded in several pre-processors that operate directly on the speech wave and perform a coarse segmentation into phonetic segments as well as a classification of segments into broad phonetic classes (vocalic, fricative, plosive and others). Performances obtained so far are about 95 % of correct segmentation in the best cases ;
- *production rules* which constitute the knowledge base of an expert system. The inference engine of this expert system carries out a reasoning similar to the one developed by a phonetician in order to label each segment on a phonemic basis and, if need be, to refine the broad classification done by the pre-processors.

Here is a typical example of rule :

IF Segment is Plosive AND Burst spectral maximum is between 3000 Hz and 4500 Hz AND Right context is /i/ ou /e/ THEN /k/.

It is worth noticing that most rules are contextual (for instance here, the right context of the segment to be labeled must be an unrounded front vowel, i.e. in French /i/ or /e/). That enables the inference engine to carry out a contextual reasoning and to propagate constraints (phonetic, phonological, etc.) throughout the process in order to finally come out with an optimal phoneme lattice. All the acoustic events mentioned in the rules (formant trajectories, burst features, friction, etc.) are extracted automatically from the speech signal by robust, speaker-independent procedures.

Experimental results show that APHODEX is capable of decoding a sentence pronounced by any unknown French male speaker with a mean accuracy of 70 %. This percentage will progressively increase when new rules are added to the knowledge base. Comparatively, several experiments carried out for English and for French have shown that an expert spectrogram

reader can reach as much as 85 % of correct labeling.

4. ISSUES IN ACOUSTIC-PHONETIC DECODING OF SPEECH

We will now present some important issues in APD and propose some elements of solution that we developed in the framework of the APHODEX project.

4.1. Segmentation of the speech wave

The continuity of the speech signal is a major difficulty of speech recognition. A segmentation is therefore necessary in order to extract units on which the labeling process will then work. The problem is two-fold :

- firstly choose one or several proper units which can be either of infra-phonemic level (phones) or of phoneme level, or else of supra-phonemic level (diphones, syllables, triplets),
- then implement a segmentation process based on the temporal evolution of acoustic-phonetic features that must yield a solution as valid and consistent as possible.

The examination of some errors made by the segmenter of APHODEX led us to propose a hierarchical multi-segmentation solution, based upon the strategy used by the human expert. This method consists in building up a multi-level segmental representation of a sentence (dendrogram) with the help of a spectral difference function. This structure is then pruned out by using acoustic cues in order to yield the final segmentation which might be unique (in non-ambiguous cases) or multiple. This pruning is carried out in close interaction with the rule-based reasoning process.

4.2. Data and knowledge gathering

As stated previously the phonetic decoding of a sentence necessitates a large amount of knowledge of various types : articulatory, acoustic, phonetic, phonological, etc. This knowledge can be implicitly integrated in a system by the examination of huge amounts of speech data, as in stochastic or connectionist models. Despite the good performances obtained by such models, it seems nevertheless necessary to design some model for the explicit storage of knowledge. It seems that a knowledge-based reasoning APD can be more easily interfaced with other processing levels of a speech understanding system (for instance the feedback from the lexical level to the phonetic decoding). Moreover, this solution provides a convenient framework for gathering all available pieces of data and knowledge related to APD (the expert knowledge involved in spectrogram reading being only one aspect). The tools and methodology provided by artificial intelligence makes it easier to incrementally build up a kind of «collective memory» of APD for a given language. That constitutes a necessity for the future of research on speech communication.

4.3. Choice of a processing unit

The choice of a processing unit, for segmentation and for labeling, is of primary importance in the design of an APD system. As a matter of fact several units can be used at different steps of the process. The present version of APHODEX is based on a phoneme-like unit. This choice was motivated by the fact that the phonetician uses this unit through out his activity of spectrogram reading. Another feature interesting on a practical point of view is the limited number of phonemes which are necessary for the description of a language. However a phonemic unit presents serious drawbacks for APD, especially due to the context dependancy of phonemes that necessitates a very large number of rules to take into account the various contexts in which a phoneme has to be identified. That led us to adopt another processing unit, the phonetic triplet which an be defined as a phone with its phonetic contexct [10]. A large amount of work is still to be done in order to collect a set of triplets representative of the language but we nevertheless consider this units as a good compromise for APD.

4.4. Decoding control strategy

The APD reasoning operation must be controlled by an efficient strategy in order to avoid unnecessary hypotheses and to focus on relevant data. We developed in APHODEX a strategy inspired from the observation of the spectrogram reader who operates in two successive steps (cf. § 3.2.) A majority of APD systems use only a bottom-up strategy (from the acoustic data to phoneme labels). However a top-down strategy is also useful in order to verify hypotheses or to interact with the linguistic levels. In APHODEX the two types of control are used concurrently in an opportunistic manner. More work is still needed in order to design more sophisticated strategies similar to those used by an expert in difficult or ambiguous cases.

It is often necessary to make assumptions about the phonetic content of an utterance and to emit alternative, competing hypotheses about the succession of sounds. That corresponds to a kind of hypothetical reasoning for which specific techniques have been developed in AI in order to maintain the overall truth and consistency of the deductions made during the decoding. We are implementing hypothetical reasoning in APHODEX, based on various types of knowledge including phonological variations. This method gives substantial improvements in the decoding accuracy, especially when there is some ambiguity or when contextual effects are important.

Two important lessons drawn from the examination of spectrogram reading activity concern the strategy of decoding. The first one consists in systematically relying phonetic labeling decisions to several acoustic features rather than a single one. The second can be called delayed decision strategy since it consists in postponing decisions until enough evidence has been accumultaed in favor of a particular label.

5. CONCLUSION

This paper has dealt with some aspects of a major problem in automatic speech recognition, i.e. acoustic-phonetic decoding of continuous speech. We have especially presented the usefulness of Artifical Intelligence knowledge-based techniques in this area and discussed important issues in the light of the APHODEX project developed in our group.

Despite the very good performances obtained so far in APD by statistical models, we consider that knowledge-based techniques have some usefulness both for gathering relevant data and knowledge and for implemnting practical systems. An explicit knowledge-based reasoning in APD also makes it easier to implement feed back links from linguistic processing levels to APD.

6. REFERENCES

[1] KLATT, D. (1977), "Review of the ARPA Speech Understanding Project", *JASA*, 62, pp. 1345-1366.

[2] SCHWARTZ, R.M. et al. (1984), "Improved Hidden Markov Modeling of Phonemes for Continuous Speech Recognition", *Proc. Int. Conf. ASSP*.

[3] BOURLARD, H., WELLEKENS, C.J. (1989), "Speech Dynamics and Recurrent Neural Networks", *Proc. ICASSP-89*, Glasgow.

[4] WAIBEL, A., SAWAI, H., SHIKANO, K. (1989), "Consonant Recognition by Modular Construction of Large Phonemic Time Delay Neural Networks", *Proc. ICASSP-89*, Glasgow.

[5] GUYOT, F., ALEXANDRE, F., HATON, J.P. (1989), "Toward a Continous Model of the Cortical Column : Application to Speech Recognition", *Proc. ICASSP-89*, Glasgow. [6] GREEN, P.D. et al. (1987), "A Speech Recognition Strategy Based on Making Acoustic Evidence and Phonetic Knowledge Explicit", *Proc. European Conf. Speech Technology*, Edinburgh.

[7] MEMMI, D., ESKENAZI, M., MARIANI, J., NGUYEN-XUAN, A. (1983), "Un système expert pour la lecture de sonagrammes", *Speech Com.*, vol. 2, n° 2-3, pp. 234-236.

[8] CARBONELL, N., FOHR, D., HATON, J.P. (1987), "An Acoustic-phonetic Decoding Expert System", *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 1, n° 2, pp. 207-222.

[9] ZUE, V.W., COLE, R.A. (1979), "Experiments on Spectrogram Reading", *IEEE-ICASSP*, Washington, D.C., pp. 116-119.

[10] LAPRIE Y., HATON, J.P., PIERREL, J.M. (1990), "Phonetic Triplets in Knowledge-based Approach of Acoustic-phonetic Decoding", *Proc. Conf. Speech and Language Processing*, Kobé, Japan.

Acknowledgment :

The author gratefully acknowledges the contribution of members of the RFIA group to the APHODEX project : Anne Bonneau, Noëlle Carbonell, François Charpillet, Jean-Paul Damestoy, Mahieddine Djoudi, Dominique Fohr, Dominique François, Ramez Hajislam, Marie-Christine Haton, Yves Laprie, Jean-Marie Pierrel.

DYNAMIC VOICE SOURCE SYNTHESIS

Sarah K Palmer (1) and David M Howard (2)

(1) Phonetics and Linguistics Department, University College London, Wolfson House, 4 Stephenson Way, London NW1 2HE, UK.
(2) Signal Processing: Voice and Hearing Research Group; Electronics Department, York University, Heslington, York YO1 5DD, UK

ABSTRACT

Analysis of the excitation waveform modelled by a four parameter model of glottal flow has revealed consistent variations in shape due to laryngeal co-articulation. LPC resynthesis using the model demonstrates that changes to the shape of the excitation affect the quality of the speech in a way predicted by the findings of the analysis. Pilot studies suggest that dynamic excitation provides a more natural LPC resynthesis than non-dynamic excitation.

1. INTRODUCTION

Previous work using the JSRU parallel formant synthesiser [8] has failed to establish a preference for utterances synthesised with a dynamically varying excitation based on the three parameter model of glottal flow [2] over those synthesised using a static excitation waveform with pre-determined spectral slope.

A number of reasons were identified which could be contributing to this lack of preference. Firstly there are difficulties in obtaining formant amplitudes for the parallel formant synthesiser which correspond to the supra-glottal tract configuration alone. Secondly, previous work was based on a three parameter model of glottal flow [4] derived from the laryngographic waveform (Lx). More detailed models of glottal flow exist (eg: [3]). In addition the use of Lx (a measure of vocal fold contact) to derive a model of glottal flow has yet to be justified.

Thirdly, the test stimuli used as a basis for naturalness testing [8] were [ɑ:hɑ:] and [ɑ:ʔɑ:]. Whilst these stimuli demonstrate clear differences in the closed quotient trends due to laryngeal co-articulation, they rarely occur in natural British English speech and are not ideal for naturalness testing. The quality of the parallel formant resynthesis of these tokens was rather poor, and changing the voice source parameters gave no reliable perceptual judgements. Whilst this could be a function of our voice source model, we believe that subjects will become sensitive to voice quality differences *only* as the overall intelligibility of the synthesis improves. This view is supported by Pickering [9].

The aim of this work is to re-analyse the natural data using a four parameter model of glottal flow [3], to investigate the changes in the time course of the flow parameters, and to study the effects of altering these parameters based on LPC resynthesis via perceptual tests. The problems of formant amplitude estimation in parallel formant synthesis are therefore avoided.

2. METHOD

The fully automatic inverse filtering programs used in this work were developed by Chan and Brookes [1]. They carry out an LPC closed phase analysis from which the inverse filter is then calculated. This filter is then applied to the speech pressure waveform and a raw estimate of the differentiated glottal flow obtained. A four parameter Fant, Liljencrants and

Lin [3] model of the derivative of glottal flow (L-F) can then be fitted to the raw inverse filtered waveform. The LPC coefficients are then re-estimated using the modelled excitation waveform.

The output of the inverse filter was modelled automatically using the L-F model of glottal flow for the utterances [ɑ:hɑ:] and [ɑ:ʔɑ:] spoken by four male and four female speakers. Figure 1 shows a typical cycle of the L-F model and the time alligned equivalent cycle of glottal flow.

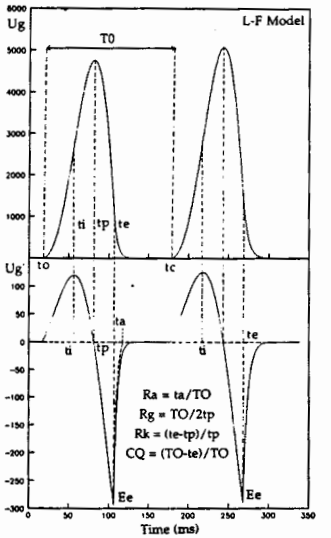


Figure 1: The four parameter L-F model of the derivative of glottal flow (Ug') with the derived glottal flow (Ug).

Analysis of the utterances was carried out in terms of five measurements: Rk (a measure of the asymmetry of the flow), Ra (the return phase ratio), Rg (the glottal frequency), CQ (the closed quotient) and Ee (the strength of the excitation). The way in which these ratios have been calculated is indicated in figure 1 along with the parameters from which they are derived: T0, tp, te and ta. Further details of the analysis ratios can be found in [5] and [7].

LPC resynthesis was carried out using the re-estimated filter function and various modelled excitation waveforms based on the five measurements above. A pilot test was carried out to compare the natural utterances of [ɑ:ʔɑ:] with three synthesised versions for three speakers, two male and one female. For each speaker the filter function remained constant whilst the excitation model was changed as follows:

(a) the shape of the excitation was varied dynamically on a cycle-by-cycle basis,
(b) the excitation was fixed according to the average values of Rk, Ra and Rg for the whole utterance,
(c) the excitation was fixed according to the average ratio values of Rk, Ra and Rg measured from the mid-portion of the second vowel in the utterance.

Perceptual testing was carried out to evaluate the naturalness of the stimuli. Four subjects were able to replay each stimulus through headphones as many times as required, and they were asked to mark down the stimulus which they perceived to be most like the natural utterance. For three of the tests subjects listened to the whole utterance whilst in a further three tests they only heard the initial vowel.

A further set of perceptual tests was carried out to study the effect of changes in the excitation on voice quality. Three stimuli were prepared based on the voice source analysis of one female speaker as follows:

(a) the excitation shape was fixed to the mean ratio values of the mid-portion of the second vowel in the utterance,
(b) Ra and Rk were increased and Rg decreased by 30% of their mean values to simulate a more breathy voice quality,
(c) Ra and Rk were reduced and Rg increased by 30% of the mean value to simulate a less breathy voice quality.

Five subjects were asked to rank the stimuli (presented over headphones in a sound-proof room) in terms of 'breathiness'.

3 RESULTS

3.1 Analysis

Results of the analysis confirm the previous finding, using Lx, that closed quotient increases before a glottal stop and decreases before a glottal fricative. This result is to be expected from an inverse filter analysis given a previously demonstrated high correlation between closed phase measurements made from Lx and inverse filtering [6].

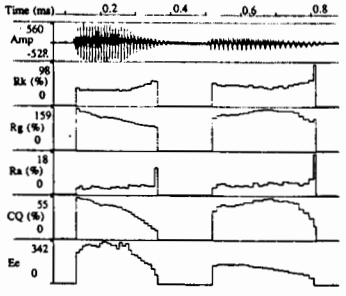


Figure 2. Analysis of [ɑ:hɑ:] showing the speech pressure waveform and changes in Rk, Rg, Ra, CQ and Ee over time.

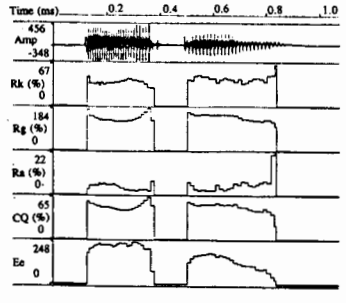


Figure 3: Analysis of [ɑ:ʔɑ:] showing the speech pressure waveform and changes in Rk, Rg, Ra, CQ and Ee over time.

Figures 2 and 3 show typical analysis findings for the utterances [ɑ:hɑ:] and [ɑ:ʔɑ:] spoken with falling intonation and the stress on the initial syllable for one male speaker.

3.2 Perceptual Tests

Results of the comparison between test stimuli and the natural utterance show that two of the subjects had a significant preference for utterances resynthesised using a dynamic excitation over those resynthesised with a fixed excitation shape (α=0.05). However, results from two further subjects were insignificant Overall there was no significant preference for the dynamic excitation. Focussing on the initial part of the utterance, where the main changes in the excitation are taking place, seems to have no effect on the preference results.

In the second test all the subjects chose stimulus (b), in which Ra and Rk had been increased and Rg decreased by 30%, as having the most breathy voice quality and stimulus (c) as having the least breathy quality.

4. DISCUSSION

The changes occuring in Rk, Ra and Rg have previously been linked to variations in the strength of the excitation which correlates strongly to the parameter Ee. It was hypothesised [5] that the sharpness of closure ratio Ra should vary inversely with Ee, (the stronger the excitation the more rapid the closure and therefore the shorter the return phase). When the glottal frequency Rg is fairly constant Rk varies with Ra in such a way that if the excitation strength Ee is large and Ra is small, the asymmetry of the pulse will increase due to a relatively shorter return phase (ie: Rk is small). On the other hand if the excitation strength is weak and Ra is high, the asymmetry will decrease and Rk will be high. Therefore at the onset and offset of voicing one would expect Rk and Ra to be higher than during the mid-point of a vowel in the utterance. This is confirmed by some of our data but the presence of a glottal stop seems to affect the relationship between these parameters.

Perhaps a clearer explanation is offered by studying the changes taking place in the closed quotient. When the closed quotient (CQ) is rising, as it does before a glottal stop (see figure 3), the length of the open phase becomes relatively shorter, resulting in a smaller value of 'tp'. This results in a rising Rg value since it is proportional to the inverse of 'tp', and the asymmetry of the pulse Rk decreases. The opposite effect is shown when CQ decreases in figure 2. CQ tends to be lower utterance initially and utterance finally and can therefore account for the changes taking place in Rk and Rg in these regions. The interpretation of the variation in the parameter Ra in our data is not clear, but it seems to depend mainly upon the the strength of the excitation Ee.

This work has demonstrated that voice source changes found in natural speech can be resynthesised and that modifications to the shape of the excitation waveform in LPC re-synthesis can alter perceived speaker quality appropriately, both for male and female speech. For some listeners dynamic excitation provides a closer perceived match to the natural speech than an excitation with a fixed waveshape. Whilst the number of subjects used in this test was limited and a more widespread study is needed, it is thought that the use of sentence level material instead of isolated utterances will produce a clearer preference for a time varying excitation due to different intrinsic properties of the phonemes within the utterance as well as co-articulation, stress and intonation effects. Our experience suggests that the overall quality of synthesis from the JSRU synthesiser requires improvement before excitation changes will affect the perceived naturalness significantly. Work is in progress to modify the methods used to specify the formant amplitude data used by the JSRU system.

5. ACKNOWLEDGEMENTS

This work was supported by the SERC research grant number GR/F/30642.

6. REFERENCES

[1] CHAN D.S.F.& BROOKES D.M, (1989), "Variability of Excitation Parameters Derived from Robust Closed Phase Glottal Inverse Filtering", *European Conference on Speech Communication and Technology*, Paris 199-202.

[2] FANT G, (1979), "Glottal Source and Excitation Analysis", *Speech Transmission Lab: Quarterly Progress and Status Report, 1*, Royal Inst. of Technology, Stockholm, 85-107 .

[3] FANT G., LILJENCRANTS J. and LIN Q, (1985), "A Four Parameter Model of Glottal Flow", *Speech Transmission Lab: Quarterly Progress and Status Report, 4*, Royal Inst. of Technology, Stockholm, 1-13.

[4] HOWARD D.M. and BREEN A. P, (1989), "Methods for Dynamic Excitation control in Parallel formant speech synthesis", *IEEE Proceedings of the International Conference on Acoustics, Speech and Signal Processing - 89, 1*, 215-218.

[5] GOBL C. and CHASAIDE A, (1990), "Linguistic and Paralinguistic Variation in the Voice Source", *Proceedings of the International Conference on Spoken Language Processing*, Japan, 85-88.

[6] HOWARD D.M, LINDSEY G. and ALLEN B, (1990), "Towards the Quantification of Vocal Efficiency", *Journal of Voice, 4*, 205-212.

[7] KARLSSON I, (1990), 'Voice Source Dynamics for Female Speakers', *Proceedings of the International Conference on Spoken Language Processing*, Japan, 69-72.

[8] PALMER S.K, ALLEN B, HOWARD D.M, LINDSEY G. and HOUSE J, (1990), 'Analysis, Synthesis and Perception of Laryngeal Co-articulation', *Proceedings of the Institute of Acoustics, 10*, 17-24.

[9] PICKERING J. B, (1989), 'Effects of Voice Type and Quality on the Intelligibility of a Text-to-Speech System', *European Conference on Speech Communication and Technology*, Paris, 637- 639.

A NOVEL MACHINE-LEARNING ALGORITHM FOR IMPROVING RECOGNITION PERFORMANCE IN A FEATURE-BASED, DELAYED-COMMITMENT CONTINUOUS SPEECH RECOGNITION SYSTEM

M. O'Kane, P. Kenne, D. Landy and S. Atkins

University of Canberra, Canberra, Australia

ABSTRACT

This paper describes an algorithm for machine learning in a feature-based continuous speech recogniser. This algorithm provides the recogniser with an automatic means of achieving high recognition scores when operating in speaker-independent mode. The learning algorithm operates at a lexical level because the fact the recogniser adheres to the principle of 'delayed commitment' in using speech signal information in making lexical decisions.

1. INTRODUCTION

The introduction by Kai-Fu Lee [1] of successful automatic learning and adaptation algorithms in Hidden Markov model speaker-independent continuous speech recognisers led to demonstrations of very significant improvements in the performance of recognisers using this architecture. To date no such similar learning algorithms have been devised for speaker-independent continuous speech recognisers using rule-based architectures. This paper presents a procedure which addresses the issue of learning techniques for rule-based recognisers.

2. THE RULE-BASED RECOGNISER

The rule-based recogniser used in this work was the FOPHO recogniser [2] which is a completely rule-based recogniser in which new lexical items and quasi-phonetic units alike are added to the system in the form of nested production rules. All rules are written in the special-purpose programming language, WAL (Wave Analysis Language) [3]. A simple rule in WAL involves associating a label with any time-segment of a speech waveform for which a specified signal processing function is true. For example the rule

```
feature
  name : stry,
  wave : speech,
  (association : zcross(20000,50)
      is
   long_high_amplitude(20,100))
end.
```

will give the label "stry" to any portion of a waveform for which the number of zero crossings averaged over a 50 msec time-window remains above 100 for 20 msec or more. Complex rules in WAL cause labels to be associated with time-logical combinations of simple signal processing functions or of other rules, simple or complex, specified either directly or by label reference. Available time-logical operators in WAL are : and, or, not, then, before, after, includes, strict_then, strict_before, strict_after. Rules can also have 'characteristics' (time_long, time_short, extendr, extendl) which can be thought of as extra conditions which rules have to meet in order to 'fire'. Thus the (complex) rule:


```
feature
  name : four,
  wave : speech,
  ((association : speech is f)
  strict_then
  (association : speech is or with
  characteristic time_long (60))
end.
```

means that in order for the lexical label "four" to be associated with a speech segment, the rule "f" must fire and then the rule "or" must fire for a time interval of at least 60 msec.

One of the fundamental principles of FOPHO is that all rules leading to quasi-phonetic labels should embody the Klatt principle of delayed commitment [3] in that they should fire on all exemplars of a given phonetic unit at the cost of overfiring on a certain percentage of near-confusion sounds to the target sound. Word hypotheses activated by incorrect fires are hopefully eliminated by the requirement that for a given lexical item to be recognised, the phonetic features have to occur in a certain sequence.

The expressive ease with which it is possible to write and test FOPHO rules using WAL, combined with the overfiring consequences of using the delayed commitment approach poses a problem for immediate introduction of formal automatic learning techniques in a system such as FOPHO. In order to proceed with the learning algorithm described here it was first necessary to automate the checking of the correctness or otherwise of rule fires against a labelled speech database and to re-structure the rule file on which automatic learning was to take place.

3. AUTOMATIC CHECKING

The Chk Program takes rule results from the WAL interpreter working on a selected speech database and compares the results with the hand-marked labels in the database. Correct fires, misses, misfires and overfires are all reported and analysed on an utterance-by-utterance basis (the verbose version) or as a summary for the whole database. If required the Chk Program also reports firing results in terms of rule components. For example, for the rule "four" given above, the program will, if the appropriate switch is set, indicate how many times and where both components fire and how many times only one or none of the components fires. This Chk Program forms the core of the learning algorithm.

4. RULE FILE RE-STRUCTURING

In order to make automatic learning easier to implement it has been necessary to re-structure the rule file before learning commences. Under the re-structuring, the "or" time-logical operator is replaced by separate rule statements. Also rule components at the simple rule level are tagged as 'capture' or 'eliminator' rules. The process of phonetic rule development is a process of 'capturing', if possible, all examples of a target phoneme while 'eliminating' as many phonemes as possible that are not in the target set. The tagging referred to above provides an explicit record of this process.

5. PRELIMINARY PROCESSING

To provide a baseline for learning, ordinary rule development for a particular vocabulary is carried out for a single speaker. Rule development for this speaker is typically done over at least twenty examples of each lexical item in spontaneous continuous speech. When the rules for the speaker are operating near-perfectly, they are used as the base set of rules for further work. This provides a suitable baseline for a speaker-

independent recogniser. For example, FOPHO rules were developed for one female speaker for the 'hard' vocabulary of the digits and the words "phone" and "double". When these rules were tested on 160 other speakers speaking a total of 250 utterances in spontaneous continuous speech with each utterance typically consisting of strings of fourteen to sixteen lexical items, the average percent correct over all lexical items was

$$\% \text{ correct} = \frac{\text{No. correct}}{(\text{No. correct} + \text{No. misses})} = 63.3$$

while the average misfire to correct ratio was

$$\frac{\text{No. misfires}}{\text{No. correct}} = 0.51$$

Of course the scores for particular lexical items in this vocabulary varied somewhat with the best case being for the word "six" for which the percent correct was 86% and the misfire to correct ratio was 0.60. The worst case was the word "double" where the percent correct was only 44.3% and the misfire to correct ratio was 1.0. By tuning the rules on extensive data for a second speaker, an improvement of 4% was achieved with no change in the misfire to correct ratio. It is on this single-speaker-developed, second-speaker-tuned recogniser that the learning algorithm was tested.

6. THE LEARNING ALGORITHM

The aim of the learning algorithm is to improve recognition scores at the lexical item level. Generally this will lead to improved scores at the quasi-phonetic level although, because of the principle of delayed commitment, this improvement will not necessarily follow.

As indicated in Section 4, the process of recognition rule development is a trade-off process of capturing as many examples as possible of a particular lexical item while simultaneously keeping misfires as low as possible. This trade-off notion is carried through into the learning algorithm.

In order to illustrate the operation of the algorithm we will trace its performance for the rule for the word "four" in the FOPHO digit-vocabulary system for which baseline data is given above. The baseline data for "four" in that system after second-speaker tuning is 67% correct and a misfire to correct ratio of 0.26.

The learning algorithm first considers the percent correct results for the large speaker set for the major (i.e. immediate sub-lexical) components of a lexical rule. In the case of the rule for "four" these components are the rule for "f" and the rule for "or" which have baseline percent correct scores of 91% and 71% respectively for occurrences within examples of "four". The algorithm uses this data to find the weakest component in a rule (in this case "or") and selects this for further processing.

In the next step the algorithm concentrates on eliminator rules for the selected component. If the Chk Program results shows that these eliminators are falsely eliminating any known "four" the learning algorithm proceeds to relax the rule parameters by a fixed percentage (currently 10%) in all the eliminator rules and keeps doing this until no examples are falsely eliminated. Results of this for "four" are given in Table 1 as 'after step 1' results. There is an 8% improvement in the percent correct results although the misfire to correct ratio does not change much because, while the number correct is improving, not as many cases as should be are being eliminated any longer.

In the next step, the algorithm concentrates on the capture rule components and proceeds to relax both rule and time parameters again by a fixed percentage (10% in this case) for all capture rules for this component. The results of doing this are the 'after step 2' results in Table 1. There has been a 12% inprovement in the percent correct score at the cost of almost doubling the misfire to correct ratio. Accordingly, the next step in the algorithm involves an attempt to lower this ratio. This is done by adding new eliminator rules designed to eliminate the major misfire lexical items (in this case "five", "phone" and "one"). These eliminator rules are found by table-lookup, thus this step is only as good as the table contents. The 'after step 3' results show that the table contents probably should be strengthened.

It is not surprising that the major misfire groups for this example are the words "five", "phone" and "one", as in Australian English the truncated start of the words "five" and "phone" do in many cases sound like continuous speech "four", as does the start of "one" when heard following a word ending in /f/. Issues such as these pose a problem in selecting a suitable stopping condition for the learning algorithm. Presently the learning algorithm arbitrarily stops after three steps. An obvious extension is to have it improve performance on the next weakest component. At present the issue of the not-too-surprising misfires is handled by lexical eliminator rules which for this example leads to the FOPHO rule "finalfour":

```
feature
  name : finalfour,
  wave : speech,
  ((((association : speech is four)
  and
  not (association : speech is five))
  and
  not (association : speech is
        phone))
  and
  not (association : speech is one))
end.
```


7. ACKNOWLEDGEMENTS

This work was supported by grants from the Industry, Research and Development Board, the Australian Research Council and Wang Australia.

8. REFERENCES

[1] LEE, K-F., *"Automatic Speech Recognition - The Development of the SPHINX System"*, Kluwer Academic Publishers, Boston, 1989.

[2] ATKINS, S., KENNE, P., LANDY, D., NULSEN, S and O'KANE, M., "WAL - A Speech Recognition Programming Language", *Proceedings of the ICSLP 90*, Kobe, Vol.1, pp233-236, November 1990.

[3] KLATT, D., "Speech perception : a model of acoustic-phonetic analysis and lexical access", *J Phonet, 7*, pp279-312, 1979

[4] O'KANE, M., "The FOPHO Speech Recognition Project", *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, Karlsruhe, pp630-632, 1983.

	% correct	misfire/correct ratio
Baseline data	67	.26
After step 1	75	.27
After step 2	87	.50
After step 3	87	.43

Table 1 : Learning algorithm results for "four".

AUTOMATIC LABELLING OF SPEECH SIGNAL INTO PHONETIC EVENTS

H. Kabré, G. Pérennou and N. Vigouroux

Institut de Recherche en Informatique
Toulouse, France

ABSTRACT

In this paper, we give the general principles behind an automatic system, developed at IRIT Lab, and capable of labelling speech signal for phonetic events. When using this sytem, the results secured on English, French and Swedish corpora demonstrate that the labelling operation becomes completely independent from either language, corpus or speaker. Moreover, this operation requires no manual adaptation or training whatsoever.

1. INTRODUCTION

Automatic labelling of speech corpora is an increasingly important problem, when considering present-day development of recorded speech databases —e.g., the DARPA Project ones.
In Europe, within the scope of the SAM ESPRIT Project —involving this kind of databases for multilingual corpora— the question has quickly arisen both as to how to adapt these various automatic labellers to different languages, and as to how to process speech material without having to resort either to a manual adaptation or to some kind of language, speaker or corpus training.
The latter problem is the one considered, here, as we are presenting SAPHO —the phonetic front-end of our automatic labeller.

2. SYSTEM STRUCTURE

Our labelling system procedes in two successive main stages:
- in the course of the first one, the signal is both segmented and labelled into phonetic events (SAPHO component);
- in the second stage, these events become aligned onto a phonetic transcription, supplied beforehand by a phonetician (VERIPHONE component).

A more detailed description of this can be found in [2], as only the general principles triggering decision are reported here.
1) The automatic alignment, arrived at, does not require any fine characterization of the phonetic events involved; macro-class labels being quite sufficient to do the job (As an extreme example, if the sequence [tom] were to be aligned onto the five-event string [+occl] [+fri] [+vow] [+voc] [+occ], the phoneme [t] would easely be identified as made up of the first two elements of the string, while [o] would readely be identified as the third element and [m] as the last two elements).
2) Under such conditions, the whole set of permanent acoustic parameters, necessary for recognition, does not have to be used up.
3) An appropriate, limited choice, among these, will bear upon the minimum subset of the most robust parameters, given the target set for the exercise; i.e., labelling that is independent from either language, speaker or corpus.
In [4], various parameters can be found, which were used in automatic labelling. Some of them describe signal amplitude, others spectrum.
We chose two parameters that are universally used by manual labellers; namely, the amplitude parameter and the zero crossing rate —both of wich contain enough information to accomplish the contemplated task.
Both afford the advantage of an identical performance over whole sets of languages, speakers and corpora.

This work is supported by the SAM ESPRIT Project and the GDR-PRC Communication Homme-Machine Program.

These parameters have to be pre-processed in order to achieve an optimal automatic segmentation.
In Fig.1, for example, three parameter forms can be seen to characterize amplitude.
In 1c, maximun amplitude is evaluated over each one of the 4ms successive frames; this amplitude has undergone a non-linear smoothing (NLSA parameter) that does preserve major instances of signal discontinuity.
In 1d, mean amplitude (energy) over each 8ms frame is evaluated.
These two amplitude values can be directly compared to the initial waveform given in 1b.
Our choice went to the logarithmic NLSA amplitude —normalized LNLSA.
This normalization occurs at two different levels:
- with respect to the whole corpus, in order for the amplitude, thus normalized, to vary within the [0, 1] interval,
- with respect to a local signal interval, by taking the ratio of amplitude to maximum amplitude within a ± 0.25 ms window that is centered upon the instant considered (whenever this maximum amplitude falls below a floor value, the latter is taken, instead, as the denominator of the ratio).

This representation of amplitude is advantageous in at least three ways :
- it preserves essential contrasts between successive phonemes (the NLSA parameter can be compared to the mean amplitude one; allowing to observe that, with the latter, the contrast "closed/nasal consonant" is all but lost, whereas it comes out enhanced with NLSA);
- amplitude comes out smooth, while essential instances of discontinuity are preserved;
- amplitude is strongly correlated to phoneme aperture; the effects of the mean sound intensity variation, within the phrase, being attenuated by local normalization.

Similar treatments are applied to the zero crossing rate given in 1e; although, in this case, normalization is global over the whole corpus.
The parameter thus obtained is LNLSZ.
Both parameters, LNLSA amplitude and LNLSZ zero crossing rate, constitute the starting basis for an evaluation of cues, enabling to label each 4ms within one of the categories appearing in the table on Fig.2.

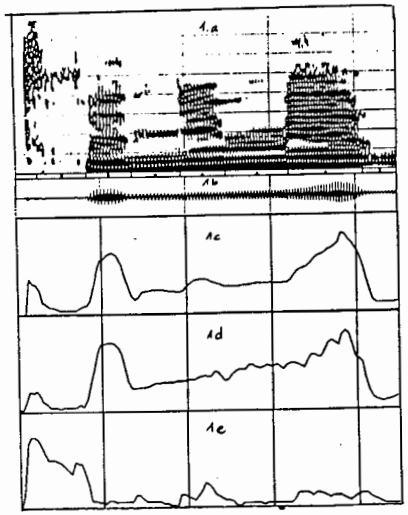


Fig.1 - Temporal parameters : 1a) Spectrogram, 1b) waveform, 1c) NLSA normalized amplitude, 1d) 8ms mean amplitude, 1e) NLSZ zero crossing rate.

Events are obtained, next, when regrouping frames by labels of equal values, while smoothing off any segment that is too short.

K	strong syll	S	s like fric
W	weak syll	C	short S
L	acute voc	Z	z like fric
U	grave voc	F	weak fric
O	voiced occl	X	x like fric
Q	unvoiced occl		

Fig.2 - Table 1 : Phonetic events.

3. EVALUATION

The quality of various events, obtained, can be evaluated thanks to the different kinds of results.
The ones given in Table 2, Fig.3, show quality of automatic segmentation, as this compares to manual labelling, when the latter is obtained over EUROM-0 French, English and Swedish corpora, without any kind of either manual adaptation(-phase) or learning.

These results remain very steady from one language to the next. Furthermore, boundary accuracy is of the same order as what is observed when comparing between 'handlabellers' performances.

Pass	Language	Phoneme Number	Speaker Number	Surseg. Ratio	QC(*) ±13 ms	QC ±17 ms	QC ±21 ms	QC ±25 ms
1	English	4476	4	2.10	0.850	0.890	0.914	0.933
2				1.44	0.794	0.843	0.872	0.888
3				1.27	0.763	0.815	0.846	0.865
1	French	2909	2	2.74	0.866	0.918	0.938	0.950
2				1.78	0.823	0.894	0.919	0.933
3				1.53	0.793	0.870	0.898	0.913
1	Swedish	1379	1	2.15	0.782	0.833	0.861	0.884
2				1.58	0.742	0.793	0.826	0.846
3				1.35	0.700	0.762	0.803	0.822

(*)QC = [n / N] , N = number of manual boundaries, n = those ones that have an approximation to an automatic boundary less than ±x ms (x=13, 17, 21, 25).

Fig.3 - Table 2 : Quality of the SAPHO segmentation.

The other results also display great steadiness, both over various corpora and from one language to the next.
The results, presented in this paper, show that segments, provided by a handlabeller in order to account for a realization of phonetic units showing up in a transcription, generally are compounds that can otherwise be broken down into a set of a few phonetic elements made available by the SAPHO automatic process.
Modelizing a given phonetic unit, belonging to a given language, boils down, therefore, to specifying the stochastic laws which pertain to it and which steer a combination of events leading up to a realization of these units. In the way of phonetic units, it is of course much better to choose contextual allophones, for a more homogeneous spread of the various realizations.
In addition to this process —which is likely to occur in every language— there are properties —also common to all languages— such as the presence of events that are specific to natural classes of phonetic units.
This is illustrated in the table on Fig.3, where stops can be seen generally to entail an event Q. It is clear, however, that in this respect languages differ from each other through their respective phonological systems, and that the stochastic laws pertaining to various phonetic units must be specifically estimated for each such unit.

ENGLISH

ph	K	W	U	L	O	Q	Z	F	X	C	S	ssr
p	0	0	0	1	2	0	6	6	6	0	0	1.2
t	0	0	0	0	0	39	35	12	8	6	40	2
k	0	0	0	5	0	6	14	19	13	0	6	1.6

FRENCH

ph	K	W	U	L	O	Q	Z	F	X	C	S	ssr
p	0	0	0	10	8	96	0	14	7	0	1	1.4
t	0	1	11	11	3	90	8	12	3	5	12	1.6
k	0	2	18	6	4	104	24	20	14	0	8	2

Fig.4 - Average number of phonetic events in [p] [t] [k] phonetic units and sursegmentation rate (ssr) for English and French.

Thus, tables on Fig.4 shows that, in English, [t] becomes realized often (probability in the order of 40%) as Q+S. This combination does occur in French, as well, but with a lesser frequency (ca. 12 % prob.). Conversely, the combination of Q with a vocalic segment (W, L or U) seldom occurs in English, whereas it is frequent in French (ca. 23 %).

The results, presented in this paper, show that segments, provided by a handlabeller in order to account for a realization of phonetic units showing up in a transcription, generally are compounds that can otherwise be broken down into a set of a few phonetic elements made available by the SAPHO automatic process.
Modelizing a given phonetic unit, belonging to a given language, boils down therefore to specifying the stochastic laws which pertain to it and which steer a combination of events leading up to a realization of these units.
In addition to this process —which is likely to occur in every language— there are properties —also common to all languages— such as the presence of events that are specific to natural classes of phonetic units. This is illustrated in the table on Fig.4, where stops can be seen generally to entail an event Q.

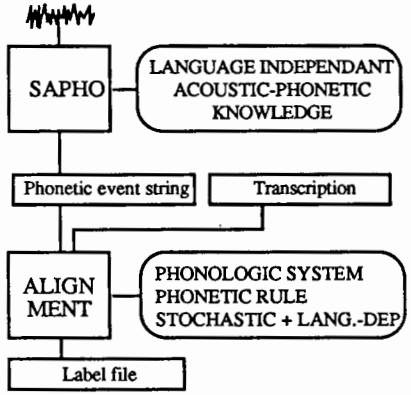


Fig.5 - The two levels of automatic labelling; language modelling implications.

It is clear (Fig.5), however, that in this respect languages differ from each other through their respective phonological systems, and that the chance laws pertaining to various phonetic units must be specifically estimated for each such unit. Thus, the table on Fig.4 shows that, in English, the phoneme [t] becomes realized quite often (probability in the order of 40%) as Q+S. This combination does occur in French, as well, but with a lesser frequency (ca. 12 % prob.). Conversely, the combination of Q with a vocalic segment (W, L or U) seldom occurs in English, whereas it is frequent in French (ca. 23 %).

4. CONCLUSION

The results we have secured over English, French and Swedish speech corpora, demonstrate the feasibility of labelling phonetic events that are language-, speaker-, as well as corpus-independent. However, these results should be reinforced both over larger corpora and over a more numerous set of languages. The results, presented here, were secured with the SAPHO System, which makes use of information relating only to amplitude and zero crossing rate.
We are now working at an efficient use of these events in automatic alignment and on pre-selection of sub-vocabularies within large lexicons.

The authors are thankful to Prof. J. F. Malet, CSU Sacramento, for this prompt translation of their original French ms.

5. REFERENCES

[1] C.J.M. Hoeckel, "The Reliability of Manual Labelling of Continuous Speech", Proceeding of ESCA Tutorial Day and Workshop on Speech Input/ Output Assessment and Speech Databases, paper 5.5.1-5.5.4.
[2] G. Pérennou, M. de Calmès, J.M. Pécatte, N. Vigouroux, "Phonetic-String Alignment for an Automatic Labelling of Speech Corpora", in Proceedings of Workshop on Speech Input/Output Assessment and Speech Databases, The Netherlands, 20-23 September, pp. 5.4.1-5.4.4.
[3] S. Seneff, V. Zue "Transcription and Alignment of the TIMIT Database," in Getting Started With The DARPA TIMIT CD-ROM.
[4] ESPRIT Project 2589, Intermediate Report, 1st March 1989 - 28 February 1990.
[5] V.W. Zue, R.M. Schwartz, "Acoustic Processing and Phonetic Analysis", in Trends in Speech Recognition W.A. Lea (ed.), pp. 101-124.l.

UN SYSTEME D'ACQUISITION D'IMAGES SIMULTANEES POUR L'ETUDE DES MOUVEMENTS DES ORGANES ARTICULATEURS.

Bernard TESTON

URA 261 CNRS, Institut de Phonétique, Aix en Provence, France.

ABSTRACT

This experimental device, is particularly valuable for obtaining a perfect correlation between articulatory and acoustic events. It allows the simultaneous recordings:

- Labio videofilms: frontal and side view.
- Radioscopic videofilms of the complete vocal tract.
- Endoscopic videofilms of velum, pharynx, larynx, oral cavity etc...

Speech signals and images are synchronized. The various images are recorded simultanously on video recorders at 50 frames per second.

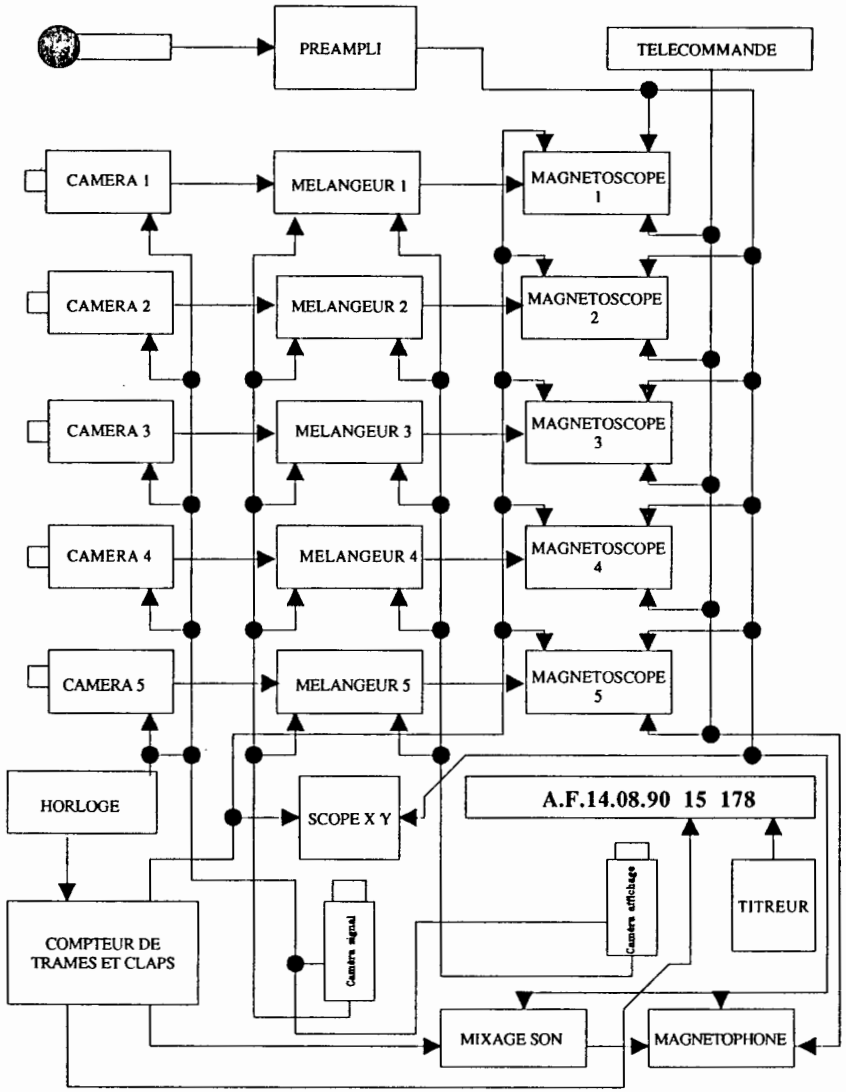
1. INTRODUCTION

Les études sur les processus d'encodage de la parole qui permettent de passer du niveau linguistique au niveau phonétique, revêtent depuis plusieurs années un intérêt croissant,et en particulier celles qui portent sur les mécanismes articulatoires. Les phénomènes de coproduction et de coarticulation, ainsi que la variabilité intra et inter locuteurs, sont les deux pôles de cet intérêt avec comme objectif, l'amélioration des systèmes de synthèse et reconnaissance vocale.

Les dispositifs d'exploration des mouvements des organes articulateurs, sont nombreux et de principes variés. Une des méthodes les plus directes pour mettre en évidence leurs mouvements, est de décomposer leur cinématique image par image, selon le concept «du récit chronophotographique avec arrêt sur l'image», imaginé par MAREY [1]. En 1986, nous avions réalisé un dispositif d'acquisitions vidéo simultanées sur des images du voile, des lèvres et de la coupe sagittale du conduit vocal [2]. Ce dispositif qui nous donnait satisfaction, a dû évoluer pour les raisons suivantes. Tout d'abord, ce système etait constitué par du materiel vidéo standard, et en particulier par des consoles de mixage utilisées au minimum de leurs possibilités et dont nous ne disposions pas en permanence. Nous devions l'emprunter ou le louer auprès de différents fournisseurs, il était de ce fait hétéroclite et onéreux. Enfin, le dispositif de synchronisation manquait de précision , même s'il nous a permis de ne pas perdre d'images. La nécessité de réaliser de multiples enregistrements sur un grand nombre de locuteurs, dans le cadre d'un contrat de recherche avec la Région Provence- Alpes -Cote d'Azur (PACA), nous a amener à développer le système suivant: Il est possible d'utiliser un maximum de cinq caméras vidéo. Toutes ces caméras sont synchronisées par une horloge générale. Un dispositif particulier permet l'incrustation sur chaque image du signal acoustique et des informations de synchronisation. le système est complété par une télecommande générale de toutes les fonctions d'acquisition, d'un dispositif d'affichage du corpus à prononcer, et d'un appareillage mécanique de positionnement du sujet et des différents capteurs d'information.

2. LA PRISE D'IMAGES

Il est possible d'utiliser des caméras de toutes provenances, à condition qu'elles disposent d'une entrée de synchronisation (genlock). Nous n'utilisons cependant, que des caméras à dispositif de transfert de charges (CCD), particulièrement bien adaptées à la décomposition du mouvement. En effet, les CCD ont, parmi d'autres avantages, une rémanence beaucoup plus faible que les vidicons[3] ce qui permet de ne pas éblouir longtemps le capteur (avantage très important en endoscopie), et surtout la possibilité d'utiliser un obturateur (shutter), qui permet de définir plus précisément l'image dans le temps, pour l'étude des mouvements rapides (lèvres) [4] ; Pour l'étude des mouvements des lèvres de face et de profil, nous utilisons des caméras PANASONIC WVF 250 tri CCD, équipées d' un objectif S12X75 BRM avec un temps d'exposition de 1 ms qui nécessite un éclairage puissant. La caméra pour l'étude des mouvements du voile est une PANASONIC M10 couleur mono CCD, caractérisée par un très faible éclairement minimum (10 lux) qui permet de pallier le manque chronique de lumière des fibroscopes. Ce dernier est un bronchoscope OLYMPUS BF20D, choisi pour sa bonne définition d'image (bien supérieure à celle des nasolaryngoscopes) et son double canal d'éclairement. Il est associé à une source de



SCHEMA DE PRINCIPE DU DISPOSITIF

lumière au xénon OLYMPUS CLV10 utilisée en mode continu (non pulsé). Il est possible d'utiliser des caméras tri CCD pour cette application mais elles sont beaucoup plus encombrantes et plus lourdes. Enfin, la caméra monochrome de vidéoradioscopie est caractérisée par une grande définition (700 points par lignes). Elle fait partie d'une installation d'angiographie GENERAL ELECTRIC (CGR) utilisée pour des interventions de micro-chirurgie intra- vasculaire de longue durée. Elle dispose d'une grande luminosité grace à un amplificateur de brillance très sensible qui permet de faibles doses de radiations X. Son champ a un diamètre de 34 cm, on peut ainsi y cadrer le conduit vocal dans sa totalité.

3. L'HORLOGE DE SYNCHRONISATION

Toutes les caméras et les dispositifs de synchronisation sont cadencés par une horloge générale. Sa fréquence est générée soit par une référence interne à quartz, soit à partir du signal vidéo d'une caméra. Cette possibilité est obligatoire dans le cas de l'utilisation de la vidéoradioscopie dont l'intallation ne dispose pas d'une entrée de synchronisation exterieure (genlock). Compte tenu de sa spécificité (12 sorties de synchronisation) cette horloge a été fabriquée localement.

4. LE SYSTEME DE SYNCHRONISATION

C'est le dispositif central de toute l'expérience. Nous l'avons réalisé spécialement pour faire des enregistrements vidéo parfaitement identifiables en toutes circonstances. Pour cela nous faisons apparaître sur chaque image:

- Le numéro de la phrase dans le corpus (clap).
- Le numéro de l'image (trame paire ou impaire).
- Le signal acoustique correspondant à la durée exacte de la trame de télévision (20 ms) de haut en bas sur la partie gauche de l'écran.
- Les références du locuteur.
- La date de l'enregistrement.

Les phrases sont considérées comme des séquences cinématographiques comprises entre deux claps. Le système de synchronisation peut compter jusqu'a 99 claps. Chaque clap remet à zéro le compteur d'images. C'est le locuteur qui déclenche le clap, qui fait apparaitre la phrase à prononcer sur l'écran d'un PC. Il peut également déclencher l'acquisition de signaux issus de différents capteurs (débits, pressions, déplacements, etc...) sur un autre PC disposant du logiciel PHYSIOLOGIA [5]. Le compteur d'images est synchronisé par l'horloge générale. Chaque image est constituée par une trame paire ou impaire du signal vidéo, sa durée est de 20 ms pour le standart européen. Il est possible d'utiliser ce dispositif à la fréquence de 60 trames par seconde si cela s'avère nécessaire. Le système peut compter jusqu'à 999 images entre deux claps.Les impulsions d'images correspondent au retour de trame (synchronisation verticale du signal vidéo). Elles synchronisent le défilement d'un scope X Y sur lequel est visualisé le signal acoustique. Ce dernier est filmé au moyen d'une caméra monochrome PANASONIC WBL200 équipée d'un objectif LA12B2 synchronisée par l'horloge générale. Les impulsions de claps et d'images sont enregistrées sur une des deux pistes son des magnétoscopes. Elles sont reconnaissables par leur différence d'amplitude. Elles peuvent moduler une fréquence de 2 kHz pour les claps et de 7 kHz pour les images pour réaliser des marqueurs sur des analyses en fréquence de type sonagramme [6]. Enfin, les impulsions de claps peuvent être mixées au signal acoustique pour parfaire la synchronisation image-son. Les informations d'identification des locuteurs et la date de l'enregistrement, sont affichées sur un indicateur alphanumérique électroluminescent, ainsi que les numéros des phrases et des images. Cet indicateur est filmé par une seconde caméra monochrome dans les mêmes conditions que précédemment. Toutes ces informations d'identification et de synchronisation sont ensuite incrustées dans chaque image anatomique au moyen d'un mélangeur vidéo.

5. LE MELANGEUR VIDEO

Nous mélangeons à chaque signal vidéo couleur PAL issu des caméras de prises de vues anatomiques, des signaux vidéo monochromes issus des caméras de synchronisation. Tous ces signaux étant en phase au moyen de l'horloge générale, nous pouvons nous contenter de simples mélangeurs sans dégrader les couleurs originales. Nous évitons ainsi d'utiliser des régies de mixage vidéo complexes et couteuses. Le mélangeur vidéo est constitué par 5 canaux à 3 entrées. Une entrée commune à tous les canaux, contient les informations sur l'enregistrement et les numéros de claps et d'images. Une entrée, sur chaque canal, est réservée à l'image du signal acoustique, qui peut être supprimée pour augmenter la dimension utile de l'écran. La dernière entrée est réservée aux images anatomiques. Les réglages de mélange ne jouent que sur la luminance de chaque entrée, elle peut être ajustée sur une échelle de + ou - 20 dB.

6. LA PRISE DE SON ET SA SYNCHRONISATION

La prise de son est effectuée au moyen d'un microphone à électret AKG C410, associé à un préamplificateur mélangeur. Le signal de parole est enregistré sur une des deux pistes des magnétoscopes; L'impulsion des claps de début des phrases est superposée au signal de parole (pour avoir un bon repérage temporel de ce dernier s'il doit être acquis sur ordinateur en vue d'un traitement acoustique). Le signal de parole peut être également mixé avec les impulsions d'image comme nous l'avons précisé précédemment, il est alors enregistré sur un magnétophone bipiste avec le signal original. L'ensemble de l'expérience a un fonctionnement silencieux ,excepté la source de lumière OLYMPUS que l'on a du insonoriser pour avoir une bonne dynamique de la prise de son.

7. L'ENREGISTREMENT DES IMAGES

Les images anatomiques contenant les informations d'identification et de synchronisation sont enregistrées sur des magnétoscopes au standart U-MATIC SONY BVU 900 avec le signal acoustique sur la piste 1 et les impulsions de claps et d'images sur la piste 2. Si l'on désire une image synthétique de quatre images anatomiques, il est possible d'en réaliser le montage au moyen d'une régie de découpage FORT A MV 24 GL par exemple et de l'enregistrer sur un cinquième magnétoscope. Un dispositif de télécommande générale, permet de déclencher simultanément toutes les fonctions des magnétoscopes et magnétophones.

8. LES EQUIPEMENTS ANNEXES ET LES PROCEDURES DE CALIBRATION

Tous ces équipements sont solidaires d'un ensemble mécanique dont le centre est occupé par le fauteuil sur lequel prend place le locuteur. La tête de ce dernier est maintenue par un céphalostat efficace et peu contraignant. Le fauteuil est contenu dans un parallélépipède dont on peut très précisément ajuster le positionnement. Il peut supporter les caméras, l'endoscope, les dispositifs de calibration, les filtres pour rayons X, le microphone d'enregistrement ainsi que divers capteurs. Cet ensemble mécanique est adapté aux dimensions des installations d'angiographie, il est démontable pour être facilement transporté.

Les calibrations sont effectuées de deux manières différentes soit en positionnant des repères réalisés par maquillage, soit à partir de références dimentionnelles mécaniques. La première procédure s'applique aux images des lèvres de face et de profil. Deux videofilms sont réalisés simultanément, l'un en gros plan sur les lèvres, l'autre de toute la face. Dans les deux cas, les lèvres sont maquillées en bleu afin d'assurer une meilleure précision à la mesure d'aperture [7]. Des repères de calibration peuvent être peints sur le visage du locuteur. La méthode qui consiste à placer des repères dans le conduit vocal pour l'étude de sa coupe sagittale s'apparente à cette procédure. Pour cela, des billes de nickel-chrome sont placées par collage sur la langue pour repérer plus facilement ses mouvements [8]. La deuxième procédure consiste à placer une règle de référence dans le plan des lèvres de face et de profil et à la filmer avant chaque séquence de prise d'images.

CONCLUSION

Ce dispositif d'étude des mouvements des organes articulateurs, est l'aboutissement d'une lente évolution, et sa mise au point, le résultat de nombreuses manipulations. De telles expériences, qui nécessitent le contrôle de très nombreux paramètres ne sont jamais simples à mettre en oeuvre. Nous commençons à bien les maîtriser. Les principes fondamentaux de ce système sont maintenant bien établis, et les améliorations actuelles ne portent que sur des points de détail. Il nous reste à dépouiller et à traiter l'énorme quantité d'informations ainsi récoltée. Ceci ne peut se faire qu'au moyen de dispositifs automatiques sur lesquels nous travaillons et qui feront l'objet d'un prochain exposé.

REMERCIEMENTS:

Nous remercions M. CHERIGUENNE, technicien au CNRS, pour l'aide constante qu'il nous a apportée dans la réalisation des matériels.

BIBLIOGRAPHIE

[1]MAREY, E.J, (1891), "La chronophotographie." Revue générale des Sciences pures et appliquées, vol 16/11 1891, 659-719.

[2] TESTON, B. et AUTESSERRE, D. (1986),"*Description d'un dispositif d'enregistrement simultané des mouvements des organes articulateurs.*",Actes des 16° J.E.P. de la Société Française d'Acoustique, AIX, Mai 1986, 65-68.

[3] HOWELLS, J.W. (1986),"*Most of what you wanted to know about charge coupled devices.*",Tutorial presented to ITVA, Toronto, Nov. 1986, PANASONIC Doc., 13 p.

[4] MIQUEL, J.C. (1985),"*L'observation en vidéo rapide.*",Chap. 3, Lavoisier, Paris. 133-212.

[5] TESTON, B. et GALINDO, B. (1990),"*Design and development of a work station for speech production analysis.*",VERBA 90, ALCATEL FACE, Jan 90, ROME, 400-408.

[6] SIMON, P., BOTHOREL, A., WIOLAND, F., et BROCK, G. (1979),"*Méthode de synchronisation image-son pour l'étude radiologique des faits de parole.*",Actes du 9° Congrés International des Sciences phonétiques, Copenhague, Vol 1, 213.

[7] LALLOUACHE, M.T. (1990),"*Un poste «visage-parole»; Acquisition et traitement de contours labiaux.*",Actes des 18° J.E.P. de la Société Française d'Acoustique, Montréal, Mai 1990, 282-285.

[8] KIRITANI, S., ITOH, K., and FUJIMURA, O. (1975),"*Tongue pellett tracking by a computer controlled X-ray microbeam system.*",J.A.S.A., vol 57, 1516-1520.

SESSIONS ORALES / ORAL SESSIONS

SESSION 1 : Production

SESSION 2 : Perception I

SESSION 3 : Perception II

SESSION 4 : Phonologie / Phonology

SESSION 5 : Linguistic aspects
Acquisition; changements linguistiques / linguistic change

SESSION 6 : Intonation I

SESSION 7 : Intonation II

SESSION 8 : Applications
Pathologie / Pathology

SESSION 9 : Applications
Didactique / Language teaching

SESSIONS AFFICHEES / POSTER SESSIONS

SESSION 10 : Applications

SESSION 11 : Acoustique / Acoustics

SESSION 12 : Technologie / Technology