SWEDISH VOICES IN MUSIC

Johan Sundberg
Department of Speech Communication & Music Acoustics, KTH,
Box 70014, S-100 44 Stockholm, Sweden

ABSTRACT

Two exotic forms of singing in Swedish folklore are presented, kölning and jojk. Data on tuning of scale tones, formant frequencies, and articulatory characteristics are presented for two representative examples of each of these singing styles. The substantial differences from Western operatic singing are discussed.

INTRODUCTION

The voice is an important music instrument in Sweden. Sweden has produced a number of great international singers in the operatic tradition, e. g., Jussi Björling, Birgit Nilsson, Ann Sofi von Otter, Gösta Winbergh and Håkan Hagegård. This style of singing has generated a considerable amount of research during the last decades.

Apart from this, choral singing is exceptionally common in Sweden. Of the Swedish population a total of about 5 to 10% is or has been a choir singer. My colleague Sten Ternström has analyzed acoustical aspects of choral singing extensively and has published the result in a great number of articles (for a review, see [1]).

The acoustic voice characteristics in speech and singing differs considerably. In the case of classical operatic singing the main reason for these differences is reasonably well understood; the need for being heard over a loud orchestral accompaniment without straining the voice.

Also in the folkloristic subcultures in Sweden the voice is commonly used as a music instrument. It is used not only in folk songs, but also without a text, with vowel sequences or nonsense syllables, more like an instrument. In some Swedish subcultures very peculiar styles of singing have evolved. Two examples will be presented here. Both offer interesting examples of differences between singing and speech.

Kölning

My first example is an exotic kind of herding song practiced by the maids in the province Dalecarlia during the summer, when the cattle was brought up to the woods in the mountains to graze. The type of singing is called kölning, a derivative of kalla (call) and has been described elsewhere [2]. The extramusical function of kölning was mainly to collect the cattle in the evening, but kölning was also used by the maids to communicate with their colleagues on other mountains.

The structure of these songs is a sequence of short melodic patterns, some of which are repeated. Kölning is typically performed on sustained vowels without interleaved consonants. The acoustic and articulatory characteristics of this type of voice use were investigated some time ago [3]. Next, typical results from these investigations will be reviewed and compared to Western operatic singing as studied in another, similar investigation [4].

The subject was a woman, born 1909, who had learned this singing style from an unbroken oral tradition. She was also a choir singer, so her kölning technique

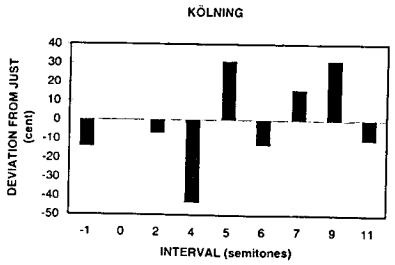


Figure 1. Mean F0 values, determined by means of histograms, for the scale tones in kölning as deviations from just tuning. The unit is cent, i. e., hundredth of a semitone.

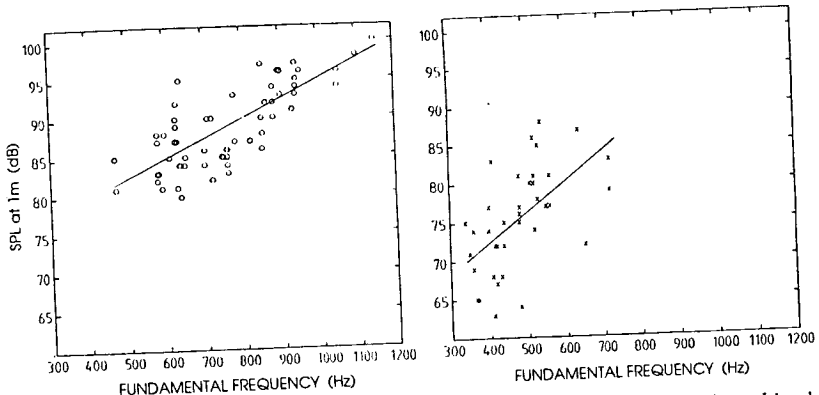


Figure 2. SPL values measured in an anechoic room at 1 m distance in the subject's kölning and singing (left and right panels). The line shows the best linear fit.

could be compared to a more regular type of singing.

The mean F0 of the scale tones she used in one recording of a kölning was analyzed by means of F0 histograms. In Figure 1, the values are shown in terms of deviations from just tuning, the unit being cents, i. e., hundredths of a semitone (st). Just tuning is obtained by multiplying the frequency of the center tone of the piece by ratios between small integers, such as 3:2, 4:3, 4:5, etc. It is used in Western traditional music along with other, rather similar tunings such as the equally tempered tuning.

In the figure, the values would all stay on the 0 deviation line, if kölning adhered to just tuning. Keeping in mind that a deviation of 100 corresponds to a semitone, some deviations from just tuning are considerable. However, the minor and major seconds, the augmented fourth, the fifth and the major seventh (1, 2, 6, 7, and 11 st) are only about 10 cents away from just. The major third is very flat indeed, nearly halfway down to the minor third. These deviations from just tuning are similar, though far from identical to those previously found in an analysis of a different example of kölning performed by another subject representing the same tradition [5]. We may conclude that kölning has developed a special kind of scale which systematically deviates from just intonation. In particular, the third is neutral, halfway between minor and major, and the fourth is sharp. Thus, in this respect kölning offers a striking

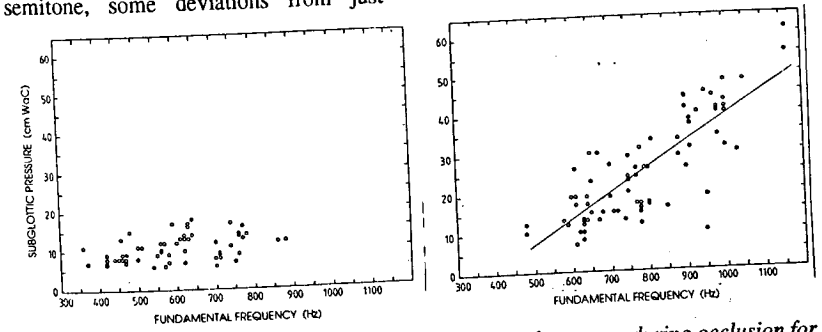


Figure 3. Mean subglottal pressures captured as the oral pressure during occlusion for [p] in the subject's kölning and singing (left and right panels). The line shows the best linear fit.

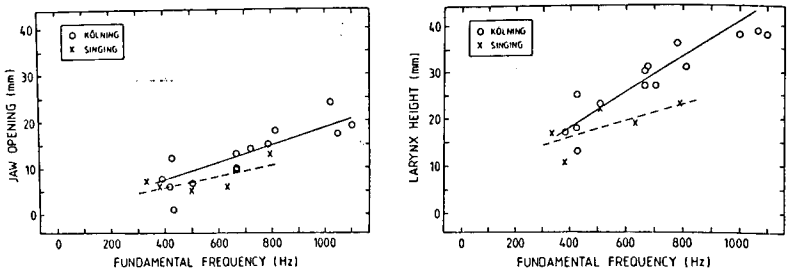


Figure 4. Jaw opening relative to clinched jaws and vertical position of the larynx relative to the resting position in the subject's kölning and singing (left and right panels).

example of the restricted applicability of the tunings used in traditional Western music.

The sound level in kölning was measured in an anechoic room. It was quite high and dependent on F0, as shown in Figure 2. In normal singing at F0=500 Hz the subject produced 76 dB @ 1 m distance, on the average, and the mean increase was about 16 dB/octave. In kölning, the sound level at 500 Hz was about 13 dB higher and the increase with F0 was only 7 dB/octave.

As expected, these high SPL values were paid in terms of high subglottal pressures as illustrated in Figure 3. The mean pressure at F0=550 Hz was 10 cm H_2O and increased linearly with fundamental frequency by about 7 cm H_2O/100 Hz. This clearly exceeds the maximum 20 cm H_2O which the subject used when she sang in a more traditional style.

Articulatory characteristics were studied from tracings of lateral X-ray images of the vocal tract taken at different moments during kölning. Thus, a material of 14 radiographs were collected from a representative choice of pitches within the relevant range. For comparison 5 radiographs were taken when the subject sang a melismatic part of a folk song, using different vowels. The pitch associated with each image was measured from an audio recording. In addition, the kölning data can be compared with corresponding data from a professional soprano singer, the technique of whom was studied in a different investigation [4]. This subject sang the vowels [a:, i:, u:] at three different pitches, F0=240, 480, 960 Hz.

There were clear articulatory differences with regard to e. g. tongue shape, jaw opening, and larynx position. The jaw opening and the vertical position of the larynx tended to increase linearly with F0, particularly in kölning, but also in singing, as illustrated in Figure 4. The larynx was above resting position in both singing styles, and the rise with F0 was quite substantial in kölning, about 23 mm between F0=400 and F0=1000 Hz. In a corresponding study of a professional soprano the larynx was also found to rise with pitch, but the larynx was constantly below the resting level, touching it only for the top F0. During both singing and kölning the subject's distance between the upper and lower lip showed a linear increase with jaw opening and the retraction of the mouth corners was linearly related to this distance. In kölning the tongue shape varied with F0 so that similar shapes were found for vowels produced at similar F0, particularly in the upper part of the pitch range, see Figure 5a. The tongue was pharyngeal with a frontal position of the tongue tip. With rising pitch the tongue root and the dorsum were substantially raised because of the increased jaw opening and the raised larynx. In operatic singing, by contrast, the tongue shape was different for different vowels except in the top part of the range, as can be seen in Figure 5b.

The summed effect of all these articulatory changes with pitch on vocal tract length was substantial, as can be

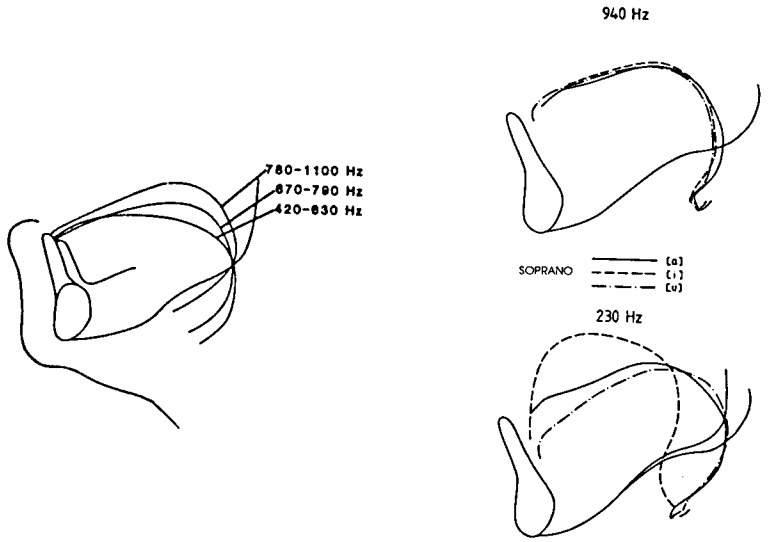


Figure 5a. Tongue shapes relative to the contour of the lower mandible observed in the F0 ranges indicated during the subject's kölning (left panel). Figure 5b (right panel) shows the corresponding data for a professional soprano singer at the F0 values indicated (data from Johansson & al., 1985 [4]).

expected. On average, the tract was shortened by 35 mm between F0=300 Hz and F0=800 Hz, saturating at about 130 mm for higher F0 values. A more moderate decrease was found in the subject's singing.

The formant frequencies were estimated from area functions derived from the radiographs. Lateral tracings of the mid sagittal vocal tract contour were made and these tracings were then converted to area functions by means of the method described by Lindblom & Sundberg [6]. The area functions were then realized in terms of tubes consisting of piles of Plexiglas washers with center holes of different sizes. The resonance frequencies of these area functions were determined by sine sweep excitation by means of the ionophone sound source [7]. The resulting formant frequencies are shown in Figure 6. In kölning, F1 tended to match F0 rather accurately while F2, F3 and F4 remained at about 1700 Hz, 2500 Hz, and 3000 Hz throughout the pitch range studied. Also in the performance of the folk tune, F1 tended to track F0, but the other formants showed more variation with F0. In the same Figure corresponding data collected from the professional soprano singer are shown, revealing a different pitch dependence of F2, decreasing with F0 for the front vowels and increasing for the back vowels.

Summarizing, as compared with normal singing and professional soprano singing, kölning seems quite special with respect to sound level, subglottal pressure, articulation, and formant frequencies. Apart from this, the intonation and the melodic patterns are also quite characteristic. It is difficult to realize why this particular singing style has developed in the Swedish herding culture. The high sound levels would reflect the need for reaching out over large distances. These high sound levels of course raise certain demands on the voice which may entail the formant frequency behavior. However, also operatic soprano singers need to produce extremely loud tones in order to be heard

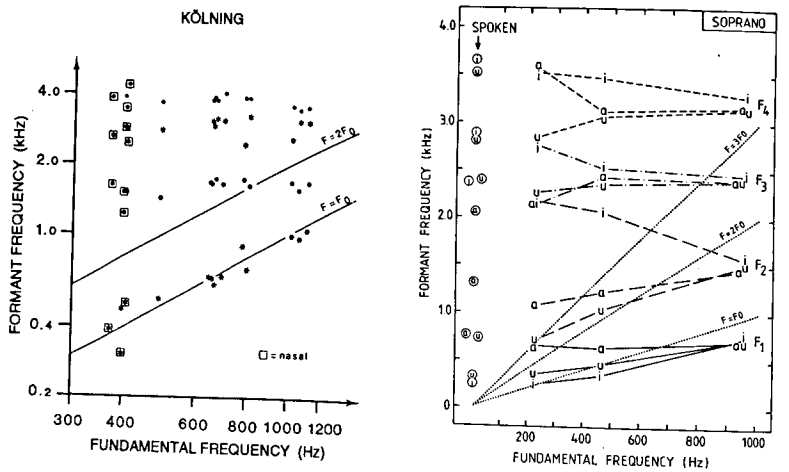


Figure 6. Formant frequencies in kölning estimated from area functions derived from tracings of the X-ray profiles during the subject's kölning (left panel). The right panel shows corresponding data for a professional soprano singer (data from Johansson & al., 1985 [4]).

in large opera houses and there, a different type of singing technique has developed. Thus, keeping in mind that kölning attracts rather than repels the cattle, we may ask whether the special kölning technique has developed to meet the esthetic and auditory demands of the cattle. This question, however, must be left for further investigation.

Jojk

Further north in Sweden another vocal peculiarity can be found, the jojk. This is a song mostly performed by males in various situations, offering another example of a rather special voice use. An acoustical investigation of jojk was carried out. The material was a documentary recording kindly supplied by Svenskt Visarkiv.

A jojk contains text and melismatic tone sequences on various vowels. The melodical structure is repetition of a short melodic sequence mostly using very few tones within a narrow F0 range. Jojks are performed with a hoarse, speech like voice quality, very far from operatic singing.

Formant frequencies were determined for some vowels from a jojk. The result is shown in Figure 7 in terms of a F1&F2 graph. The most striking difference as compared with speech is that F2 remains close to 1000 Hz for all back vowels while for front vowels a more speech like pattern is observed, starting around 2000 Hz for [i:] like vowels and approaching 1600 Hz for the [ae:] vowel. F3 is generally quite low, suggesting a constantly retracted tongue tip.

F0 patterns also are quite special. As in most folkloristic styles of singing there is no trace of the Western diatonic scale. In addition, however, F0 continually glides upward. An example is offered in Figure 8 showing measured F0 values of the four scale tones contained in this jojk. In the figure, the boundaries between

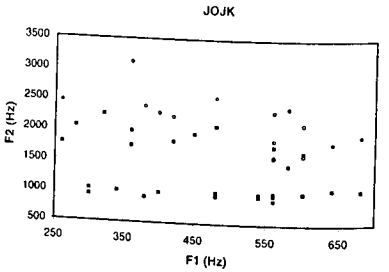


Figure 7. F2 (filled squares) and F3 (open circles) as function of F1 measured in a jojk.

verses are marked by a break in the heavy line joining the data points for the fifth. F0 for all scale tones are seen to increase linearly with time. The rise for the top pitch is almost perfectly linear; the mean rate across all scale tones is 1 Hz/sec. The greatest variability is observed for the lowest tone, probably because of measurement difficulties.

The tonal center of this jojk is the fifth, the tuning of which shows a systematic variation with melodic structure. The tuning is gradually sharpened during the three final tones of each verse, and then starts somewhat flatter at the beginning of next verse. The dashed lines represent pure intervals relative to the central scale tone, the fifth. The minor sixth is almost perfectly tuned, while the third grows increasingly flat.

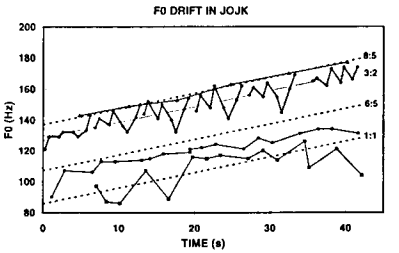


Figure 8. Mean F0 values, determined by means of histograms, for the scale tones in a jojk shown as deviations from just tuning. The unit is cent, i. e., hundredth of a semitone. The thin line shows the best linear fit for the fifth of the scale, and the heavy dashed lines represent pure intervals for the other scale tones.

Discussion

It is interesting to observe these very special data on articulation, formant frequencies and intonation in these two folkloristic examples of Swedish voices in music. In earlier investigations great differences have been revealed between speech and Western operatic singing [8]. However, these folkloristic singing styles deviate from speech in a different way than operatic singing. We have analyzed only one example of kölning and one example of jojk. Still it is interesting to observe the very special characteristics of these two examples of singing styles. They certainly invite to speculations as to the background of these characteristics in these two particular examples. Also, as the two examples analyzed were published as documentary recordings, there seems to be no reason to doubt that they are representative.

With respect to the special tuning of the scale tones the differences between kölning and jojk, on the one hand, and operatic singing on the other, the reason may be acoustical. The main reason would be that, although harmonic spectra are produced in both cases, kölning and jojk are performed without accompaniment. In operatic music, singing is mostly accompanied by instruments which also produce harmonic spectra. Under such conditions beats will occur in consonant intervals if the tuning differs too much from just. The reason is that some partials are common to the tones produced, and these partials will differ in frequency and thus generate beats, if the tuning does not approach just. In kölning and jojk the singers are free to chose intonation as they please. It is frequently assumed that part of the beauty of music relies on the use of ratios between small integers for the scale tone frequencies. The departures from just tuning in kölning and jojk shows that this is by no means true.

The steady rise of the tuning reference in the jojk is interesting. We may speculate that the perceptual effect is an exciting conflict between the perceived intervals, i. e., the pitch difference between the scale tones which reamains basically constant, and the perceived tuning reference which constantly is drifting upward. It is also interesting that small perturbations of the change in reference was used for marking the structure; the drift was increased toward the end of each verse and started with a relatively lower pitch at the beginning of each verse.

The reason for the differences between speech and operatic singing seems to be the need for a loud voice capable of successfully competing with the sound of a loud orchestral accompaniment. A loud voice is needed also in kölning. Yet, the formant frequency strategies used in these cases are not the same. In both cases the principle is used that F1 is raised to a

frequency close to F0 as soon as F0 would otherwise exceed F1. In the front vowels, on the other hand, F2 is lowered with rising F0 in operatic singing while in kölning it seems to remain close to 1700 Hz regardless of F0. The reason may be that only front vowels seem to be used in kölning which results in a homogenity of vowel timbre. Thus, there is no need to reduce vowel quality differences between vowels. This would be important in operatic singing.

Loud voice production is not a concern in jojk. Here, both back and front vowels occur, although F2 values lower than 1000 Hz were not observed. It is possible that the simple melodic structure allows a greater freedom with regard to vowel quality.

The extremely high larynx positions used in kölning is another interesting finding. An elevated larynx position is generally regarded as harmful to the voice among singing teachers. This is certainly true under certain conditions. The behavior of our subject, who had performed kölning during most of her long life and suffered from no phonatory problems, indicates that an elevated larynx position does not necessarily cause damage to the voice.

Conclusions

In conclusion, these folkloristic styles of singing offer striking examples of man's desire to decorate reality with ornaments and patterns. The great differences from operatic singing suggest that the conditions under which music is performed represent a factor of relevance to the articulatory and acoustic characteristics of vocal art. The study of folkloristic singing styles is likely to widen the views and complement our understanding of both music and voice function.

Acknowledgements

The assistance of Monica Thomasson and Bo Lantz in gathering data and preparing the camera ready copy is gratefully acknowledged. The research was supported by a grant from the Swedish Research Council for Technical Sciences.

References

[1] Ternström, S (1991) Physical and acoustic factors that interact with the singer to produce the choral sound, *Journal of Voice 5*, 128-143.

[2] Johnson A (1986) *Sången i skogen*, dissertation, Uppsala University, Department of Musicology.

[3] Johnsson A, Sundberg J & Willbrand H (1985) "Kölning". Study of phonation and articulation in a type of Swedish herding song, in A Askenfelt, S Felicetti, E Jansson & J Sundberg, ed:s, *SMAC 83, Proceedings of the Stockholm Music Acoustics Conference 1983*, Stockholm: Royal Swedish Academy of Music, Publication No 46:1, 187-202.

[4] Johanson C, Sundberg J & Willbrand H (1985) X-ray study of articulation and formant frequencies in two female singers, in A Askenfelt, S Felicetti, E Jansson & J Sundberg, ed:s, *SMAC 83, Proceedings of the Stockholm Music Acoustics Conference 1983*, Stockholm: Royal Swedish Academy of Music, Publication No 46:1, 203-218.

[5] Tjernlund P, Sundberg J & Fransson F (1972) Grundfrequenzmessungen an schwedischen Kernspaltflöten, in E Stockmann, ed., *Studia Instrumentorum Musicae Popularis II*, Stockholm: Musikhistoriska Museet, Publication No 4, 77-96.

[6] Lindblom & Sundberg (1971) Acoustical consequences of lip, tongue, jaw, and larynx movement, *Journal of the Acoustical Society of America 50*, 1166-1179.

[7] Fransson G & Jansson E (1973) The STL-Ionophone: Transducer properties and construction, *Journal of the Acoustical Society of America 58*, 910-915.

[8] Sundberg J (1987) *The Science of the Singing Voice*, DeKalb, Lillinois, USA: N Illinois University Press.

PHONETICS - A LANGUAGE SCIENCE IN ITS OWN RIGHT?

K. J. Kohler
IPDS, Kiel, Germany

ABSTRACT

This paper starts with some remarks on the history of the ICPhS and argues for a phonetic paradigm in two stages: the heuristics of phonetic phonology, and phonetic explanation. It speaks in favour of phonetics as a language science in its own right on the basis of this paradigm.

ON THE HISTORY OF THE ICPHS

This scientific meeting is the thirteenth since its inception in Amsterdam in 1932, and it has always been called the International Congress of Phonetic Sciences. At closer inspection, two things are noticed about this name: (a) it refers to a plurality of phonetic sciences and (b) it views this plurality as an open class. In this respect, also because this plurality is meant to include parts of such subjects as psychology, acoustics and linguistics, our Congress differs in a striking way from what is practised in representative disciplines of the Humanities, such as history, or of Science, such as physics. Since a scholarly conference of international dimensions mirrors the theoretical foundations of an academic discipline and the recognition, or absence, of a unified research paradigm [1] constituting a science in its own right, we would have to conclude from the way our Congress has been conceived that the answer to the question of this plenary address is negative.

So I could stop here, and we could all go for a cup of coffee instead. But let us look into this matter more broadly and more deeply and arrive at the proposal of a better-reasoned answer which can at the same time justify - or reject - that we call ourselves phoneticians and that what we do - namely phonetics - is something special. Since the actual state-of-the-art in a subject is always the result of historical incidents and developments it will help understanding to have a brief look at how and under what auspices this Congress originated.

At the First International Congress of Linguists (The Hague 1928), de Groot proposed that an international periodical of Experimental Linguistics be started, "in order to further the cooperation of Experimental Phonetics, Experimental Psychology and Linguistics, for the study of Language" [2]. In the "Explanatory memorandum" [2] he says: "Instrumental methods are of great importance in nearly every chapter of Linguistic Phonetics, but they need improvement...the phonetician does not always start from a definite linguistic problem; he sometimes even confines the field of Experimental Phonetics to what is of no interest to the linguist at all;...his chief interest is often concentrated upon instruments and curves, instead of upon the elements and the functions of speech;..."

The type of experimental phonetics de Groot had in mind was the one practised at his time by such scholars as E. W. Scripture and G. Panconcelli-Calzia. The former saw the 'nature of speech' in measurement-numbers and characterized the phonetic scientist as someone that "might be - and preferably should be - congenitally deaf and totally ignorant of any notions concerning sound and speech." [3, p.135]. The latter explicitly incorporated phonetics into physiology as part of the study of motion, like walking, running, jumping, and therefore regarded phonetics as a natural science, noticing with great satisfaction that the 'philologus auricularius furibundus' of late was getting rare [4, pp. 8,18]. Scripture was present at this congress and succeeded in founding the International Society of Experimental Phonetics

on 11 April, 1928, the day after de Groot made his proposal.

As the president of the International Society of Experimental Phonetics, Scripture planned a first congress of experimental phonetics, which was then held at the Bonn Phonetics Institute in 1930 and organized by P. Menzerath [5]. The second congress of the Society was scheduled in Amsterdam for 1932. However, the Dutch Organizing committee under the chairmanship of the psychologist van Ginneken, and with the phonetician Louise Kaiser as the secretary and the linguist de Groot as a member, decided to invite the "Internationale Arbeitsgemeinschaft für Phonologie" of the Prague Circle, which constituted itself in 1931, and the Amsterdam congress was, therefore, to be "The Second Congress of the International Society of Experimental Phonetics and the First Meeting of the Arbeitsgemeinschaft für Phonologie" as parts of an "International Congress of Phonetic Sciences".

The intentions were clear: the narrow field of the science of experimental phonetics was to be broadened by bringing in the linguistic orientation. This is in keeping with de Groot's summing up [2]: "Phonetics has up to now been too "practical", too didactical; instrumental Phonetics too physiological, too physical, too materialistic; Linguistics too much afraid of instruments." Practical phonetics, of course, referred to the activities of the International Phonetic Association and its prime concern with transcription and pronunciation teaching in foreign languages. So the Dutch organizers had three phonetics branches in mind right from the start: practical phonetics, experimental phonetics and phonology. It was consequently only a small step to broaden the field even further: "After some deliberation and in view of the recent reorganization of the Dutch Society of Phonetics [which in 1931 replaced the Dutch Society of Experimental Phonetics, founded in 1914] we decided that it would be wise to make the sphere of activity of the congress as extensive as possible and to have phonetic sciences treated in the widest sense." [6] The aim was "that all those who were interested in any aspect of speech sounds should meet and work together" [7].

A circular announcing the congress and its scope was sent out at the end of December 1931, upon which Scripture decided not to hold a Congress of the International Society of Experimental Phonetics. This was the birthday of the International Congress of Phonetic Sciences for short. I think it has now become obvious why the plural was used in the Congress name. At the outset, it refers to no more than a juxtaposition of disciplines, which were still to find the common thread uniting them. This was a task for the future; for the 1932 Congress we are reminded of what Peter Ladefoged said with reference to the IPA: "[it behaves] somewhat like the Church of England - a body whose doctrine is so diffuse that one can hold almost any kind of religious belief and still claim to be a member of it." [8]

DEVELOPING A PARADIGM OF PHONETICS: FIRST STAGE

Integration of phonetics and phonology

We have now explained how the infelicitous name of our Congress originated (which, by the way, also shows a linguistic oddity, no doubt due to an insufficient proficiency of English on the part of the Dutch congress organizers, who translated "wetenschappen" into English, not realising that, contrary to continental usage, English "science" refers to natural science and would normally be in the singular). Other academic disciplines started their congresses after they had reached a common theoretical grounding for all their subsections, expounded in handbooks and expected of anybody wanting to be a member of the same academic circle. In Phonetics it worked just the other way round, and therefore the vital distinction - for the

integrity of a subject - between parts of scholarly activity areas belonging to the same conceptual core, and cooperation of disciplines across their boundaries in questions of mutual concern and interest was blurred.

The question now is whether phonetics has taken this great opportunity of being embedded in an interdisciplinary environment to develop a unifying paradigm that allows a straightforward definition of the subject and its research questions, the setting up of teaching programmes and the publication of comprehensive handbooks of the subject as a whole. The first scholar to reflect thoroughly on the relationship between experimental phonetics and phonology and their integration into what he called the "system of scientific disciplines", was E. Zwirner [9]. His answer was phonometrics [9], which established two essentials: the allocation of measurements to units of language and their statistical evaluation. This view that measurable speech signals are not primarily a physical phenomenon per se but a physical carrier structured for the transmission of meaning in communication has been repeated several times and in various places.

Over the past sixty years the leading centres of phonetic research in the world have established the integration of instrumental and experimental techniques into the context of speech communication. It is a corner stone of modern phonetics that both aspects of human pronunciation, the physical/physiological and the linguistic, are prerequisites of each other. Phonology without a detailed description of the physical manifestation of speech is abstract, and instrumental measurements without their projection onto categories of human communication, linguistic categories among others, are empty and meaningless. Under this view, phonetics includes phonology, albeit a phononlogy that is at least as closely linked to the laboratory as to the scholar's desk.

The linguistic view: linguistic phonology

So we have certainly advanced since 1932 and created the outlines of a scientific paradigm for phonetics. But in the eyes of linguists, especially of phonologists that are proud of being within the linguistic rather than the phonetic camp, and who advocate - even during keynote addresses at phonetics meetings - that they are not phoneticians the dichotomy between phonetics (conceptualized exclusively as experimental phonetics) and phonology, between Science and the Humanities, persists. Linguists and linguistic phonologists (to coin a term referring to linguists, rather than phoneticians, doing phonology) still regard phonetics as nothing more than the supplier of instrumental data and analyses for the structural slots they have established, i. e. an ancillary appendix to autonomous linguistics, which alone is thought to be capable of giving explanatory accounts of human language. If phonetics is thus devoid of this potential of explaining speech and language phenomena, of the essential ingredient in a scientific discipline, it cannot be a language science in its own right. So, although phonetics has begun to define its own unified basis the attitude of the linguistic world is still that of the thirties.

Even the institution of the Conferences in Laboratory Phonology does not contradict this statement because it is linguistics that is to be taken into the lab to substantiate its categories. The alternative procedure of phonetic measurements obtained and evaluated in the lab being taken into linguistics to confirm, adjust or refute phonological categories by independent assessment is not considered a possibility within this framework.

I would like to buttress this contention with an example from the phonology of German that illustrates the type of epiphenomena that may be created by this 'phonetics-in-phonology' approach. Until Mitleb's thesis of 1981 [10], it was a basic tenet of German phonology that

there is neutralization between word-final voiced and voiceless obstruents. In the interim generative phonology had provided a different account: because of correspondences in morphological paradigms (*Bund* vs. *Bunde*, *bunt* vs. *bunte*) the opposition is postulated at an abstract underlying level for all word positions. Mitleb took this new systematization of the same language phenomena into the lab and tested it with native German speakers who had lived in the US for various lengths of time and who were asked to read word lists containing such unusual items as "Alb" vs. "Alp", but also "weg" vs. "Weck", where there is no morphologically conditioned alternation and "Weck" represents a regionally restricted word. Mitleb being a student of Robert Port's, who in turn learned his phonetics from Leigh Lisker, it is natural that the parameters of 'voicing' he measured were vowel and consonant durations. He found statistically significant differences between the word pairs in the direction expected from the generative description and therefore concluded that the underlying morphophonemic voice distinctions are retained in the production of phonetically voiceless finals through a systematic difference in the length of the preceding vowel: quod erat demonstrandum.

However, this finding is the result of a poor methodology of data collection and processing and does not prove anything about the differentiation between these classes of obstruents in the speech of Germans in their native environment, and, of course, says nothing at all about the perceptual relevance of the statistically significant differences as a discriminative function in the communication with a listener. As long as phonology is taken to the lab in this way it will not advance our understanding of how speech communication works, but will simply constitute a self-fulfilling prophecy of autonomous linguistics, which might just as well continue to work with symbolic representations at scholars' desks. That is what Dinnsen [11] did when he claimed that careful phonetic studies would reveal the non-neutralizing character of perhaps all rules heretofore identified as neutralizations.

But sad to say, even phoneticians fall into this trap set by the way the supremacy of linguistics conceptualizes phonological form and its relation to substance. Francis Nolan [12], at the Second Conference in Laboratory Phonology, having investigated apparent assimilations of final apical to following labial or dorsal stops by electropalatography proposed that differences in lexical phonological form will always result in distinct articulatory gestures, even if overlapped and/or reduced or not discernible in the instrumental record. Here again the questions are as to how good the methodology of data collection was and what these instrumental data can teach us about reduction processes in speech production and their function in communication.

This influence of phonological categorization on the empirical and theoretical work phoneticians do is even more far reaching in the case of Browman and Goldstein's Articulatory Phonology [13]. Their postulates that gestures specified by sets of related tract variables function as primitives of phonological contrast and that gestures are never changed into other gestures, nor added, were undoubtedly triggered by the representation of lexical items with the help of contrastive invariant phonological elements, which are set up in autonomous phonology independently of any function they might have in varying environments of speech communication and which consequently remain invariant. This phonological invariance is extrapolated via the gestural score to the gestures unfolding in articulation.

The phonetic view: phonetic phonology

I have argued against this stand and will do so again in the Symposium on Speaking Styles at this Congress. In

essence my criticism runs as follows. If we, as phoneticians, are interested in gaining insight into how speech communication works, thus transcending the dichotomy of competence and performance, we need to take variants at the phonological level within the same lexical items into account because speakers produce them and listeners successfully decode them. Thus in the German utterance *nun wollen wir mal kucken* ("now let's see") from a dialogue of the Kiel Corpus of Spontaneous Speech [14], displayed in the spectrogram of Figure 1, the phonological citation form representation would be (in IPA transcription) /'nu:n 'vɔln 'vi:ɐ 'ma:l 'kʊkŋ/, but what the speaker pronounced may be symolized as [nũ: ɔ̃n̆ʷ ɛ̆ ɱɑ 'kʰʊkŋ]. This can only give a filtered replica of the articulatory movements that may be deduced from auditory and graphic evaluation of the acoustic record as having taken place. The apical gesture for the second /n/ as well as the two lateral gestures of the phonological representation of canonical word forms have disappeared, whereas the nasalization extends during the whole of what remains of the first three words; the first occurrence of the approximant /v/ has probably left its trace in a labiodental approximation during the vowel sequence of the first two words; the third /n/ is extremely short (37ms) and realised as a tap, and superimposed on it seems to be a labiodental approximation, which may be a continuation of the preceding lip configuration as well as an advance of the same feature in the following /v/, which is not realized as a separate segmental unit; the following vowel is again very short (30ms), and the labiodental gesture continues and tightens to a closure for the following /m/.

I find it impossible to relate this intricate articulatory control to the same invariant gestures as they are to be deduced from the gestural score for the citation forms, in particular the postulate of a gestural reorganization with regard to the apicals at a higher processing level than the actual articulatory execution seems to be inescapable. Simple temporal sliding and amplitude variation of gestures in the realization of the invariant score cannot explain the empirical facts fully and adequately. On the other hand, separate lexical entries for the words in different communicative environments is out of the question: they are decoded as the same words by the listener (and are, therefore, different from such instances as *zu dem* in *er kam zu dem Schluß, daß...* ("he reached the conclusion that...") versus *zum* in *er kam zum Schluß* ("he came to/at the end")).

Phonology as descriptive heuristics: complementary phonology

In view of these problems with the postulate of invariant phonological units, e. g. phonemes, the question arises as to whether it has contributed anything to the study of language and speech communication. The answer is: "Of course it has!" But the linguistic categories of any phonological model can, at best, only function as heuristic devices, 'As Ifs' in Vaihinger's sense [15], that provide a preprocessing of spoken language data for them to become accessible to further phonetic analysis [16]. Particularly in the case of connected speech, be it read text or spontaneous dialogue, the phonological categorization allows the reduction of a large variability to a small number of entities in canonical word forms that may be listed in a lexicon and to which actually occurring pronunciations are referred. Especially the segmental concept of the phoneme is extremely useful here, e.g. for the labelling of acoustic data bases and for subsequent data retrieval in computer data banks, provided it is integrated with long componential features in a complementary phonology. So in the Kiel Corpus of Spontaneous Speech [14], the utterance of Figure 1 is represented in SAMPA notation as

n u: -MA n-+ v- O- l- @- n+ -MA
v- i:6-6+ m a: l-+ k -h 'U k @- n-N.

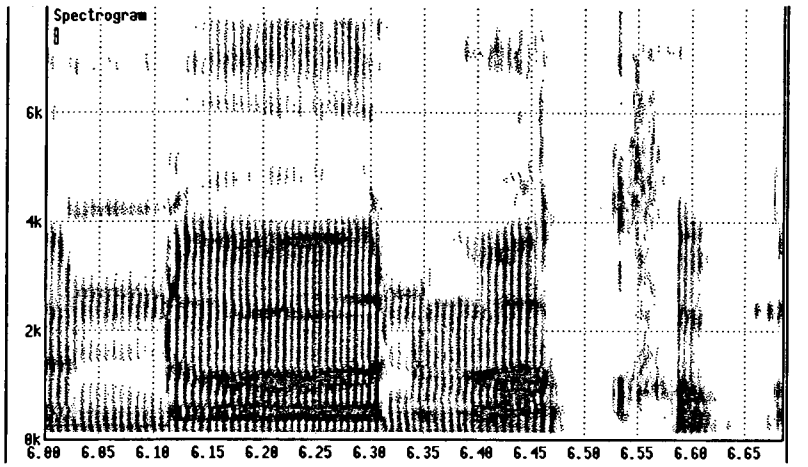


Figure 1. Spectrogram of the German utterance "nun wollen wir mal kucken" from the Kiel Corpus of Spontaneous Speech [14]

For the use of '-' and '-MA' see my contribution to the Symposium on Speaking Styles.

PHONETIC EXPLANATION

But none of these phonological devices are explanatory, they are heuristic and descriptive (and, unfortunately, in a large number of purely linguistic phonological studies they are not even that because, as autonomous symbols on paper, they lack the connection with the spoken word). Björn Lindblom has argued on many occasions, e.g. [17], and will certainly develop this point further in the Plenary Symposium on Saturday, that the explanatory questions about speech communication are not answered by phonology as we know it, because it lacks the functional viewpoint with regard to the communicative purpose of speech. It cannot explain why sound systems are the way they are, why speakers change their phonetic output in different situations in the ways they do, and why listeners are still able to decode extremely reduced speech production with great ease. To be able to provide insightful answers to these fundamental questions about speech and language, phonologists would have to step outside their autonomous linguistics field and set up hypotheses that are on the one hand independent of the data to be explained and that are on the other hand related to the biological and social conditions of humans communicating by speech.

In connection with the example of Figure 1 one decisive question is whether any reduction could have taken place at random, or whether the output is structured in highly constrained ways that only allow certain types of deviations from citation form utterances, and this question must be seen under a perspective that goes beyond the individual language, but relates to the physical make-up of the human sound producing system. So the question of language and speech universals is intimately linked to the explanation of individual language data. The specific example comes under three principles: the general instability of apical gestures [18], the greater reduction in word-final than in word-initial position for reasons of word detectability by a hearer, and the greater reduction in non-prominent function words for rhythmical reasons in a stress-timed language like German. Since apical laterals require greater muscular coordination than apical

closures in nasals and plosives, they are more easily dropped than the latter in gestural sequence, when the special conditions for reduction obtain. As the movements of the velum are more sluggish for physiological reasons the nasalization between several nasal-oral-nasal sequences occurs as a matter of course, particularly in fast speech. The superposition of labiodental constriction on tongue body and tip gestures over a relatively large stretch of articulation is made possible by the anatomical and physiological independence of the lower lip and by its slow execution of movements, especially in a repetitive frame /v...v/. So the articulatory manifestations found in this utterance in relation to the canonical forms can be deduced from general principles which would also be applicable to other languages, given the same rhythmic structure and the same tolerance of hearers under social constraints. Historical sound change exemplifies these developments over and over again in the most diverse languages, as John Ohala has pointed out on several occasions, e.g. [19].

The other pertinent explanatory questions related to the utterance in Figure 1 are: "How far do listeners allow degradations of this sort to go before they have to ask for repetition because they do not understand?" and "How do speakers manage to decode such reduced speech correctly and with such ease?" No answers are available as yet. But here is a specific task for phonetics, which falls outside linguistic phonology, which the latter could not handle, and which it would not even be interested in proposing.

A PARADIGM OF PHONETICS: SECOND STAGE

So the paradigm of phonetics is taking shape. The integration of phonology and the physics of speech in a phonetic phonology, as expounded above, constitutes the first part of this paradigm: a heuristic framework for phonetic descriptions of languages in all their speech manifestations. Built on this is the second part of the paradigm: the functional view of speech production and reception - the explanation of the speech communication process between a speaker and a listener (and we may add, the acquisition of language and speech) with reference to the physics, biology and social environment of homo loquens. No other discipline has or wants to have such a paradigm. Linguistics is content within its autonomous framework detached from the purpose it may be put to in communication; acoustics and engineering (except for the engineers that have adopted, or in the case of the colleagues at KTH even assisted in creating, the phonetic paradigm) are only interested in the physical perspective, as is illustrated, for example, by the way they deal with automatic speech recognition or with building block synthesis.

Picking up the theme of Francis Nolan's paper on "Phonetics in the next ten years" at the last Congress [20], I would now venture to say that the coming years will see a consolidation of this paradigm of modern phonetics as a unitary discipline of the spoken medium of language, an essential interface between the pure and simple signal approach of physics and engineering and the symbolic orientation of semantics, syntax and linguistic phonology in lingistics. Phonetics will thus occupy a key position in enquiries into the functioning of speech communication at the levels of pure research as well as application. Of course, there must and always will be interdisciplinary cooperation with neighbouring fields that have different paradigms, but can contribute special expertise which the phonetician does not have, e.g. acoustics, physiology, psychology, linguistics.

This paradigm also necessitates the training of phonetics students in symbol as well as signal aspects of speech and language, including analytic listening and transcription techniques, speech signal processing and experimental methods. A common core curriculum will be developed and a "Handbook of Phonetic Science" will be written in accordance with the paradigm. There are already initiatives for a phonetics curriculum at the European level of the ERASMUS programme of Phonetics and Speech Communication, although its compilation of subject areas is still too encyclopedic and juxtapositional with not enough reference to the phonetic paradigm.

CONCLUSION

I can now go back to the title of this talk. Yes, phonetics, in my view, is a language science in its own right by virtue of its subject matter, and it is well on its way towards asserting itself as such. There is still a good deal of hard work ahead of us. Let's begin with an evolution of our historical traditions and drop just one letter and two phonemes, at the end of the Congress name!

REFERENCES

[1] Kuhn, T.S. (1970), *The Structure of Scientific Revolutions*, Chicago: University of Chicago Press.

[2] de Groot, A. W. (1928), "Proposition 2 - Explanatory memorandum", *Actes du Premier Congrès International de Linguistes*, pp. 6-9, Leiden: Sijthoff's Uitgeversmaatschappij.

[3] Scripture, E.W. (1935), "Bulletin of the International Society of Experimental Phonetics III", *Arch Neerl Phon Exp*, vol. 11, pp. 133-147.

[4] Panconcelli-Calzia, G. (1948), *Phonetik als Naturwissenschaft*, Berlin: Wissenschaftliche Editionsgemeinschaft.

[5] Menzerath, P. (1930), *Bericht über die I. Tagung der Internationalen Gesellschaft für experimentelle Phonetik*, Bonn: Universitäts-Buchdruckerei Gebr. Scheur.

[6] (1932), *Proceedings of the International Congress of Phonetic Sciences*, Amsterdam.

[7] Fischer-Jørgensen, E. (1984), "Some aspects of the 'Phonetic Sciences', past and present", in M.P.R. van den Broecke, A. Cohen (eds.), *Proc Xth Inter Cong Phon Sc*, pp. 3-11, Dordrecht: Foris Publications.

[8] Ladefoged, P. (1990), "Some reflections on the IPA", *Journal of Phonetics*, vol. 18, pp. 335-346.

[9] Zwirner, E., Zwirner, K. (1936), *Grundfragen der Phonometrie*, 2nd ed., Basel: Karger, 1966.

[10] Mitleb, F. (1981), *Segmental and non-segmental structure in phonetics: evidence from foreign accent*, PhD. diss. Indiana University, Bloomington.

[11] Dinnsen, D.A. (1983), *On the characterization of phonological neutralization*, Bloomington: Indiana University Linguistics Club.

[12] Nolan, F. (1992), "The descriptive role of segments: evidence from assimilation", in J. Docherty, B. Ladd (eds.), *Papers in Laboratory Phonology II*, pp. 261-280, Cambridge: CUP.

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

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

[15] Vaihinger, H. (1920), *Die Philosophie des Als Ob*, Leipzig: Meiner. (Transl. Ogden, C.K. (1965), *The Philosophy of 'As If'*, London: Routledge & Kegan Paul.)

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

[17] Lindblom, B. (1980), "The goal of phonetics, its unification and application", *Phonetica*, vol. 37, pp. 7-26.

[18] Kohler, K.J. (1976), "Die Instabilität wortfinaler Alveolarplosive im Deutschen - eine elektropalatographische Untersuchung", *Phonetica*, vol. 33, pp.1-30.

[19] Ohala, J. (1983), "The origin of sound patterns in vocal tract constraints", in P.F. MacNeilage (ed.), *The Production of Speech*, New York/Heidelberg/ Berlin: Springer.

[20] Nolan, F. (1991), "Phonetics in the next ten years", *Proc XIIth Inter Cong Phon Sci*, vol. 1, pp. 125-129, Aix-en-Provence: Université de Provence.

THE PERCEPTION OF STOP CONSONANTS: LOCUS EQUATIONS AND SPECTRAL INTEGRATION

A. Eek and E. Meister
Laboratory of Phonetics and Speech Technology
Institute of Cybernetics, Tallinn, Estonia

ABSTRACT

Formant transitions did not provide the primary context-independent cues for place of articulation. Locus equations showed relational invariance for stop categorization in the production space but they had not the same relevant role in perception. The connection of the strongest peak of the gross shape of the spectrum sampled at the stop release and the gravity centre of the following vowel demonstrated a reliable cue for stop categorization in the perception space.

INTRODUCTION

Although the classical locus concept is applicable to two-formant synthesis, it does not reflect adequately the reality in natural speech because it fails to document an invariant F2 or F3 loci for different vowel contexts. Another approach - the concept of locus equations was recently investigated as a potential metric capable of illustrating relational invariance for stop categorization in cross-linguistic perspective [1]. The perceptual relevance of locus equations has not been systematically studied.

A lot of data have been collected about differences in spectral energy distribution immediately after the burst release or about relative changes in distribution of energy from the burst release to the onset of voicing. The gross shape of the spectrum sampled at the stop release has showed an invariant shape for each place of articulation. The gross shape peculiarities provide the primary context-independent cues whereas the formant transitions from the stop release to the vowel nucleus provide secondary context-dependent cues linking the abrupt transient to the syllable nucleus and creating a perceptual impression of the syllable as an integral unit [2]. Below we shall test some aspects of stop perception.

SPEECH MATERIAL

The speech material consists of the Estonian CVV syllables beginning with *p, t, k* (voiceless unaspirated stops with average burst durations 18, 31, 37 ms, resp.) and followed by 9 long vowels *i, e, ä, ü, ö, õ, a, o, u*. Such syllables were read as one-syllable sentences by 1 male speaker. The speech samples were digitized at 10 kHz and autocorrelation LPC spectra were computed in Kay CSL 4300 system (Hamming filter, high-frequency preemphasis, 14 coefficients). Spectra for the vowel onset and nucleus were computed with a 10 ms time window by centering the window at the last third of the first half of the vowel F0 period; measurements were repeated on wideband spectrograms (the data are plotted in Fig. 1). Burst spectral shapes were computed with a 25 ms window by centering the window at the burst release. *k* in unrounded front vowel contexts has its strongest transient peak near F3 initial frequency of the vowel, while in back vowel contexts it lies at F2 initial frequency (the latter is also valid for rounded front vowel contexts). F2 and F3 diverge during the transition to the vowel nucleus, creating thus at the burst release a 'bottle neck'-like formation. The strongest peak of *k*'s burst in midfrequency region between 1000-2900 Hz stands out dominantly from any other peaks. Such strong compactness of spectra is unambiguously valid in unrounded front vowel contexts whereas in rounded front vowels and back vowels contexts there is another outstanding but weaker peak at 4000-4400 Hz. *t* shows the strongest burst peak at high frequencies between 3000-4000 Hz, while lower peaks have gradually weakened. *p* has the strongest burst region between 350-500 Hz in back vowel contexts. Before front vowels two first peaks of the burst are of equal intensity (the second peak coincides with F2 initial frequency) and higher transient peaks have gradually damped.

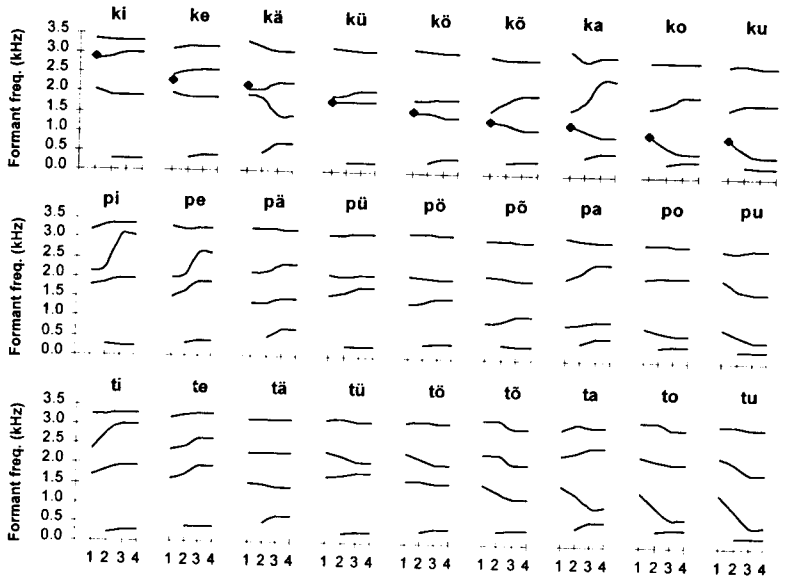


Figure 1. Formant trajectories in CVV syllables with syllable-initial stop consonants. Measurement points: 1 - stop release; 2 - vowel onset; 3 - vowel nucleus; 4 - vowel nucleus continued; ◆ - the strongest peak of k burst measured at stop release.

PERCEPTION EXPERIMENTS

Experiment 1. In this experiment we used 27 original CVV syllables + the same syllables without burst (four-formant acoustic patterns of without-burst-stimuli were described by the data of measurement points 2, 3 and 4 in Fig. 1). 54 different stimuli were presented 4 times in random order to 13 listeners; their task was to identify a stop consonant at the beginning of each syllable.

The direction of the F2 transition is not invariant (e.g. in Fig. 1: F2 rises in the syllables *pi, pe* and *ti, te,* but falls in *po, pu, to, tu* and *ko, ku*). Presumably the degree of movement freedom of the tongue body is the biggest in *p*-syllables and the least in *t*-syllables (cf. locus equations in Fig. 2). Therefore, in the case of without-burst-stimuli, labial stops should receive the lowest identification scores.

All with-burst-stimuli were correctly recognised. But for without-burst-stimuli we obtained the results opposite of what we expected: only *p* was recognised in all vowel contexts. Interpreting listeners' responses we cannot ignore particularly the relations between F2 and F3. Despite the fact that in front vowel contexts *p* transitions were moderate, a labial stop was identified 90-100%. We suppose that in these cases marked F3 rising transitions to the direction of a gravity centre of front vowels take over the function of weakly marked F2 transitions. This can also explain why all *t*-syllables in front vowel contexts were recognised 70-93% as beginning with *p*. This supposition is indirectly confirmed by a simple test: after the removal of frequencies higher than F2 from the vowel spectra of with-burst *pi, ti, ki* syllables, all listeners perceived the remaining original F1 vs. F2 spectrum of

the syllables as if these were *pü, tü, kü*, just as they had perceived the original with-burst *pü, tü, kü* on the basis of F1 vs. F2 spectrum. Syllable-initial *t* was perceived only in the context of back vowels *õ* and *a* (the biggest fall from the highest F2 onset frequency), while *t* before *o* and *u* was perceived ca 50% as *p* (the fall from the lowest F2 onset). *k* was perceived before *ä* (the only 'bottle neck' formation preserved after deleting the bursts), while for all other front vowel contexts without-burst-syllables were perceived as long vowels (due to preserved negligible transitions). Before back vowels, *k* was mostly identified as *p* (a fall from low-frequency F2 onset; weak higher formants probably have no essential role). The removal of bursts destroyed the entirety of transitional trajectories. Will the results improve if we complete CVV transitions by adding F2 changes without noise components of the transitional part of the burst to the vowel transitions (see below)?

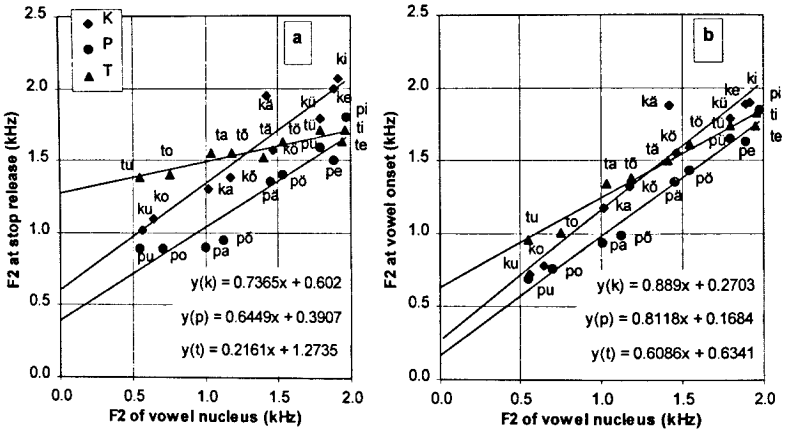


Figure 2. Locus equations for k, p, t. The extent of consonant-vowel coarticulation in CVV syllables. a - F2 measured at stop release; b - F2 measured at vowel onset.

Experiment 2. We used the same 27 without-burst-stimuli known from the previous experiment as base syllables. F2 onset frequencies of the four-formant base syllables were altered in both directions limited by F1 and F3 frequencies. The stimuli were generated in Kay CSL 4300 system using LPC synthesis. Each stimulus was repeated 5 times in succession with 1 s pauses; there was a 3 s pause between different stimuli for marking responses. The listeners' task was to mark a stop consonant at the beginning of syllables; it was also allowed to mark unidentifiability when the syllables were perceived as long vowels. Results have been presented in Fig. 3.

As a rule, rising F2 transitions were preferred for *p* responses, falling transitions for *t* responses (in back vowel contexts) as well as for the identification of *k* (in front vowel contexts if a 'bottle neck' formation was created). *t* before front vowels and *k* in back vowel contexts (except before *a*) were not identified (for probable reasons see above). It should be noted that in the perception space *p* may be represented even by a locus equation slope of 0 (y intercept about 500 and 900 Hz). F2 transitions did not provide the primary context-independent cues for place of articulation. Locus equations showed relational invariance for stop categorization in the production space but they had not the same relevant role in perception.

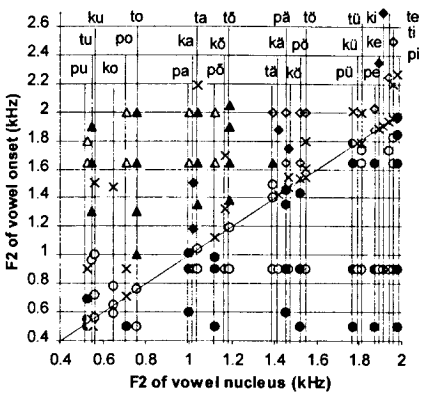


Figure 3. Perception of stop consonants in CVV syllables without bursts. Each point represents a stimulus that was identified as p(● ○), t(▲ △), k(◆ ◊), or as a long vowel (x) for more than 50% of the responses. Diagonal line displays the cases where F2 of vowel onset equals to F2 of vowel nucleus (no F2 transitions).

Experiment 3. 9 *p*VV base syllables without bursts were used generating two-formant vowel patterns (F1 and its transitions were unchanged and corresponded to F1 of the vowel; F2' of the corresponding vowel type was fixed as the value of F2; for spectral integration in vowel perception see [3]). Two series of vowel stimuli were generated: (a) vowels with a straight F2' (no transitions); (b) vowels with 50 ms transitions directed to the strongest peak of the preceding stop bursts in connection with the corresponding vowel type (for *p*-stimuli - rising transitions; for *k*-stimuli - straight transitions in front vowel contexts and falling in back vowel contexts; for *t*-stimuli - falling transitions). The corresponding original stop burst was added to each vowel pattern. The listeners' task was to identify a stop consonant at the beginning of each syllable. Results have been presented in Fig. 4.

Identification scores were 75-100% for all cases. There were no essential differences between scores given to the intended consonant with moving and straight transitions (only for *p*-syllables 10-15% higher identification scores were registered with rising transitions). The

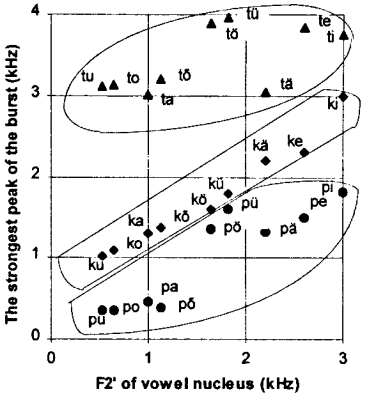


Figure 4. The perception space of stop consonants.

connection of the strongest peak of the burst and the gravity centre of the following vowel provides a reliable cue for stop categorization in the perception space. Supposedly listener's perception mechanism fixes the gravity centre both in the gross spectral shape of the stop burst and in the following vowel; linking of these centres supports listener with sufficient information for making decisions about syllables as a whole.

ACKNOWLEDGEMENTS

This work was supported by Estonian Science Foundation, Institute of Cybernetics, Eduard Treu Memorial Foundation and Aleksander Kaelas Foundation. We are thankful to Mall Laur and her students acting as listeners.

REFERENCES

[1] Sussman, H.M., K.A. Hoemeke, F.S. Ahmed (1993), "A cross-linguistic investigation of locus equations as a phonetic descriptor for place of articulation", *JASA* 94, 3: 1256-1268.

[2] Blumstein, S.E., K.N. Stevens (1980), "Perceptual invariance and onset spectra for stop consonants in different vowel environments", *JASA* 67, 2: 648-662.

[3] Eek, A., E. Meister (1994), "Acoustics and perception of Estonian vowel types", *Phonetic Experimental Research*, Institute of Linguistics, University of Stockholm, *PERILUS* XVIII: 55-90.

ACOUSTIC STUDY OF VC TRANSITIONS FOR HINDI STOPS

Manjari Ohala
San Jose State University, San Jose, California, USA

ABSTRACT

Do languages like Hindi with large segment inventories show less contextual variation? VC transitions of stops in different vocalic environments were examined to see if (a) the different places of articulation were cued similarly, and (b) if the five places of articulation were well-differentiated. The results show that acoustic cues for place are highly variable and context dependent -- even in a language with a large segment inventory.

INTRODUCTION

Hindi has a large and, by some measures, a crowded segment inventory: 33 consonants and 11 vowels. If distinctive geminate consonants and nasalized vowels were included, the total number of phonemes would be 85. Of the 33 singleton consonants, 20 are stops or affricates, produced at 5 different places of articulation: labial, dental, retroflex, palatal, and velar (the palatal stops are affricates). See Table 1 (where geminate consonants are not indicated; see [4, 5] for further details). This raises the question of how well these sounds are differentiated; how is coarticulation managed? Following some ideas of Lindblom [2] and Lindblom and Maddieson [3], one might expect Hindi with its crowded phoneme space to permit less allophonic variation than might be the case with a language with a smaller phoneme inventory. To the extent that two different phonemes or even two different sequences of phonemes exhibit similar acoustic patterns, it would presumably make the task of the listener more difficult. In general, the amount of variability and thus the ambiguity inherent in the signal should be inversely related to the inventory of possible message units.

p	t	ʈ	t͡ʃ	k		
pʰ	tʰ	ʈʰ	t͡ʃʰ	kʰ		
b	d	ɖ	d͡ʒ	g		
b̤	d̤	ɖ̤	d͡ʒ̤	g		
f	s		ʃ			
	z					
m	n					
w			j			h
	r	ɽ				
	l					
	i	ĩ		u	ũ	
	ɪ	ɪ̃		ʊ	ʊ̃	
	e	ẽ		o	õ	
	ɛ	ɛ̃	ə ə̃	ɔ	ɔ̃	
	æ		ɑ ɑ̃			

Table 1. Segment inventory of Hindi (excluding 25 geminate consonants).

To address this issue I examined VC transitions of stops in different vocalic environments and asked (a) are these places cued similarly in different vocalic environments? and, (b) are the five places of articulation well differentiated by formant transitions alone i.e., without the benefit of stop or affricate releases? Additionally, I noted whether formant patterns characteristic of place are similar to those found in other languages. Although it is now known that other cues such as rate of formant movements, stop bursts, etc., also play a role [6], they have not been examined in this study, and thus in this respect the study is limited and should be treated as a preliminary investigation.

METHOD

Speakers and speech corpus

I recorded three male native speakers of Standard Hindi uttering syllables of the form /pVC/ where V = one of the following eight front/central/back vowels [i, ɪ, ɛ, ə, u, ʊ, ɔ, a] and C = a voiced or voiceless (unaspirated) stop that was bilabial, dental, retroflex, palatal, or velar. (The palatal stop is actually a palato-alveolar affricate.)

The recordings were made in the Language Laboratory of the Jawahar Lal Nehru University, Delhi, using high-quality analog portable equipment. All test words were read in two different random orders in the frame vo____aya 'he____came'.

Analysis Methods

The recorded speech was band-pass filtered at 68 Hz to 7.8 kHz, digitized at 16.7 kHz and analyzed with the aid of waveform and LPC spectral displays produced by the CSRE speech analysis software and related programs. The VC formant transitions (the last 100 msec of the vowel) was extracted from computed spectrograms and analyzed. The results given below are for the most part based on 9 tokens per utterance (3 tokens X 3 speakers) (In a few cases there are fewer tokens, but never less than 7.)

Results

Fig. 1 gives the averaged formant tracks for three vowels /i a u/ before 5 different places of voiced stop (or affricate). (The formant tracks for voiceless consonants are not given due to space limitations but they were similar to those for the voiced consonants.) The rightmost column gives a superimposition of the formant patterns from the leftmost three columns. This last column of formant tracks is difficult to read for the sake of isolating the patterns for specific VC combinations but it does show global patterns better, e.g., presence or absence of a restricted range of terminal frequencies for the VC transitions. Mid vowels are not represented but, in general, their patterns were interpolated between those shown here, e.g., the pattern for /ɛ/ is approximately in between those for /i/ and /a/.

DISCUSSION

The following patterns can be noted (I also give the patterns for the vowels [ɪ, ʊ, ɛ, ɔ, ə] even though they have not been included in Fig. 1).

Characteristic differences in VC transitions: bilabials showed the characteristic lowering of F2 (and/or F3) after front vowels but not after back vowels where F2 and F3 were essentially flat. Velars exhibited the familiar convergence of F2 and F3 only after [i]. After other vowels the transitions were more or less flat. For dentals, F2 showed the typical bending toward mid-frequency terminal values in the range 900 - 2000 Hz; F3 had an even narrower range of terminal values, 2500-2800 Hz. Retroflexes showed a marked convergence of F2 and F3 and also a lowering of F4, except after [i] and [ɪ] (though even in these cases there was a noticeable lowering of F4). Palatals had a rising F2 and F3 except after [i] where the transitions were flat; terminal values for F2 ranged from 1000 to 2300 Hz.

Characteristic similarities in VC transitions: after [i] and [ɪ] the bilabials and dentals have very similar patterns--lowering of F2 and F3. After [ɛ] retroflexes and velars were similar in their F2 and F3 patterns, however, retroflexes had a lowered F4. After [u] and [ɔ] bilabials and velars had similar and essentially flat transitions for all formants. After [u,ɔ, a, ə], dentals and palatals had similar transitions except that for palatals F2 rises higher and starts this rise earlier.

The formant frequencies I obtained for the VC transitions are fairly similar to those obtained for Gujarati by Dave [1] for the subset of the data that lent itself to comparison.

CONCLUSIONS

These results reinforce the accumulating evidence that the acoustic cues for place are highly variable and context-dependent -- even in a language with a

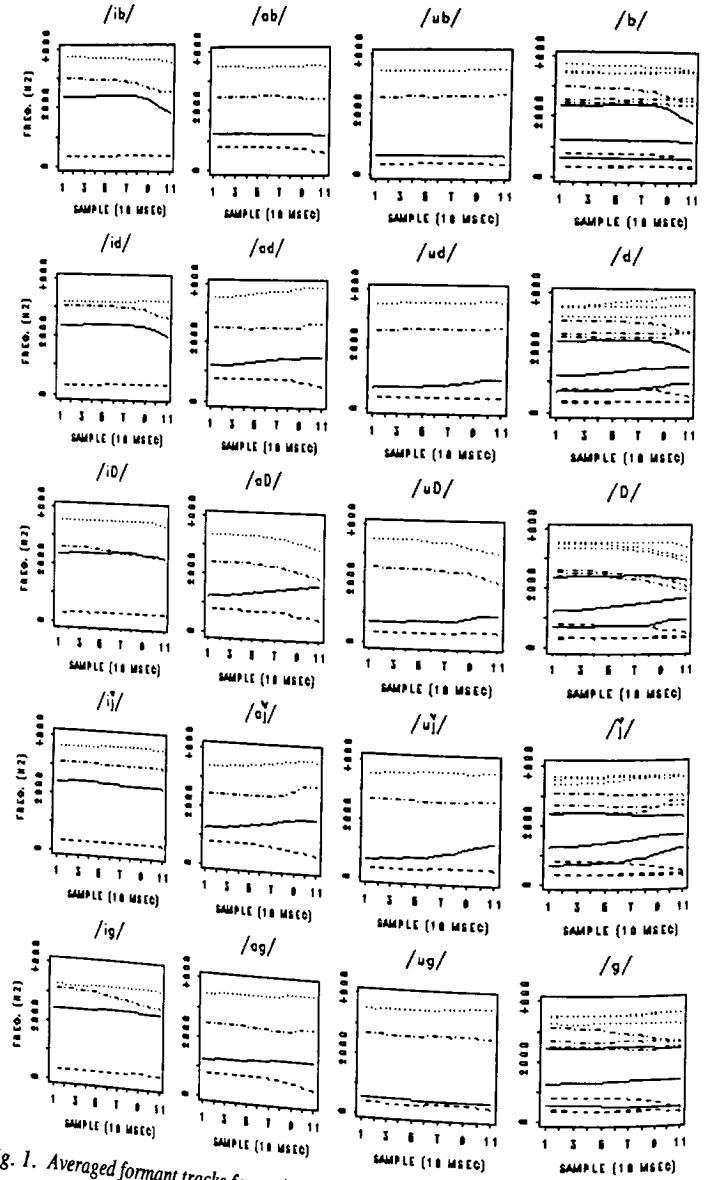


Fig. 1. Averaged formant tracks for various vowel + voiced stop sequences. Abscissa: measurement points: 10 msec intervals for a total of 11 measurements immediately preceding the vowel offset. Ordinate: frequency in Hertz. Parameters: F1 = dashed lines, F2 = solid line, F3 = dash-plus-dot lines, F4 = dotted lines. Columns from left to right: /i/, /a/, /u/, superimposition of those three vowels. Rows from top to bottom: bilabial, dental, retroflex (symbolized 'D'), palatal (symbolized 'ɲ'), velar.

crowded segment inventory. It is also of interest to note that velars (except after [i] do not show the characteristic F2-F3 convergence found in other languages. On the other hand, convergence of these two formants is found for retroflexes in some vocalic environments. F4, which is not usually considered an important cue for place, seems to exhibit a highly consistent lowering for retroflexes (as also noted by Stevens and Blumstein [6]). Finally the following caveat must be given: the VC environment (in final position) is known for neutralizing various distinctive features including place of articulation. Thus the fact that the transitions in the data reported were quite similar for a number of consonants is perhaps not so surprising.

ACKNOWLEDGMENTS

I thank A. Abbi and V.V. Sahni for their help in the use of facilities at Jawahar Lal University and T. Nearey and T. Wells for help in using the speech analysis systems at the University of Alberta. I thank John Ohala for his help with the data analysis and figures. The research was supported by a sabbatical leave from SJSU and a research grant from the U of Alberta.

REFERENCES

[1] Dave, R. (1977), "Retroflex and dental consonants in Gujarati: a palatographic and acoustic study". *Annual Report of the Institute of Phonetics, University of Copenhagen*, vol. 11, pp. 27-156.

[2] Lindblom, B. (1986), "Phonetic universals in vowel systems" In: J. J. Ohala & J. J. Jaeger (eds), *Experimental Phonology*. pp. 13-44. Orlando, FL: Academic Press.

[3] Lindblom, B. & Maddieson, I. (1988), "Phonetic universals in consonant systems," L. M. Hyman and C. N. Li (eds), *Language, Speech, and Mind*, London: Routledge. pp. 62-78.

[4] Ohala, M. (1983), *Aspects of Hindi Phonology*. Delhi: Motilal Banarsidass.

[5] Ohala, M. (In press), "Hindi", *J. Int. Phonetic Association.*

[6] Stevens, K. N. & Blumstein, S. E. (1975), "Quantal aspects of consonant production and perception: a study of retroflex stop consonants", *Journal of Phonetics*, vol. 3, pp. 215-233.

LOCUS EQUATIONS AS A METRICS FOR PLACE OF ARTICULATION IN AUTOMATIC SPEECH RECOGNITION

Eugenio M. Celdrán and Xavier Villalba
Laboratorio de Fonética, Universitat de Barcelona

ABSTRACT

In this communication it is showed that locus equations are a powerful metrics for classifying Spanish stops regarding place of articulation. 10 subjects (5 male and 5 female) produced a series of labial [p]-[b], dental [t]-[d], and velar [k]-[g] tokens for 5 vowels. The resultant three locus equations nicely characterized the three places. Moreover, a discriminant analysis using both slopes and *y* intercepts and slopes alone yielded a 100% correct classification.

INTRODUCTION

After having failed in the search of reliable invariant cues for place of articulation of Spanish stops, following Blumstein and Stevens's [1] steps, we decided to look at Sussman's [2] and Sussman et alii's [3], [4] new proposal based on locus equations. This concept was originally conceived by B. Lindblom [5], who sensed that, although variations caused by coarticulation were noticeable, there seemed to exist a close relationship between the onset F2 values —which would roughly correspond to the vowel transition— and the F2 values of the midvowel nucleus. In order to prove it, he calculated the correlation between these F2 values. Thereby he obtained a series of lineal functions such as the following:

$$F2_{onset} = k * F2_{vowel} + c$$

where the constants *k* and *c* stand for the slope and the *y* intercept, respectively. Lindblom found that the values of the constants were clearly different for each place of articulation. In other words, the slope of the regression line varied depending on the place of articulation. Therefore, the various lineal functions obtained represented different places of articulation, more precisely, different locus equations.

[3] presented a particular proposal about the basis of voiced English stops. The results they achieved were enormously encouraging: using locus equations as metrics, they obtained 93% classification rates. That is why we found it necessary to carry out a study about the usefulness of locus equations in the discrimination of place of articulation for Spanish stops. This would serve us to see the universality of this method and we would be able to study in a quantitative manner, the effects of coarticulation of vowels on stop consonants. The degree of success of this method has direct consequences on the studies about automatic speech recognition: it provides us with a metrics to classify place of articulation starting from a specific sound stimulus.

METHOD

Material of study

The study was based on the speech of five male and five female subjects of ages 20 to 30. The subjects produced the sequence [kan'CVna], where C={[p], [b], [t], [d], [k], [g]} and V={[i], [e], [a], [o], [u]}; they repeated this sequence five times for each vowel. We obtained 150 stimuli for each subject (6 stops x 5 vowels x 5 productions = 150). The informants' productions were recorded in a soundproof booth using a "Shure SM58" microphone and a cassette recorder (Marantz, model CP430).

Measurements

The stimuli were reproduced and analyzed with the Kay CSL 4300 B. The formant measurements were based on measurements via cursor on a wide-band spectrogram, with the additional LPC derivation values of each formant.

Points of analysis

Considering that Spanish stops are non-aspirated, we measured the first glottal pulse after the burst (F2 onset) and the midvowel nucleus (F2 vowel). When the vowel formant trajectory was either ascending or descending, we took the value of F2 at a middle position in the formant. In the cases in which the formant trajectory was ascending-descending or descending-ascending, we took either the maximum or the minimum value, respectively.

RESULTS

We generated thirty locus equations (3 places of articulation x 10 speakers). The results are presented in Table 1.

Table 1. Slope and y intercept values for all speakers and place of articulation

	/p/-/b/		/t/-/d/		/k/-/g/	
Subjects	Slope	Intercept	Slope	Intercept	Slope	Intercept
M1	0.74	224.42	0.59	606.83	0.93	65.62
M2	0.79	151.06	0.69	516.53	0.95	60.7
M3	0.9	-1	0.52	803.27	1.19	-242.74
M4	0.88	45.04	0.65	545.86	1.12	-167.64
M5	0.81	195.56	0.51	829.94	0.97	74.43
F6	0.87	71.62	0.64	706.28	0.95	149.37
F7	0.83	134.33	0.63	662.41	0.99	5.66
F8	0.82	94.96	0.53	920.51	0.99	79.54
F9	0.80	241.86	0.53	966.08	1.01	42.13
F10	0.84	102.99	0.50	899.88	0.92	227.21
mean	**0.83**	**126.08**	**0.58**	**745.76**	**1**	**29.43**

The mean labial slope was 0.83 (s.d. 0.05) and the labial *y* intercept mean was 126.08. The dental mean slope was 0.58 (s.d. 0.07) with a mean *y* intercept of 745.76. Finally, the velar mean slope was 1 (s.d. 0.09) with a mean *y* intercept of 29.43. In all the cases

Slope variability

To test the variability of slope values, two ANOVAs were performed. The first one made a comparison with respect to gender and showed no significant difference: $F(1,28) = 0.138$, $p<0.7172$. The other ANOVA made a comparison with respect to stop place and significant difference was found: $F(2,27) = 93.013$, $p<0$. Further comparisons between place pairs also showed significant difference: labial vs. dental yielded $F(1,18) = 88.22$, $p<0.01$; labial vs. velar yielded $F(1,18) = 30.92$, $p<0.01$; and finally dental vs. velar yielded $F(1,18) = 145.01$, $p<0.01$. To sum up, the variability of slopes was not significantly affected by gender but by place of articulation.

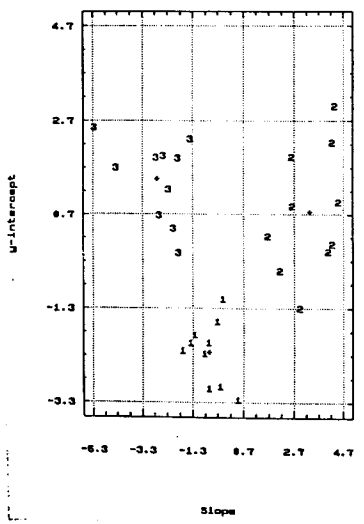
Y intercept variability

Again, in spite of the great variability of *y* intercept values, ANOVAs analysis showed non significant effects due to gender ($F(1,28) = 0.699$, $p<0.419$), but very significant effects due to place of articulation ($F(2,27) = 88.024$, $p<0$).

Discriminant analysis

Following [3], we decided to set up a discriminant analysis taking place as the classificator and the values of all locus equations (3 places × 10 speakers = 30 locus equations). Using only *y* intercepts the correct classification rates were 70% for labials, 100% for dentals, and 70% for velars (an overall mean of 80%). Nevertheless, using either slopes alone or both slopes and *y* intercepts the correct classification rates were 100% for all three place categories. Figure 1 shows the means plot for both slope and *y* intercept values.

Figure 1. Means plot of the discriminant analysis. 1 corresponds to labial stops, 2 to dental stops, and 3 to velar stops. Points show group centroids.



DISCUSSION

The experiment has shown that three clearly distinguished locus equations describe place of articulation of Spanish stops. Note that, as Table 2 shows, Spanish stops yield no slope overlapping, even though the minimum slope value of velar stops and the maximum one of labial stops are very close. This is consistent with the discriminant analysis results, which offered a 100% classification rate just using the slope values.

Table 2. Range slope values across place of articulation in Spanish stops.

	labial	dental	velar
minimum	0.74	0.5	0.93
maximum	0.9	0.69	1.19

However, before raising any conclusion, it must be taken into account the phonetic nature of Spanish stops. Firstly, Spanish voiceless stops show no aspiration, unlike the English ones. Moreover, their VOT values are quite low (6,5 ms for [p]; 10,4 ms for [t] and 25,7 ms for [k], according to [8]), which makes them quite similar to English voiced stops at initial position. Spanish voiced stops are also quite different from their English counterparts: they only occur at initial position and after nasals, and they always have a negative VOT value. Finally, we should also emphasize that the place of articulation of [t] and [d] in Spanish is dento-alveolar —i.e., the tip of the tongue touches the upper teeth, whereas the tongue blade is attached to the alveolar ridge—, as opposed to other languages like English, where they are clearly alveolar. All these differences probably explain why [1]'s method for the establishment of invariant cues for place of articulation of stop consonants failed in Spanish (see [9], [10]). Nevertheless, even though the differences, the resultant locus equations for Spanish stops have proved to be very consistent cross-linguistically, as Table 3 shows.

Table 3. Cross-linguistic comparison of mean slope values.

	labial	dento-alveolar	velar
Spanish	0.83	0.58	1
Thai [4]	0.70	0.30	
English [3]	0.87	0.43	0.66
Swedish [1],[7]	0.63	0.32	0.95
Arabic [4]	0.77	0.25	0.92
Urdu [4]	0.81	0.50	0.97
mean	**0.77**	**0,39**	**0,9**

Both labial and velar Spanish stops are very consistent cross-linguistically. The most important difference appears with dental stops. However, even such a difference can be explained by the above mentioned fact that Spanish [t]-[d] are clearly dento-alveolar and never alveolar, unlike e.g. English. That this might be the reason for the difference in dental slope values is supported by Urdu, a language showing both dental and alveolar stops: dental slope values were higher than the corresponding alveolar values, and hence closer to Spanish values.

One of our primary aims was to find invariant cues which would allow us to carry out automatic speech recognition based on phonetic features. From our statements, we believe that locus equations give us a good basis for the achievement of this aim.

ACKNOWLEDGEMENTS

We are indebted with O. Julià, L. Rallo, and H. Sussman for their help. This work has been supported by grants PB91-0278 from the DGICYT and AP92-46654991 from the Spanish Ministry of Education and Science (MEC).

REFERENCES

[1] Blumstein, S.E. & K.N.Stevens (1979), "Acoustic invariance in speech production", *JASA*, vol. 66, pp. 1001-1017.

[2] Sussman, H. (1991) "The representation of stop consonants in three-dimensional acoustic space", *Phonetica*, vol. 48, pp. 18-31.

[3] Sussman, H., H.A. McCaffrey, & S.A. Matthews (1991), "An investigation of locus equations as a source of relational invariance for stop place categorization", *JASA*, vol. 90, pp. 1309-1325.

[4] Sussman, H., K.A. Hoemeke, & F. S. Ahmed (1993), "A cross-linguistic investigation of locus equations as a phonetic descriptor for place of articulation", *JASA*, vol. 94, pp. 1256-1268.

[5] Lindblom, B. (1963) "On vowel reduction", Speech Transmission Lab. Report #29, The Royal Institute of Technology, Stockholm.

[6] Krull, D. (1988) "Acoustic properties as predictors of perceptual responses", *PERILUS*, vol. VII, pp. 66-70.

[7] Krull, D. (1989) "Second formant locus patterns and coarticulation in spontaneous speech", *PERILUS*, vol. X, pp. 87-108.

[8] Castañeda, M. L. (1986) "El VOT de las oclusivas sordas y sonoras españolas", *Estudios de Fonética Experimental*, vol. II, pp. 91-110.

[9] Núñez-Romero, B. (1995) "Invariación acústica en las oclusivas del español", to appear in *Estudios de Fonética Experimental*, vol. VII.

[10] Villalba, X. (1995) "Los invariantes acusticos y el punto de articulacion de las oclusivas en español", to appear in *Estudios de Fonética Experimental*, vol. VII.

PROCEDURAL INFLUENCES ON THREE MEASURES OF ARTICULATORY CONTROL

Ruth Huntley, Allen Montgomery, Theresa Herron, Heather Clark, Kathy Kugel and Laura Iden
University of South Carolina, Columbia, SC USA

ABSTRACT

This project was designed to study the stability of three measures of articulatory control: a relative timing task, the F2 slope index and locus equations. College-age subjects were asked to read traditional stimulus types at conversational, slow and fast rates at two separate testing sessions. Statistical analyses indicated that a change in rate significantly affected two out of the three measures while the test-retest factor was not significant. However, normalization procedures revealed that rate continued to be a factor in the F2 slope index while exerting little to no influence on the other techniques.

INTRODUCTION

The articulatory process is indeed a complex one that is mediated by both central and peripheral factors. Numerous investigators have attempted to identify the essential components of this process in an attempt to draw conclusions regarding the planning, programming and execution of phonemes in a sequence. Three analysis procedures have appeared in the recent literature on articulatory control that merit further examination because they offer some support to the notion of relational invariance. These techniques include a relative timing task [1,2], the F2 slope index [3,4] and the locus equation [5,6].

Relative timing refers to the inherent temporal organization of phonetic units across the time dimension. Several studies have indicated that relative timing does remain constant throughout changes in speech rate and stress patterns [1,7]. These investigators have concluded that timing characteristics are inherent in the motor programs for speech and not superimposed on the signal at the level of the articulators. The F2 slope index, on the other hand, has been used to demonstrate the importance of formant trajectories in the determination of speech intelligibility [3,8]. It is postulated that a flatter slope than "normal" would be perceived as speech that was less intelligible [4]. Finally, locus equations are used to demonstrate relational invariance (i.e., independent of vowel context) of place of articulation for initial voiced stop consonants [5]. These equations are so robust that they even hold across languages [6].

While all of these measures seem to be well established, there is little information regarding their effectiveness in conditions where the rate of speech is either faster or slower than normal. Since rate has been shown to influence segment duration and formant frequency values [9-11], it seems important to test this factor more carefully. Furthermore, little is known about the stability of such measures across times of testing. Therefore, it is the purpose of this project to assess the influences of rate of speech and test-retest reliability on each of these measures of articulatory control. Finally, as a test of intersubject variability, the investigators normalized the rate and formant values obtained in their measurement procedures. The normalization process should highlight the factors that may be changing as a function of speech rate. Since timing ratios are presumed to reflect the central organizing activity of the articulatory process, they should not be affected by changes in rate. On the other hand, the F2 measures represent movement of the articulators. Therefore, statistical normalization of rate may affect the F2 slope measures while F2 normalization (adjusting for vocal tract length) may influence F2 slope and F2 regression lines.

SUBJECTS

Twenty subjects (10 males and 10 females) participated in this study. They ranged in age from 22-33 years of age and were enrolled at the University of South Carolina. All of the subjects spoke Standard American English. They had no known speech, voice, fluency or hearing disorders and were screened for minimal dialectal variation. None of the speakers had received professional voice training; however, the talkers were given a chance to practice the tasks prior to commencement of the procedure.

PROCEDURE

The voice recordings were made with a Digital AudioTape (DAT) recorder in a sound-treated booth. The subjects were presented with a set of cards containing the stimulus items. Each of the speakers read the cards aloud into the microphone. They were required to produce each stimulus set, starting with the first phrase and continuing on to the last one, before moving on with the next repetition. The subjects read the stimulus sets first at a normal, conversational rate, followed by a slow rate and finally at a fast rate. To achieve a "slow" rate, subjects were asked to speak at what they considered to be half their normal rate of speaking. They were also provided with a spoken model by the experimenter, as well as cues to slow down throughout the procedure. Similarly, to achieve a fast rate, subjects were asked to speak at twice their normal rate, with a model again provided by the experimenter. The entire experimental procedure was repeated two days later as a measure of test-retest reliability.

STIMULUS SETS

They were asked to read the following stimulus types: three repetitions of five sentences constructed for the ratio task [1], three repetitions of a standard phrase with a target word taken from the acoustic signature literature [3] and five repetitions of /b,d,g/ paired with four vowels in a /CVt/ context [5].

DATA ANALYSIS

All measurements were made from computer-generated spectrograms using Sensimetrics SpeechStation (version 3.0) software on an IBM-compatible 486 computer. The Sensimetrics program allows the user to identify various acoustic correlates, such as time (in msecs.) and frequency corresponding to specific points on the stimuli.

Measurement Procedures

For the relative timing ratios, the investigator constructed a set of nine measurements (one durational and four sets of ratio measurements) by which each utterance was analyzed. The ratio measurements were based on constructs developed by Weismer, et al. [1] and Prosek, et al. [2]. The boundaries for each ratio occurred at a vowel-consonant (VC) or consonant-vowel (CV) interface.

The F2 slopes were derived by identifying the last frequency value of F2 before the transition and the value of the first glottal pulse of the leveling off point of the steady state portion. The procedure was somewhat different for "wax" and "blend" in that the glides were also measured as they rose into the vowels. This procedure seemed to include the most frequency change. Once the starting and stopping points were identified, the frequency and millisecond values for each of these points were recorded and a slope was computed by dividing the amount of frequency change by the duration of that change (i.e., rise over run).

The locus equations were generated as follows: measurements were taken at the first glottal impulse of F2 and then during the vowel steady state. Each stimulus item was measured three times at each rate and joined on a scatterplot which represented each consonant by plotting the F2 onset by the F2 steady state frequencies of all four vowels. The regression line, which results from this procedure, is known as the locus equation. From these scatterplots, slopes were derived for intersubject comparisons. Lastly, all measurements were subjected to intra- and inter-judge

reliability testing and all correlations derived from these procedures were excellent ($r > 0.85$).

Normalization Procedures

Rate normalization was carried out on the timing ratios and the F2 slope measurements. The timing ratios were recomputed for the fast and slow rates with corresponding durations from the normal rate serving as the denominator. These new component fractions were used to calculate "normalized" ratios. The intent was to factor out the influence of rate across the ratio portions. If speech rate was a constant factor across the generation of a phrase, the ratios generated for each component should be equal and the ratio should be equal to one. Thus, normalized ratios that deviated from one would indicate differential increase (or decrease) in rate across the sentence.

In adjusting the rate of F2 slopes, the investigator utilized the "normal" duration value as the denominator for all of the slopes. This process would provide information about F2 rise while holding time constant. Now, one could talk about the influences of coarticulation independent of rate.

F2 frequency normalization was also employed as a means of controlling for individual differences in estimated vocal tract length. Therefore, the extent of frequency change noted here could provide insight into the process of coarticulation while controlling for one factor known to vary across individuals.

RESULTS AND DISCUSSION

Two or three-way ANOVAs (*stimuli x* rate x testing time) were conducted on the preliminary data to determine the effects of rate and test-retest reliability on each of these measures of articulatory control. Rate was noted to be a significant factor in two out of three of these measures. However, the test-retest factor was not significant. Therefore, one can conclude that the articulatory process associated with each of these measurements is fairly consistent across times of testing.

Effects of Speaker Rate

The use of a slow rate proved to be troublesome for the relative timing ratios in that three of the four ratios did not remain constant in this condition. Furthermore, one of the sentences generated variable ratios in both the fast and slow conditions for at least two of the ratios. While the sentences were chosen to represent different types of phoneme transitions, this factor did not seem to be the most pertinent. However, the semantic naturalness of the sentence seemed to disrupt relative timing. This factor needs further investigation.

Rate, once again, was a factor in the interpretation of the F2 slopes. The mean slopes for each rate were significantly different from one another. Only the analysis of the words "wax" and "blend" provided different results. In these cases, the mean slope values for the normal and fast rates were not found to be significantly different from one another; however, slope values for the slow rate were distinctive from the other two speeds.

Finally, a change in rate did not affect the derivation of a locus equation for any of the voiced plosives tested. This finding was to be expected since the locus equation is not a time dependent measurement. However, it is interesting to note that the possible target undershoot and overshoot that occurred in the fast and/or slow conditions did not significantly affect the slope of the regression line for /b/, /d/ or /g/. Furthermore, as noted in previous research, the slope of the /d/ regression line was significantly different from the others.

Effects of Normalization Procedures

Rate normalization revealed that the effects of rate on the relative timing ratios are consistent throughout the sentence with mean ratios for both fast and slow conditions approximating 1.0. That is, when differences attributable to talker variation and sentence context were removed, the ratios revealed no differential lengthening or shortening of sentence components on the average. However, specific sentences or ratios showed some modest effects of rate.

As might be expected, rate normalization of the F2 slope reduced differences attributable to talker speed. However, significant differences for the fast and slow rates for certain words still remained indicating the continuing presence of articulatory over/undershoot as a contributory factor.

The F2 frequency normalization procedures are currently underway and no conclusions are available at this time. In general, the three measures employed in the present study appear to be useful tools for assessing articulatory behavior and further refinements and enhancements of these techniques would appear to be justified.

REFERENCES

[1] Weismer, G. & Fennell, A. (1985), "Constancy of (acoustic) relative timing measures in phrase-level utterances," *J. Acoust. Soc. Amer.*, vol. 78, pp. 49-57.

[2] Prosek, R., Montgomery, A. & Walden, B. (1988), "Constancy of relative timing for stutterers and nonstutterers," *J. Sp. Hear. Res.*, vol. 31, pp. 644-658.

[3] Kent, R., Weismer, G., Kent, J. & Rosenbek, J. (1989), "Toward phonetic intelligibility testing in dysarthria," *J. Sp. Hear. Dis.*, vol. 54, pp. 482-499.

[4] Kent, R., Kent, J., Weismer, G., Martin, R., Sufit, R., Brooks, B. and Rosenbek, J. (1989), "Relationships between speech intelligibility and the slope of second-formant transitions in dysarthric subjects," *Clin. Ling. and Phon.*, vol. 3, pp. 347-358.

[5] Sussman, H., McCaffrey, H. & Matthews, S. (1991), "An investigation of locus equations as a source of relational invariance for stop place categorization," *J. Acoust. Soc. Amer.*, vol. 90, pp. 1309-1325.

[6] Sussman, H., Hoemeke, K. & Ahmed, F. (1993), "A cross-linguistic investigation of locus equations as a phonetic descriptor for place of articulation," *J. Acoust. Soc. Amer.*, vol. 94, pp. 1256-1268.

[7] Tuller, B., Kelso, J. & Harris, K. (1983), "Converging evidence for the role of relative timing in speech," *J. Acoust. Soc. Amer.*, vol. 76, pp. 1030-1036.

[8] Weismer, G., Kent, R., Hodge, M. and Martin, R. (1988), "The acoustic signature for intelligibility test words," *J. Acoust. Soc. Amer.*, vol. 84, pp. 1281-1291.

[9] Lindblom, B. (1963), "Spectrographic study of vowel reduction," *J. Acoust. Soc. Amer.*, vol. 35, pp. 1773-1781.

[10] Gay, T. (1968), "Effects of speaking rate on diphthong formant movements," *J. Acoust. Soc. Amer.*, vol. 44, pp. 1570-1573.

[11] Gay, T. (1978), "Effect of speaking rate on vowel formant movements," *J. Acoust. Soc. Amer.*, vol. 63, pp. 223-230.

The Amplitudes of the Peaks in the Spectrum as Acoustic Attributes of the Place of Articulation

Anna Esposito
Massachusetts Institute of Technology 02139 Cambridge (MA) USA
Universitá di Salerno, Dept. Fisica Teorica (SA) Italy

Abstract[1]
This work is devoted to the study of the properties of the sound spectrum at the release of Italian stop consonants in vocalic contexts. The aim is to check if the amplitudes of the peaks in the spectrum can be used as acoustic attributes of the place of articulation of the consonants. Moreover, different measurements have been performed in order to define what of measure retains more information about peak amplitudes.

Materials and procedures

The recording and measurements were made in the Research Laboratory of Electronics, Speech Communication Group, MIT, Cambridge, Usa. The materials consists in VCVC utterances produced by seven adult Italian speakers (three females and four males) in a sound-treated room and recorded on a high-quality magnetic tape recording system. The speakers were selected from different parts of Italy. The utterances are embedded in a carrier phrase: *Prendi VCVC se vuoi (Take VCVC if (you) want)*. The measurements were made for the consonant between vowels. Data have been collected for all the official Italian vowels embedded in stop contexts. However, the results reported in this paper come from the analysis of the stop consonants in the context of the vowel /*i*/. A more detailed description of the speech materials and the procedures can be found in Esposito and Stevens [2]. The spectral representations used include a DFT spectrum, a smoothed version of the DFT, a spectral averaging method. The analysis window (Hamming window) was set to 3.1 msec for each measurement. The spectrum at the release of each consonant, the averaged spectrum during the first 4 msec (for /*b, d, g*/) and 10 msec (for /*p, t, k*/) after the release and the *k*-averaged[2] spectrum were computed. All spectra are preemphasized, i.e. spectra of the first difference of the waveform are calculated. Moreover, the spectral amplitudes were also enhanced by changing an overall spectral gain control parameter. The amplitudes of the maximum peaks in the frequency ranges of 1-3kHz, 3-5kHz, 4-6kHz, 5-7kHz, 0-2kHz, and 0.8-1.5kHz were measured from cursor amplitude readouts via mouse position placed on the spectrum display. The spectrum display shows, superimposed, both the smoothed spectra and the DFT spectrum. However the peak amplitudes were measured only on the DFT spectrum.

The amplitude attributes

The peaks amplitudes measured in the different frequency ranges described above were compared in order to identify properties that can be useful to discriminate the place of articulation of each consonant. Initially averages of the maximum peak amplitudes in different frequency ranges were computed. However, even though some of these averages differ significantly from one consonant to another, the standard deviations were high and they overlapped. This effect is mostly due to the variability of the peak amplitudes among the speakers. For this reason we decided to exclude these measures and we start to look to the amplitudes of the maximum peaks in specified frequency ranges compared to the amplitudes of the maximum peaks in other frequency ranges. This comparison seemed more reasonable to us because it is possible to reduce the amplitude variability among speakers and repetitions. We carried out several attempts, comparing the maximum peak amplitudes in some frequency ranges with the maximum peak amplitudes in some other frequency ranges or comparing the differences between the maximum peak amplitudes in the different frequency ranges examined. In each attempt we defined a set of acoustic features based on these comparisons and tested this set of features on the consonants in order to verify if it accepted the consonant under examination and rejected the others. The final results of this trial and error process are the following set of acoustic attributes for each place of articulation:

Velar amplitude attributes:
a1) The differences between the maximum peak in the 0-2kHz and the maximum peak in the 4-6kHz frequency ranges must be lower than 2dB;
b1) The differences between the maximum peak in the 1-7kHz and the maximum peak in the 0-2kHz frequency ranges must be greater or equal to 9dB;
c1) The differences between the maximum peak in the 3-5kHz and the maximum peak in the 4-6kHz frequency ranges must be greater or equal to 0dB;
d1) The differences between the maximum peak in the 3-5kHz and the maximum peak in the 5-7kHz frequency ranges must be greater or equal to 0dB.

Labial amplitude attributes:
a2) The differences between the maximum peak in the 0-2kHz and the max-

[1]Supported by CNR-IIASS contratto quinquennale and INFM Salerno University. Acknowledgements goes to Maria Marinaro, Carmen D'Apolito and Kenneth N. Stevens for their comments and suggestions.

[2]This spectrum was computed measuring, for each voiceless consonant, the VOT length. Then the cursor was placed on the waveform at the temporal sampling point corresponding to the half of the VOT length and the spectrum averaged on 5 msec to the left and 5 msec to the right of this sampling point was computed. We call this spectrum the *k*-averaged spectrum because *k* is the command to compute it. *k* is a parameter of the analysis tool (Klatt tools [3]) previously set to 150 samples, corresponding to 10 msec of signal duration, at the sampling rate of 16.000 Hz.

Table 1: *Amplitude feature-matching results for velar consonants. The entries give the mean percentage of utterances of each consonant (based on 21 utterances of each consonant, occurring in /i/ vowel environment, and obtained from seven speakers) that were correctly accepted or rejected by the set of acoustic features defined above.*

Spectrum at release	
Correct Acceptance	Correct Rejection
/k/ 57.1	/p/100 /t/95.2
/g/ 90.4	/b/100 /d/85.7
Averaged Spectrum	
Correct Acceptance	Correct Rejection
/k/ 95.2	/p/100 /t/95.2
/g/ 80.9	/b/90.4 /d/76.2
***k*-Averaged Spectrum**	
Correct Acceptance	Correct Rejection
/k/ 95.2	/p/95.2 /t/95.2

Table 2: *Template-matching results obtained using the Blumstein and Stevens compact template.*

Spectrum at release	
Correct Acceptance	Correct Rejection
/k/ 33.3	/p/ 95.2 /t/ 90.4
/g/ 66.6	/b/ 57.1 /d/90.4
Averaged Spectrum	
Correct Acceptance	Correct Rejection
/k/ 47.6	/p/ 61.9 /t/ 85.7
/g/ 52.3	/b/ 47.6 /d/66.6
***k*-Averaged Spectrum**	
Correct Acceptance	Correct Rejection
/k/ 65	/p/ 61.9 /t/ 76.2

Table 3: *Amplitude feature-matching results for labial consonants.*

Spectrum at release	
Correct Acceptance	Correct Rejection
/p/ 33.3	/k/ 95.2 /t/ 100
/b/ 66.6	/g/ 100 /d/95.2
Averaged Spectrum	
Correct Acceptance	Correct Rejection
/p/ 90.4	/k/ 100 /t/ 90.4
/b/ 57.1	/g/ 100 /d/100
***k*-Averaged Spectrum**	
Correct Acceptance	Correct Rejection
/p/ 90.4	/k/ 100 /t/ 100

Table 4: *Template-matching results obtained using the Blumstein and Stevens diffuse-falling template.*

Spectrum at release	
Correct Acceptance	Correct Rejection
/p/ 38.1	/k/ 95.2 /t/ 66.6
/b/ 57.1	/g/ 90.4 /d/85.7
Averaged Spectrum	
Correct Acceptance	Correct Rejection
/p/ 61.9	/k/ 100 /t/ 90.4
/b/ 66.6	/g/ 90.4 /d/80.9
***k*-Averaged Spectrum**	
Correct Acceptance	Correct Rejection
/p/ 52.3	/k/ 100 /t/ 85.7

imum peak in the 4-6kHz frequency ranges must be greater than 1dB;

b2) The differences between the maximum peak in the 1-7kHz and the maximum peak in the 0-2kHz frequency ranges must be lower than 9dB;

c2) The differences between the maximum peak in the 1-3kHz and the maximum peak in the 5-7kHz frequency ranges must be greater than 8dB.

Alveolar amplitude attributes:

a3) The differences between the maximum peak in the 1-3kHz and the maximum peak in the 5-7kHz frequency ranges must be lower than 9dB;

b3) The differences between the maximum peak in the 1-7kHz and the maximum peak in the 0-2kHz frequency ranges must be lower than 10dB;

c3) The differences between the maximum peak in the 3-5kHz and the maximum peak in the 4-6kHz frequency ranges must be lower than 0dB;

d3) The differences between the maximum peak in the 3-5kHz and the maximum peak in the 5-7kHz frequency ranges must be lower than 9dB.

Table 5: *Amplitude feature-matching results for alveolar consonants.*

Spectrum at release	
Correct Acceptance	Correct Rejection
/t/ 90.4	/k/ 80.9 /p/ 33.3
/d/ 80.9	/g/ 100 /b/90.4
Averaged Spectrum	
Correct Acceptance	Correct Rejection
/t/ 85.7	/k/ 95.2 /p/ 90.4
/d/ 66.6	/g/ 90.4 /b/90.4
***k*-Averaged Spectrum**	
Correct Acceptance	Correct Rejection
/t/ 95.2	/k/ 90.4 /p/ 95.2

The set of acoustic attributes defined above are the same both for voiced and voiceless consonants. However, for voiced consonants we have to changes the 0-2kHz and 4-6kHz frequency ranges to 0.8-1.5kHz and 3-5kHz respectively. These frequency changes can be justified considering that in order to allow vocal-fold vibrations during the production of voiced consonants the larynx is lowered, the pharynx is expanded and the walls of the vocal tract are compressed. This could cause small shifts in the vocal tract resonances such as a lowering in frequency.

Results

We used the set of features defined above and the templates defined by Blumstein and Stevens [1] and we tested their discrimination performances. We obtained the results reported in the tables. These preliminary

results show that the amplitudes of the peaks in the spectrum computed during the first 10 msec after the release and in the *k*-averaged spectrum can be used to discriminate among the voiceless consonants /p, t, k/ (see tables 1, 3, 5). What is mostly useful to discriminate /p/ from /k/ is the property *a2* (even though also *b2* plays an important role) whereas *c2* is mostly useful to discriminate /p/ from /t/. The properties that allow to discriminate /k/ from /p/ are *a1, b1, d1*, whereas /k/ is successfully distinguished from /t/ by *b1*. The opposite of *a2 (a3)* is mostly used to discriminate between /t/ and /p/ and the opposite of *b1 (b3)* is used to discriminate between /t/ and /k/. This information can be used to define an automatic algorithm which discriminates successfully among /p, t, k/. Using the Blumstein and Stevens templates on the same data (see tables 2, 4, 6) the discrimination performances are less good in most of the cases. This result is expected in the case of the alveolars because of the different point of constriction of Italian /t, d/ with respect to American /t, d/. However, the results for labials and velars does not seem to be better suggesting some language specific influences on the gross shape of the spectrum.

In the case of voiced consonants, the set of attributes defined above can be used to identify /g/ and to discriminate /g/ from /b, d/ (at the release). However, for /b, d/ similar information does not identify the two consonants, even though they discriminates /b/ from /g, d/ and /d/ from /b, g/. In such cases, information about formant transitions is required. The voicing, which is always present in Italian, causes pressure fluctuations that lead to variability in the peak amplitudes.

With regard to the particular spectra computed it is possible to say that, in the case of voiceless consonants, the better performances of the acoustic attributes defined above and the Blumstein and Stevens template are obtained when the spectra during the first 10 msec after the release and the *k* averaged spectra are used for the comparisons. These spectra seem more useful to retain information about amplitude features. The spectra at the release retain more information about the amplitude attributes of voiced consonants.

Table 6: *Template-matching results obtained using the Blumstein and Stevens diffuse-rising template.*

Spectrum at release	
Correct Acceptance	Correct Rejection
/t/ 42.8	/k/ 47.6 /p/ 38
/d/ 52.3	/g/ 61.9 /b/80.9
Averaged Spectrum	
Correct Acceptance	Correct Rejection
/t/ 71.4	/k/ 71.4 /p/ 85.7
/d/ 42.8	/g/ 57.1 /b/85.7
***k*-Averaged Spectrum**	
Correct Acceptance	Correct Rejection
/t/ 61.9	/k/ 70 /p/ 90.4

These results are restricted to the consonants in the /i/ vowel environment. We will test the set of acoustic attributes defined in this paper to the consonants in the other vowel environments. We expect that there will be changes in their definition in order to improve their performances in the other vowel environments.

References

[1] S.E. Blumstein, K.N. Stevens, 1979, *Acoustic Invariance in Speech Production:...*, JASA, Vol. 64(4), 1001-1017.

[2] A. Esposito, K.N. Stevens, 1994, *Note on Italian Vowels:...*, (in press on MIT Speech Com. Work. Prog.).

[3] D.H Klatt, 1984, *MIT SpeechVax User's Guide*, Copyright 1984 by Dennis H. Klatt.

ON THE PHONETIC INTERPRETATION OF THE YORUBA TONAL SYSTEM

Akin Akinlabi
Rutgers University

Mark Liberman
University of Pennsylvania

ABSTRACT

Yoruba is a tone language, with three lexically contrastive levels H(igh), M(id) and L(ow). Various phonological and phonetic properties of these tones are explained in terms of the view that M is phonologically unspecified, and that adjacent H and L join to form "derived pitch accents."

BACKGROUND

Yoruba has three phonemically distinctive tones—H(igh), M(id), L(ow). H occurs in word-initial position only in (marked) consonant-initial words, which reveal an implicit initial vowel when preceded by another word in a genitive construction. Most nouns and adjectives start with a vowel, which is L or M but not H. Except for these minor tonotactic restrictions, any lexical vowel can have any one of the three tonal specifications. There are no underlying tone glides.

ra H "to vanish"	**ra M** "to rub"	**ra L** "to buy"
ọkọ MH "hoe"	**ọkọ MM** "husband"	**ọkọ ML** "vehicle"
ilu LH "town"	**ilu LM** "opener"	**ilu LL** "drum"
pako HH "plank"	**kese HM** *place-name*	**pako HL** "chewing stick"

Thus Yoruba presents itself as a fundamentally tonal language, in which tonal features have a lexical distribution about as free as that of any other phonological features.

There are several reasons to believe that Yoruba M(id) tone is underlyingly just the absence of tonal features ([1], [12]). We will mention just one of these: tones L and H remain when their lexically-associated vowels delete, but M does not. Thus in the case of a verb followed by a vowel-initial object, one of the two adjacent vowels obligatorily deletes. The tonal consequences are simple to calculate if we assume that M is just the name for lack of tone—then all "real" tones remain stable under vowel deletion.

(1) a. wa (H) + ẹkọ(LH) ⇒
look (for) education
wẹkọ (H LH)
look for education

b. mu (H) + iwe (LH ⇒
take book
muwe (H LH)
take a book

c. jọ (M) + ajẹ (LH) ⇒
resemble witch
jajẹ (LH)
resemble a witch

d. sin (M) + oku (LH) ⇒
bury dead (body)
sinku (LH)
bury the dead

SOURCE OF DATA

There have been several earlier instrumental studies of Yoruba tone (e.g. [3], [14], [6], [4], [5]). In order to apply to Yoruba the scaling technique previously applied to English in [8] and to Igbo in [9], we devised 78 Yoruba phrases exhibiting an appropriate range of tone sequences, with texts that avoid consonants likely to interrupt or strongly affect F0. These phrases were read a total of 18 times each, six in each of three pitch ranges, by three native Yoruba speakers. Pitch range was varied by instructing the speaker, in each utterance, to address one of three (imaginary) interlocutors, placed immediately adjacent to the speaker, across the room, or out the door and down the hall. Within each recording session, the list of phrase/pitch-range combinations was randomized.

In this paper, space does not permit us to report fully on this experiment or to discuss its relation to previous work. Instead, we will focus on three key points and a general conclusion that is suggested by them.

TIMING OF TONE GLIDES

As was first noted in [15], Yoruba HL and LH sequences postpone the falling or rising F0 glide to the second syllable. By comparison, the transitional glide for sequences involving M (HM, ML, MH, LM) occurs significantly earlier.

Figure 1 shows the F0 tracks for the initial LH sequences in the six narrow-pitch-range repetitions of **ọ̀rúnlàmí lèmi** "I am Orunlami"[1] as produced by one male speaker.

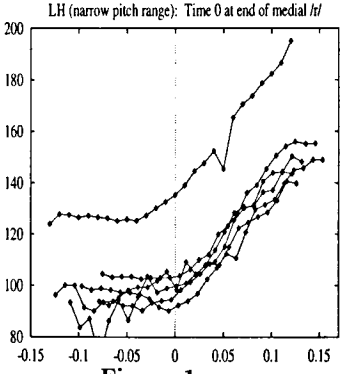


Figure 1

Figure 2 shows the initial LM sequences in six wide-pitch-range repetitions of **ọ̀runlèyẹ lèmi** "I am Orunleye," produced by the same speaker as in Figure 1. In both Figure 1 and Figure 2, the x-axis presents time in seconds, with zero set at the opening of the **/r/** in each utterance, while the y-axis shows F0 in Hz.

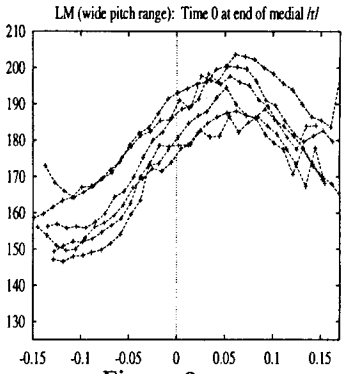


Figure 2

This postponement of the HL and LH glides in Yoruba is a well-established fact. Our contribution is to compare it systematically to HM, ML, MH, LM glides across pitch ranges, and to verify the stability of the difference under this manipulation.

TONE DISSIMILATION

In Yoruba, H is raised before L (as opposed to before M), and L is lowered before H (as opposed to before M). The raising of H before L has been cited in [2], [4], [5]. The lowering of L before H is (we believe) a new observation.

Table 1 exemplifies the raising of H before L, by showing the means and standard errors of peak F0 measurements in each of the three pitch range conditions, for one of three subjects.

	narrow	middle	wide
mean H/_ML	125	154	250
standard error	3.2	8.8	2.6
mean H/_LM	143	182	287
standard error	3.3	7.0	2.9

Table 1

For evidence of the lowering of L before H as opposed to before M, see Figure 3, which plots the relationship of successive L tones in the sequences HLHLM (plotted with squares) and HLHLH (plotted with pluses).

When both L tones are followed by H, the second L tone is considerably

[1] All examples in this paper use the standard Yoruba tone marking convention, according to which H is marked with an acute accent, L with a grave accent, and M with no accent. Vowels with a dot underneath are non-ATR.

lower than the first. This is consistent with the general expectation of downdrift in such sequences; dissimilatory lowering applies to both L tones in this case. When the first L is followed by an H, while the second L is followed by an M, the expected downdrift effect is almost completely nullified. This is because the first L is lowered because it is in an LH sequence, while the second L does not experience this effect. Thus dissimilatory lowering of the first L, and downdrift lowering of the second L, leave them at about the same level.

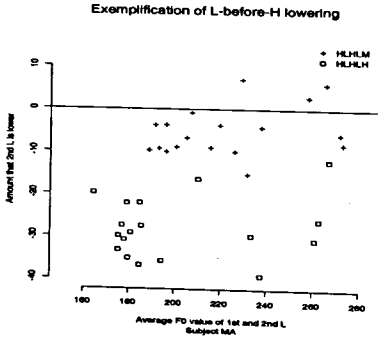


Figure 3

DOWNDRIFT

Since [16] and [13] it has been understood that the tendency of pitch to fall in the course of phrases in tone language like Yoruba is not a sort of phrasal wave on which tonal ripples ride, but rather is connected specifically with alternating high and low tones. Sequences of like tones, especially H tone sequences, remain more or less level.

Since [2], it has been known that in Yoruba, this *downdrift* does not extend to sequences in which H or L alternate with M (HMHM...or MLML...). At least, the amount of downtrend is much lower in these latter cases.

For a quantitative picture of the difference between the amount of lowering in HLHL vs. HMHM or MLML, see Figure 4. Here we show the relationship between adjacent F0 maxima in the sequences HLHL (plotted with squares), HMHM (plotted with pluses) and MLML (plotted with triangles). The x-axis gives the average height of the two F0 maxima, while the y-axis shows the difference between them. In the case of HLHL, this difference is about 15–35 Hz., showing a healthy amount of downdrift, the tonally-conditioned effect that has been called "catathesis." In the case of HMHM and MLML, the difference is about 5–15 Hz., perhaps reflecting the more general downtrend sometimes distinguished as "declination."

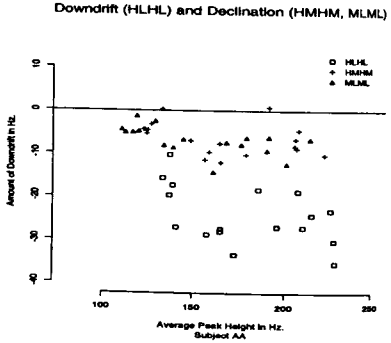


Figure 4

PHONOLOGICAL REPRESENTATION

The postponement of HL and LH tone glides in Yoruba has been treated as a case of tone spreading: an H or L associated with a given syllable comes to be simultaneously associated with the following syllable iff the following syllable bears the opposite tone. This account offers no motivation for the circumstances of the spreading: why should a tone spread to the following syllable if and only if the following syllable is already specified for (a different) tone?

We suggest that the motivation is simple. HL and LH tone glides are phonologically and phonetically natural entities, just as CV syllables are. Yoruba tone spreading is exactly analogous to the re-syllabification of a syllable-final consonant to fill the empty onset of a following onsetless syllable. Thus Yoruba tone spreading is a natural process because it forms cognitively favored structures. There are various plausible ways to express this notion formally, for which space is lacking here.

This idea says, in effect, that Yoruba forms "derived pitch accents" out of adjacent HL and LH tones. Although this does not in itself explain tonal dissimilation and the special status of HL or LH sequences in downdrift, it suggests a direction of research by connecting them to comparable phenomena in Japanese and other languages.

Japanese accent is interpreted (e.g. by [10] as a lexically-specified HL sequence that functions as a unit. *Catathesis* is triggered only by accents in Japanese according to [11]. Each accentless "minor phrase" has an (ungrouped) H and L tone pair, but accentless sequences rise and fall with only a small amount of *declination*. Japanese accentual H is higher than non-accentual H ("accentual boost") according to [7], even though Japanese accent is not stress-like, does not cause greater segment durations, and is not considered a strong position for alignment with music.

REFERENCES

[1] A. Akinlabi. *Tonal Underspecification and Yorùbá Tones.* PhD thesis, University of Ibadan, Nigeria, 1985.

[2] A. Akinlabi and Y. Laniran. Tone and intonation in Yorùbá declarative sentences, 1987.

[3] J. Carnochan. Pitch, tone and intonation in Yoruba. In D. A. et al., editor, *In honour of Daniel Jones*, pages 397–406. Longmans, 1964.

[4] B. Connell and D. R. Ladd. Aspects of pitch realization in Yoruba. *Phonology*, 7:1–30, 1990.

[5] Y. Qlabisi Laniran. *Intonation in Tone Languages: The Phonetic Implementation of Tones in Yorùbá.* PhD thesis, Cornell, January 1992.

[6] J.-M. Hombert. The perception of bisyllabic nouns in Yoruba. *Studies in African Linguistics*, Supplement 6:109–121, 1976.

[7] H. Kubozono. *The Organization of Japanese Prosody.* Kurosio Publishers, Tokyo, 1993.

[8] M. Liberman and J. Pierrehumbert. Intonational invariance under changes in pitch range and length. In M. Aronoff and R. Oehrle, editors, *Language Sound Structure*, pages 157–234. MIT Press, 1984.

[9] M. Liberman, J. M. Schultz, S. Hong, and V. Okeke. The phonetic interpretation of tone in Igbo. *Phonetica*, 50(3):147–160, 1993.

[10] J. B. Pierrehumbert and M. E. Beckman. *Japanese Tone Structure.* MIT Press, Cambridge, MA, 1989.

[11] W. J. Poser. *The phonetics and phonology of tone and intonation in Japanese.* PhD thesis, Massachusetts Institute of Technology, 1984.

[12] D. Pulleyblank. *Tone in Lexical Phonology.* Reidel, Dordrecht, 1986.

[13] J. M. Stewart. Niger-Congo: Kwa. In T. A. Sebeok, editor, *Current Trends in Linguistics 7: Linguistics in Sub-Saharan Africa.* Mouton, The Hague, 1971.

[14] C. L. Velle. An experimental study of Yoruba tone. *Studies in African Linguistics*, Supplement 5:185–194, 1974.

[15] I. Ward. *An introduction to the Yoruba language.* W. Heffer and Sons Ltd., Cambridge, 1952.

[16] W. E. Welmers. Tonemics, morphotonemics, and tonal morphemes. *General Linguistics*, 4:1–9, 1959.

A REPRESENTATIONAL BASIS FOR MODELLING ENGLISH VOWEL DURATION

G. N. Clements, CNRS, UA 1027, Paris
Susan R. Hertz, Eloquent Technology, Inc. and Cornell University, Ithaca, N.Y.
Bertrand Lauret, Université de Paris III, UA 1027, Paris

ABSTRACT

This paper proposes a representational basis for modelling the durational behavior of syllable nuclei in General American English. It examines two lengthening patterns, one in which all portions of the nucleus are affected uniformly, and another in which primarily the beginning and end portions are affected. On the basis of this distinction, a classification of nuclei into one-phone vs. two-phone nuclei is proposed.

1. AN INTEGRATED REPRESENTATIONAL BASIS FOR PHONOLOGY AND PHONETICS

Our aim in this paper is to present the broad outlines of a working model of the phonology-phonetics interface, with an illustration from certain facts of General American English. Our approach is based on the premise that phonetics should be viewed as an essential component of the theory of grammar, and that as such, it can be studied in terms of much the same type of theoretical modelling that we find elsewhere in linguistic theory. In this view, which received a preliminary formulation in Clements and Hertz [1], the phonetic component of a grammar does not consist of descriptions of physical patterns as such, but involves a symbolic representational system defined at a level of some abstraction from physical data.

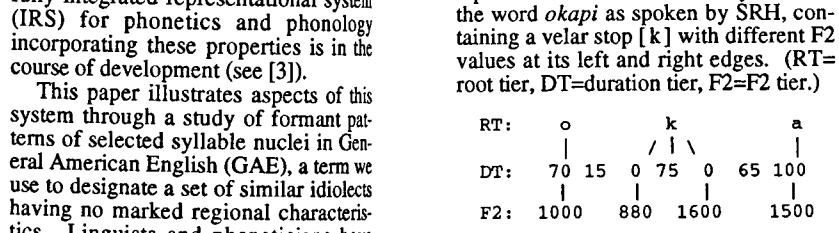
Specifically, we propose that the categorial feature representations of the phonological level are projected directly into the acoustic phonetic level, where they provide the basis for specifying acoustic parameter values in terms of which speech output can be accurately modelled. Like phonological representation, acoustic representation involves partially-specified, multitiered arrays of units related by often complex patterns of association. Acoustic representation differs from phonological representation primarily in introducing new acoustic and duration tiers, required to account for language- and speaker-specific regularities in the acoustic output. By allowing acoustic units to be only partially specified, we allow rising and falling ramps between extrema to be modelled in terms of a target-and-interpolation model [2], while the use of multiple tiers allows for the description of regular patterns of overlap within and across segments. A fully integrated representational system (IRS) for phonetics and phonology incorporating these properties is in the course of development (see [3]).

This paper illustrates aspects of this system through a study of formant patterns of selected syllable nuclei in General American English (GAE), a term we use to designate a set of similar idiolects having no marked regional characteristics. Linguists and phoneticians have long disagreed on the classification of the long gliding vowels of words like *beat, boot, bait,* and *boat*, some treating them as a single segment and others as two. Researchers have also disagreed as to whether the liquids in words like *belt* and *Bart* should be treated as part of the syllable nucleus, or assigned to the margin.

We address these questions within the framework of the integrated approach to phonological and phonetic analysis just outlined. One component of this approach is the phone-and-transition segmentation strategy outlined by Hertz [2]. This strategy is based on the view that speech sounds ("phones") are not necessarily adjacent to each other in phonetic representations, but may be separated from each other by time intervals ("transitions") during which the articulators (lips and tongue) move from the target configuration appropriate for one sound to that appropriate for the next. Phones appear on spectrograms as the time intervals that correspond to such target configurations, while transitions are the time intervals that connect them.

Following these assumptions, we may represent the acoustic structure of an utterance as follows. The root nodes of the phonological representation constitute the *phone tier* of the acoustic representation. Root nodes dominate duration values on a *duration tier*, whose function is to assign each phone a certain duration. Between any two root nodes having different oral tract places of articulation we introduce *transitions*, formally represented as duration values unlinked to root nodes. Duration values dominate appropriate acoustic parameter values on further tiers (F0, F1, F2, aspiration, voicing, etc.). These values can serve as a basis for interpolation across segments unspecified for these parameters.

We illustrate this model with a partial representation of the first two syllables of the word *okapi* as spoken by SRH, containing a velar stop [k] with different F2 values at its left and right edges. (RT= root tier, DT=duration tier, F2=F2 tier.)



This graph represents a pattern with (i) a 70-msec F2 steady state at 1000 Hz characterizing the [o], (ii) a 15-msec transition to the [k] during which F2 falls continuously to a target value of 880 Hz, (iii) a 75-msec period of silence during the [k], (iv) a 65-msec transition to the [a] during which F2 falls from 1600 Hz to 1500 Hz, and (v) a 100-msec F2 steady state at 1500 Hz characterizing the [a]. This representation treats this [k] as a "contour phone", analogous to the contour segments of phonology.

2. DURATIONAL ASPECTS OF ENGLISH SYLLABLE NUCLEI

With this background, we report on a preliminary study of a variety of syllable nucleus types in GAE. Our goal is to find out whether their durational behavior can help us decide whether a given nucleus consists of a single unit or two.

It is well known that GAE syllable nuclei are often lengthened before voiced obstruents, especially phrase-finally [5]. What is less well understood is whether all nuclei lengthen in a uniform manner, or whether they show different patterns of lengthening. Our hypothesis is that if GAE contains a distinction between one-segment nuclei and two-segment nuclei at the phonological level, this distinction might be reflected in different patterns of lengthening at the phonetic level. Such differences, if they exist, might help to answer the questions concerning the analysis of GAE nuclei raised above.

To test this hypothesis, we collected data on four sets of paired monosyllabic words differing only in the voicing of the final consonant: *bit/bid, bait/bade, bite/bide, felt/felled.* One female and three male speakers of GAE were recorded; we report on data from one of the latter (GNC) here. Each test word was pronounced in the frame *say ___ for me*, and the full sequence was repeated ten times. Recordings were digitized at 16 kHz and analyzed by means of the CSRE formant tracker. Aberrant values were discarded, accounting for the occasional gaps in the formant tracks. Segmentation was carried out mainly on the basis of second formant (F2) tracks, since we have found these to provide the most consistent basis for analyzing the temporal properties of vowels and diphthongs.

Representative F2 tracks are displayed in Figure 1. All graphs are reproduced at the same scale. Each one displays an overlay of F2 tracks extracted from the first five tokens of each word. The ordinate represents formant frequency in Hz, and the units of the abscissa represent 8-msec time intervals. Overall, we see that all items were produced with considerable consistency from token to token.

All pairs in Figure 1 have rising or falling F2 ramps, showing that they are phonetic diphthongs. The diphthongal nature of the nucleus of *bit/bid* for this speaker is confirmed by the fact that the F1 track (not shown here) rises as F2 falls, showing that the nucleus ends in a central offglide. Although discussed in [3], the difference in formant values at the beginnings and ends of the nuclei in these words cannot be attributed to coarticulation with the neighboring consonants.

All pairs of nuclei in Figure 1 exhibit F2 lengthening in their second member. However, the first and second columns show distinct patterns. The F2 tracks in the first column resemble a straight line, with some examples showing a sharp drop at the very end. Disregarding these drops, the tracks can be modelled to a fair approximation by positing two duration-less target points at their beginnings and ends, and perfoming a straight-line inter-

polation between them. For these diphthongs, lengthening does not change the overall shape of the F2 track, although its slope is somewhat reduced in the lengthened form.

The diphthongs in the second column display a different F2 lengthening pattern. For these diphthongs, lengthening visibly changes the shape of the F2 track, especially at the ends. Comparing *bite* and *bide*, we see that the durationless initial target of *bite* is replaced by a quasi-plateau some 80 msec long in *bide*; its final steady state is also somewhat longer. In contrast, the F2 transition between the initial and final extrema has about the same duration in both cases (the sharper rise in *bite* can be attributed to its higher final target value). Similar remarks hold of the second pair. The durationless initial target of *felt* is replaced in *felled* by a steady state approximately 30 msec long, and the final portion expands similarly (in three tokens, final F2 values were too low in amplitude to be read). The transitions between these relatively stable portions have about the same duration in both cases.

3. DISCUSSION

We propose that these two patterns can be analyzed as one-segment and two-segment diphthongs, respectively. Note first that the nucleus of *bit*/*bid* is uncontroversially a single vowel at the phonological level, while that of *felt*/*felled* just as clearly constitutes a two-segment sequence. We can explain the fact that the nucleus of *bait*/*bade* patterns with that of *bit*/*bid* by treating them both as one-segment nuclei, and the fact that *bite*/*bide* patterns with *felt*/*felled* by treating both as two-segment nuclei. In addition, if we considered the first-column nuclei to have a two-segment structure, we would have to allow that their first segments are durationless even in lengthening contexts, contrary to our observations elsewhere. Further arguments in support of this analysis are given in [3].

Using the representational system outlined earlier, we can interpret the nuclei [ɪ] and [e] as single phones (= root nodes), even though they show different F2 target values at their left and right edges. Their analysis parallels that of the "contour" phone [k] in *okapi*, observed earlier. Typical values for the [e] of *bade*, for instance, are shown below:

```
RT:        e
          / | \
DT:     0 140 0
        |     |
F2:  1600   1850
```

The representation of [ɪ] differs from that of [e] both in its choice of F2 values and in the fact that its root node is linked to one, instead of two positions on the phonological skeleton (not shown here).

The nuclei [ay] and [ɛl], in contrast, are analyzed as phone sequences, as shown below for the [ay] of *bide*:

```
RT:     a      y
        |      |
DT:    80  70  20
        |      |
F2:  1030    1670
```

Given these analyses, we may state the following generalization: lengthening before voiced stops affects all phones within the syllable nucleus, but affects the transitions between them little, if at all (see [2], [3], [4] for fuller discussion). We can directly explain the fact that [l] lengthens in *felled* by considering that it, too, belongs to the syllable nucleus.

These preliminary observations are offered in illustration of our approach to the study of the phonetics/ phonology interface. Our current project is to examine a fuller set of data, involving further nucleus types, more contexts, and other speakers, in order to determine the generality of these observations, and refine and improve them as necessary.

ACKNOWLEDGEMENT

This work has been supported in part by grant NIDCD R44 DC00758 to Eloquent Technology, Inc..

SELECTED REFERENCES

[1] Clements, G.N. and S.R. Hertz (1991) "Nonlinear Phonology and Acoustic Interpretation", *Proc. of the 12th Int. Congress of Phonetic Sciences*, Aix-en-Provence, vol. 1, 364-73.

[2] Hertz, S.R. (1991) "Streams, Phones, and Transitions: Toward a New Phonological and Phonetic Model of Formant Timing," *J. of Phonetics* 19, 91-109.

[3] Clements, G.N. and S.R. Hertz (1995) "An Integrated Model of Phonetic Representation in Grammar," ms.

[4] Hertz, S.R. and M. Huffman (1992) "A Nucleus-based Timing Model Applied to Multi-dialect Speech Synthesis by Rule" in J.J. Ohala et al., eds., *Proc. of ICSLP 92*, Edmonton, vol. 2, 1171-4.

[5] Chen, M. 1970. "Vowel Length Variation as a Function of the Voicing of the Consonant Environment," *Phonetica* 22, 129-59.

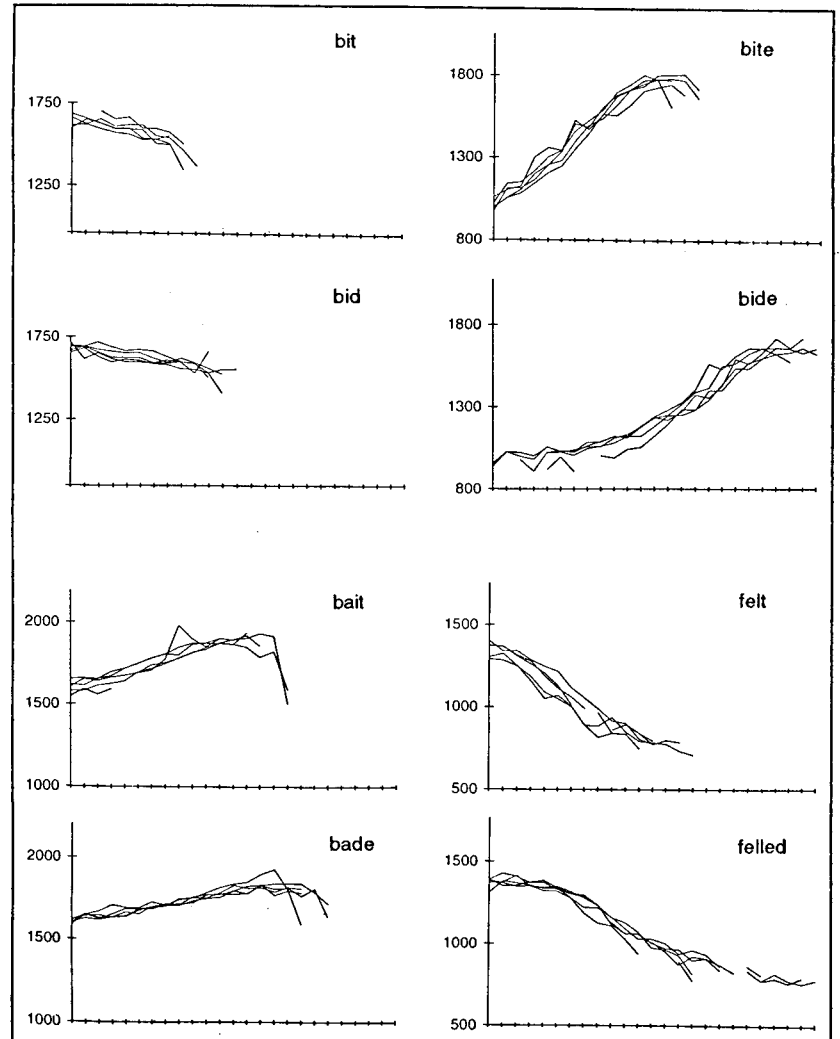


Figure 1.
Lengthening patterns of four types of GAE syllable nuclei, spoken in the context *say __ for me*. Graphs show overlaid F2 tracks of five typical tokens for one speaker.

THE INTERRELATIONSHIP BETWEEN VOT AND VOICE ASSIMILATION

Daan Wissing[1] and Justus Roux
Research Unit for Experimental Phonology (RUEPUS), Private Bag X5108, 7599 Stellenbosch, South Africa.

ABSTRACT

The main aim of this contribution is to enlighten the phenomenon of regressive assimilation of voicing (RVA), referring especially to Afrikaans and Tswana Afrikaans. The strong interrelationship between negative VOT values on the one hand, and RVA on the other hand is being pointed out, using production experiments. In conclusion, we speculate on the possible explanations of the results and the theoretical implications thereof. We suggest a clear interrelationship between hard core phonetic and abstract phonological considerations.

INTRODUCTION

The phenomenon of assimilation of voicing has been researched extensively in the case of the Germanic languages Dutch [1, 2] and Afrikaans [3,4]. Few cross-linguistic studies (with the exception of Van Dommelen [5] and Elshout [6] in the case of Dutch and German) has been done, and not much is known about the possible phonetic causes either. Van Dommelen [5] implies a causal role of negative voice onset time (-VOT) in the case of regressive voicing assimilation (RVA) in Dutch. In this contribution we will test this hypothesis cross-linguistically, and present preliminary results which could serve as a starting point to fill this gap in our knowledge. We will concentrate on Afrikaans L1 as well as Afrikaans L2 (of Tswana speaking persons, 'n Bantu language of Southern Africa). We will also refer to a variety of (mainly) European languages.

Assimilation of the type given in (1) is very common in many languages:

(1) Type 1 languages:
Afrikaans: o/pd/aag → [bd]aag
Dutch: a/sb/ak → a[zb]ak

Other languages include, Hungarian Russian French, and Spanish.[2]

When an underlying voiceless consonant precedes a voiced consonant, the first (C_1) assimilates as to voicedness to the second (C_2). This is called regressive assimilation of voice, or regressive voicing assimilation (henceforth RVA). Both English and German are absent from this list of languages (cf. Roach [7] for English, and Elshout [6] for German). Instead, these languages both prefer progressive voicing assimilation (PVA)[3], as in (2):

(2) Type 2 languages:
English: it i/z/ → it[s]
German: au/f d/em → au[ft]em

It is a puzzling fact that languages such as Afrikaans and Dutch are being characterized by the existence of RVA of Type 1, but that this type of assimilation is absent in the related English and German. In Figure 1 we present some measurements of VOT in a variety of languages, serving as a starting point for discussing the possible relationship between the presence of negative VOT values and RVA in any given language. RVA is known to exist in the cases of Afrikaans, Dutch, Russian, Spanish, but not in German or English. It was tested in this study for Tswana Afrikaans.

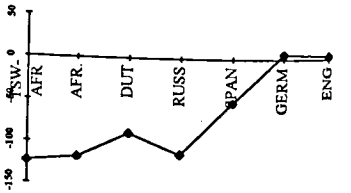


Figure 1. Comparison of VOT values (neg. and pos. - indicated on Y-axis) for voiced plosives in languages exhibiting RVA, i.e. Tswana Afrikaans, Afrikaans, Dutch, Russian and Spanish. In English and German RVA is absent. The precise values are (from left to right: -123ms, -117ms, -88ms, -112 ms, -51ms, and German +6ms, English +4ms. Measurements are of one L1 speaker per language. (n=10+).

The voiced consonants of German [5] and English [7] are strictly speaking not voiced, but rather lenis (tense), in contrast with voiceless consonants, which are fortis (lax), e.g. [p t k].

In Afrikaans, RVA and PVA (progressive voicing assimilation), sometimes alternate in a given word (e.g. se[zd]e / se[st]e - "sesde" (*sixth*)). Voice assimilation thus is an optional process in Afrikaans, unlike the situation in Russian, where it is an obligatory process. This, together with the above-mentioned hypothesis, leads to the expectation that, even in the case of Type 1 languages, the -VOT has to be of a certain minimum value in order to enforce RVA, otherwise either no assimilation at all will take place, or C_2 would be weakened with respect to voicedness to such an extent that RVA would be impossible. This was tested in the following experiments.

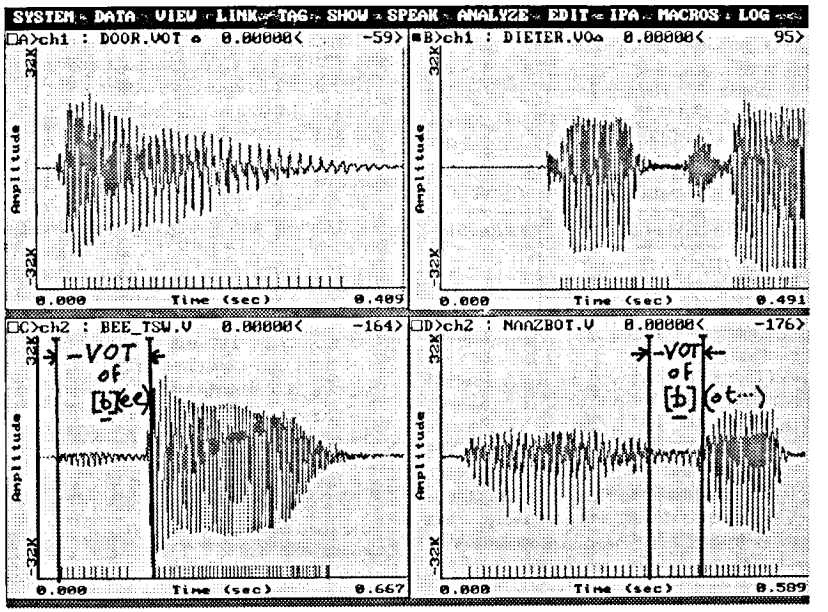


Figure 1: Waveforms of typical examples of initial voiced (lenis) consonants of British English ("door" in Window A) and German ("Dieter" in B) with no negative VOT values, compared to languages with large -VOT values: Tswana Afrikaans ("bee" in C) and L1 Afrikaans ("Naa[zb]otha" in D).

THE CASE OF AFRIKAANS

One speaker of Afrikaans, known from a previous experiment [3] to exhibit RVA, was used. The phrase *Naas Botha* (name and surname) was read 50 times at a fast but comfortable rate. The words *sesde*, *elfde* and *liefde* were read ten times each. In all of these words both RVA and PVA are possible, though PVA would be expected to occur rarely in *Naas Botha*, and more readily in the former three derivatives [3]. All instances of the two types of assimilation were registered. *Naas Botha* were pronounced either as *Naa[zb]otha* (RVA) or as *Naa*[sb]*otha* (no assimilation), but never as *Naa*[sp]*otha* (PVA). See Figure 2 (D) for an example of RVA. The words *sesde*, *elfde* and *liefde* exhibited either PVA or RVA, or no assimilation at all. C_2 consonants were isolated and their -VOT durations measured, using the CSL speech editing system of KAY.

Results

Table 2. Values in milliseconds (ms) of negative VOT (-VOT) of C_2 ([b]) in the Afrikaans phrase **Naas Botha**. *Ranges of values are also mentioned in brackets.*

(One L1 Afrikaans speaker)	C_2 (-VOT)
RVA (C_1=[z]) (33 cases)	80 (40-122)
No ass. (C_1=[s]) (17 cases)	48 (15-83)

There is a statistically highly significant difference (p= 0,000003) between the -VOT values involved in RVA (80 ms) and those of C_2's not involved (48ms).

Table 3. Values in ms of negative VOT of C_2 ([d]) in Afrikaans **sesde, elfde** *and* **liefde**. *Ranges of values are also mentioned in brackets.*

(One L1 Afrikaans speaker)	C_2 (-VOT)
RVA (C_1=[z]) (14 cases)	52 (0-75)
No ass. (C_1=[s]) (2 cases)	33 (16-49)
PVA (C2=[t] (10 cases)	7 (0-65)

There is a statistically highly significant difference (p = 0.007) between the -VOT values involved in RVA (60ms) and those of C_2's not involved (33ms). In two of the RVA cases a VOT value of 0 was found, the rest alternated between 45 and 75ms (see Discussion). There is a definite tendency for the -VOT's to shorten progressively from PVA through No Assimilation to RVA. This also goes for the durations of C_1, but in reversed order: ([s] in PVA = 63ms and in No Assimilation = 60ms; in RVA [z] = 51ms). Both these tendencies are consistent with results of experiments on RVA so far [3]. For the significance of the results in broader perspective, see Discussion.

THE CASE OF TSWANA AFRIKAANS

Four male Tswana speakers, all of whom were competent speakers of Afrikaans (their second or third language), were asked to repeatedly read a few sentences, among others, containing the phrases *ek dink* ("I think") and *mos dat* ("certainly that"), at a comfortable rate. Out of the 320 possible instances of RVA, the subjects assimilated 247 times (either *e[**gd**]ink* or *mo[**zd**]at*). 42 phrases were not taken into account, due to mispronunciations, yielding only 31 instances which were in fact not assimilated regressively (i.e. a mean of 89% RVA's, ranging from 77% to 95% for the four subjects). When taking into account the fact that Tswana is characterized by strong -VOT values (123ms in the case of the above-mentioned subject - see also Figure 1 (C) for a waveform example), these results clearly support the hypothesis postulated. (See next section for discussion)

DISCUSSION

It is quite clear that there is a direct relation between the presence of negative VOT's in C_2 consonants and the appearance of regressive voicing assimilation in the languages studied or referred to in this contribution. The question is, however, what kind of a relationship this is. More specifically: are large -VOT values a prerequisite for RVA to surface in any given language? The strong presence of RVA in Tswana Afrikaans certainly is an indication that this is the case, especially when taking into consideration the fact that RVA is not a possibility in Tswana itself, because of the total non-existence of the relevant of C_1+C_2 combinations - Tswana has mainly a CVCV syllable structure. The statistically significant difference between the -VOT values in the case of "Naas Botha" involved in RVA (52ms) and those of C_2's not involved (33ms) in the case of the first experiment on Afrikaans(see Table 2) strongly supports this hypothesis. The same goes for the second Afrikaans experiment (Table 3). The magnitude of -VOT's in the latter instance (52ms for RVA, 33ms for No Assimilation, and 7ms for PVA) surely highlights the plausibility of the hypothesis that the low-level phonetic VOT values does indeed interrelate with the presence or absence of voice assimilation in the languages under consideration (and perhaps in any given language). On the other hand, the presence of two 0ms values involved in RVA (see Table 3 and accompanying text) suggests that this explanation cannot be absolute. A possible explanation lies on a mental level. Speakers of languages characterized by large negative VOT values sometimes might not actually produce voiced plosives distinguished by -VOT's large enough to further RVA, but such speakers might nevertheless be directed by the tacit knowledge of, in this case, the presence such large negative VOT's typically of their language, be it actualized in particular instances or not. Such an explication implies a clear interrelationship between hard core phonetic and abstract phonological considerations. This, however, is merely a suggestion, which has to be followed up. The optionality of this phonological process in languages such as Dutch and Afrikaans, as well as the fact that males are more inclined to assimilate regressively that females [1] have to be accounted for in subsequent studies of this nature.

As to the presence of RVA in Tswana Afrikaans, this cannot be explained in terms of Natural Phonology (as was suggested by some e-mail reactions per Linguist List). According to Stampe [8] processes such as voicing assimilation are found in the speech of children universally and have to be unlearned for those languages which violate them. More in particular, Natural Phonology (NP) posits a set of innate, vocal tract physiology-driven phonological natural processes, of which RVA would be one. However, vocal tract physiology has to be (very much) the same for all humans, so that it is highly unlikely that the vocal tracts of German and Dutch speakers, for instance, would differ to such an extent that RVA will be present in the latter language but absent in the former.

Neither can NP explain the alternation of RVA and PVS (and, in fact, no assimilation) as is the case in the Afrikaans words **sesde, elfde** and **liefde** (see Table 3).

According to Universal Grammar [9], each language may fix the VOT parameter differently. UG, however, was, up till now, restricted to syntax. Another possibility is that large negative VOT values simply evolve in certain languages but not in others, due to unknown factors.

REFERENCES

[1] Slis, I. (1986), "Assimilation of voice in Dutch as a function of stress, word boundaries, and sex of speaker and listener", *Journal of Phonetics*, vol 14(2), pp. 311-326.

[2] Trommelen, M. & W. Zonneveld (1979), *Inleiding in de generatieve fonologie*, Muiderberg: Coutinho.

[3] Wissing, Daan. (1991), "Regressiewe stemassimilasie in Afrikaans", *South African Journal of Linguistics*, Supplement 11, pp. 132-156.

[4] Wissing, Daan. (1990), "Progressiewe stemassimilasie - 'n "nuwe" Afrikaanse fonologiese reël?", *South African Journal of Linguistics*, vol. 8(2) pp. 88-97.

[5] Van Dommelen, W.A. (1983), "Some observations on assimilation of voicing in German and Dutch", in M. Van den Broecke et al. *Sound Structures*, Dordrecht: Foris Publications.

[6] Elshout, M. (1983), *Assimilationserscheinungen: eine kontrastive Analyse zur Assimilation in Niederländischen und im Deutschen*. Utrecht.

[7] Roach, Peter. (1991), *English Phonetics and Phonology*, Cambridge: Cambridge University Press.

[8] Stampe, D. (1979), *A dissertation on Natural Phonology*.

[9] Chomsky, N. (1986), *Knowledge of Language: Its Nature, Origin, and Use*, Praeger: New York.

---

[1] The first author is also affiliated to the Dept. of General Linguistics, Univ. of Potchefstroom, S.A., and the second author to the Dept. of African Languages, Univ. of Stellenbosch, S.A.

[2] Confirmed by different e-mail users of Linguist List.

[3] But see Van Dommelen [5], who argues in favour of the term **fortition** instead of **devoicing** of C_2

PALATALITY AS A PROSODY IN TUNDRA NENETS

Richard Ogden
Dept. of Language & Linguistic Science,
University of York,
Heslington, York YO1 5DD, UK.
rao1@york.ac.uk

ABSTRACT

This paper describes some of the phonetic characteristics of palatalisation and non-palatalisation in Tundra Nenets. I argue that palatalisation is best treated as a prosodic property, with implications for place and manner of articulation, manner of release of secondary stricture, and tongue body shapes. The categories y and w are set up as terms contrastive over CV structures, and exponents are stated for them in the manner of the Firthians [1, 2, 3].

INTRODUCTION

Tundra Nenets is a language of the Nenets (formerly known as Yurak) subbranch of the Samoyed branch of the Uralic family, and is spoken by approx. 25,000 people in an area of tundra in Arctic Russia and Siberia. There are three dialect groups, of which the Eastern one is exemplified here.

The material presented in this paper was collected from Anastasia Lapsui, a Nenets woman who comes from Nyda in Yamal Nenets district, part of the Russian federation. She works as a foreign correspondent in Helsinki.

TRADITIONAL ACCOUNTS OF PALATALISATION IN NENETS

There are essentially two accounts of palatalisation in Nenets. The first one, typified by Décsy [4], treats palatalisation as a property of consonants: relevant consonants have palatalised and non-palatalised forms. Décsy sets up a system of seven vowels, of which /i e ɑ o u/ occur after palatalised consonants and /i e ɑ ə o u/ after non-palatalised consonants.

In the other account of palatalisation, adopted by Collinder [5], two groups of vowels are set up, one of which invokes palatalisation. Palatalisation under this analysis is an allophonic property of consonants in conjunction with any one from a set of five vowels.

PALATALISATION AS A SYLLABIC PROPERTY

The traditional descriptions of Tundra Nenets contain the following weaknesses:

- they do not give any detail about what the phonetic implications of 'non-palatalised' consonants are;
- they do not describe in any detail the relationship between the presence or absence of palatalisation and the accompanying off-glide as the secondary articulation is released.
- they arbitrarily assign palatality as a property of consonants or vowels.

Table 1 contains some impressionistic records, which were checked for accuracy by analysis of the speech with a sound spectrograph. Particular attention is drawn to the resonance properties of the syllables, and the ways in which portions of consonantal constriction are connected to vocalic portions. The records were not made with the intention of recording stress, intonation nor duration in any detail.

Syllables can be described as overall more front or more back; and sometimes the frontness or backness is dynamic rather than static, giving rise to vocalic portions of changing qualities on the front-back dimension (exx. 4, 9, 13). Note also that fronter syllables frequently contain consonants written with palatal symbols, such as [c ɲ ʎ j] (exx. 9, 16, 18), which imply articulations made with the tongue blade rather than the tongue tip. Fronter syllables also typically contain closer vocalic portions than backer ones.

[s] is followed by vowels of rather front quality (eg. 21), but [sʲ] has a very noticeable palatal off-glide (eg. 19), and when final in the utterance frequently sounds ejective as the secondary articulation is released. Other apico-dental sounds also are followed by front vowel qualities (exx. 2, 3).

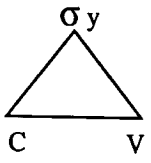
For many of the backer items, the degree of constriction for the 'secondary' velarising articulation is quite close, sometimes even giving the percept of complete dorso-velar closure.

Table 1. Some impressionistic records of Tundra Nenets.

1. mˠɑðᵊɜ	his hammer	2. mɣɑnᵊᵽ	paw
3. mˠʏɛǫ̈e mˠɛ̤ǫ̈æ	his foodstuff	4. mˠʏˈëⁿk̲iⁱ	seems to be there
5. mˠɯi̤ᵊðᵊᵽ	his foodstuffs	6. mʲjɐðᵊɣ̞ŋɑðᵊᵽ	broke something
7. mʲadᵊõsʌ̈ˡ	present (n.)	8. pʲɔðᵊŋɑkiⁱ pʲɔdᵊˈŋɑkiⁱ	seems to fight
9. ʎɐᵊðᵊëᵽʲɪɛ̤	trembling	10. (gˀ)lˠɛkʰ	lazy
11. lʷʸɔ̤ŋ̤gʌɪɐ̣	metal button, stud	12. ʎiuˠᵉrˠt̲ɑ	bushy (tail)
13. lˠɯ̤i̤ᵊðˠɜ	his bones	14. ɲiːðᵊɜ	his friend
15. pˠɑlˠi̤i	sword	16. cɯː	sleeve
17. tˠɑtˠiⁱ	elder wife	18. ɲebʲɛɐ̣	mother
19. sʲanˠɯ̃i̤	piece (in drafts)	20. tʷʸuʔˀ	fire
21. saũɲe	magpie	22. jakʷʸõ	it itches

There is a mutual dependency between the consonantal and vocalic articulations of Tundra Nenets syllables. The most satisfactory approach to dealing with palatalisation in Nenets is to treat it as a prosodic feature contrastive over the whole CV 'piece'. ('Piece' is a Firthian term, meaning an indeterminate amount of phonological material, [6].) Fig. 1 shows in graphic terms what is proposed.

Figure 1. Proposed treatment of palatalisation in Tundra Nenets.



The opposing term of **y** (which stands for 'palatalisation') will be labelled ¬**y** ('non-**y**').

This statement has the advantage that **y** vs. ¬**y** alternations, which are an important part of Nenets morpho-pholonogy [4, 7], can easily be handled. Furthermore, the number of **C** and **V** terms which commute in **y** pieces is different from the number which commute in ¬**y** pieces [7]; thus a structural motivation for this form of statement also exists.

PHONETIC EXPONENTS OF PALATALISATION IN NENETS

Table 2 gives a summary of the phonetic exponents of **y** and ¬**y** according to place of articulation of the exponents of **C**.

*Table 2. Summary of the exponents of **y** and ¬**y**.*

exponents of **y**	exponents of ¬**y**
overall fronter quality of the syllable	overall backer quality of the syllable
With bilabial constriction Open approximation of tongue body in palatal region, close and front in the mouth; maximal closeness timed to coincide with any complete closure; rather quick release of palatal gesture. Fronter subsequent vocalic portions; relatively closer vocalic portions	With bilabial constriction Open approximation of tongue dorsum at velum, which is rather slowly released, giving backer and generally more open vocalic portions.
With tongue-front constriction Articulations made with tongue blade and predorsum, with the tongue tip down: [j c ɲ ʎ]; followed by a palatal off-glide [sʲ]; momentary articulations made with the tongue tip against the upper teeth, and the tongue body close and front, followed by palatal glide	With tongue-front constriction Apico-dental articulations accompanied by central or back resonance; followed by vocalic portions with generally front quality; momentary articulations made with the tongue tip against the upper teeth and the tongue body close and back.
	With dorsal articulation Either complete dorso-velar closure with a slow release from velarity, [k ŋ]; or more open articulations [x̲]. Backer subsequent vocalic portions.

DISTINCTIVE PROPERTIES OF THE PROPOSED ANALYSIS

The proposed analysis keeps phonological and phonetic categories separate, thus avoiding any pseudo-phonetics in the phonology. It also allows the phonology to be accountable to the phonetics, since without an accompanying statement of phonetic exponency, the phonological categories are meaningless.

The proposed statement gives as the exponents of **y** not just 'secondary' tongue body gestures, but also the tongue shapes required to produce the exponents of the **C** terms in a **y** piece, and correspondingly in a ¬**y** piece. This is done without recourse to statements of allophonic variation.

By treating **y** and ¬**y** as categories applicable to **CV** structures, and by stating the exponents of **y** and ¬**y** over the whole consonant-vowel stretch, the question of whether to allot palatalisation to the consonant and spread it to the vowel, or whether to allot palatalisation to the vowel and spread it to the consonant, becomes redundant.

ACKNOWLEDGEMENTS

I would like to express my thanks to Tapani Salminen for his willingness to let me share his data, and to Anastasia Lapsui for acting as informant with such patience and good grace.

REFERENCES

[1] Firth, J. R. (1948). Sounds and Prosodies. Reprinted in Palmer, F. R. (ed.) (1970). *Prosodic Analysis*. London: Oxford University Press. pp. 1-26.

[2] Firth, J. R. (1957). A synopsis of Linguistic Theory, 1930-1955. In *Studies in Linguistic Analysis* (1957): Special Volume of the Philological Society, 2nd edition, 1962. pp. 1-32.

[3] Ogden, Richard & John Local (1994): Disentangling prosodies from autosegments: a note on the misrepresentation of a research tradition. To appear in *Journal of Linguistics*.

[4] Décsy, Gyula (1966): Yurak Chrestomathy. *Indiana University Publications, Uralic & Altaic Series* Vol. 50.

[5] Collinder, Björn (1957): *Survey of the Uralic Languages*. Stockholm: Almqvist & Wiksell.

[6] Sprigg, R. K. (1961). Vowel harmony in Lhasa Tibetan: prosodic analysis applied to interrelated vocalic features of successive syllables. Bulletin of the School of Oriental and African Studies 24. 116-138. (Reprinted in Palmer (1970), 230-252.)

[7] Salminen, Tapani (forthcoming): Tundra Nenets. In Daniel Abondolo: *The Uralic Languages*. London: Routledge.

DYNAMIC ARTICULATORY PHONOLOGY AND THE SUPERVISION OF SPEECH PRODUCTION

Mark Tatham
University of Essex, Colchester, U.K.

ABSTRACT

Articulatory Phonology unifies the domains of phonetics and phonology, linking utterance planning and execution by common units of control. It links with the **Task Dynamic Model** of speech production, forming a smooth data pathway from the most abstract level to the physical level of articulatory configuration. This paper reviews the need for refinements to the model and proposes task *supervision* to explain some data previously overlooked.

ARTICULATORY PHONOLOGY

Articulatory Phonology was proposed by Browman and Goldstein [1] partly as an attempt to unify phonetics and phonology by treating them as 'low and high dimensional descriptions of a single system' [2]. They come together by the idea that the constraints of the physical system underlie the phonological system, and by making the units of control at the planning level the same as those at the physical level; planning and execution are seen as more closely related than in other theories of phonetics and phonology.

The plan of an utterance is formatted as a *gestural score* (see Figs.2 and 3 for examples) which provides an input to a physically based model of speech production — the **Task Dynamic Model** [3]. The gestural score graphs locations and degrees of constrictions within the vocal tract, as well as time markers as an utterance progresses. The sequencing of gestures and their durations, and the timing relationships between the various vocal tract variables involved are critical to the score. The tract variables form a parametric framework which is manipulated in the **Task Dynamic Model**. Lip aperture, location and degree of tongue tip constriction, location and degree of tongue body constriction, velar aperture and glottal aperture are examples of tract variables.

As an example the gestural score for the utterance of a single [æ] would show that for a certain time the tongue body constriction is to be in the area of the pharynx and wide, with the velar aperture closed to prevent nasalisation, and the glottis closed to promote vocal cord vibration. Other tract variables may or may not be specified depending on how crucial they are to the utterance.

The gestural score graphs the utterance plan — an abstract representation related to vocal tract movements. Since they are abstract score gestures are correctly represented as discontinuous. Thus they capture the cognitive discreteness of phonological segments while indicating how they are to be organised within the plan.

THE TASK DYNAMIC MODEL

In the **Task Dynamic Model** gestures have a functional goal, called the *task* and executed by *coordinative structures* [4]. Coordinative structures are groups of articulators or their underlying musculature which are able to *internally communicate*. The model derives its phonological perspective from the expression of functionality, and its phonetic perspective from the task specification.

Within the **Task Dynamic Model** the individual tasks are independent of each other, though they are related functionally in the gestural score representation. The model's dynamic perspective is achieved through the control of movement towards the specified physical goals. The **Task Dynamic Model** focuses on the task itself, rather than on the parts of the articulatory system involved in executing it.

PLAN AND EXECUTION

Articulatory Phonology seeks to unify phonetics and phonology though a common framework and a formal statement of low level constraints on cognitive processes. The constraints are *prior* conditions on planning; the planner knows about them in a general sense before undertaking to score a particular utterance. The constraint knowledge base is formally static in nature.

Tatham [5] attempted to show that phonetic constraints fall broadly into two types: those which are obligatory and those which are optionally contollable. The optionality of a physical constraint rests in its ability to be itself limited or enhanced. Constraints which are not optional are not able to be manipulated in this way. The recognition of two major categories of physical constraint on articulation had been proposed much earlier [6].

Some consequences arise from modelling constraints in this way:

1. the planning mechanism must be aware that a class of constraint is manipulable;
2. the manipulation takes place at a phonetic rather than phonological level;
3. the universal set of linguistically usable phonetic possibilities is augmented by the manipulative processes.

Tatham and Morton [7] claimed that the internal behavioural properties of a phonetic object (the **Task Dynamic** *coordinative structure*) could be interfered with (re-*tuned* in **Task Dynamic** terms) dynamically during the course of an utterance. The interference is planned into the utterance.

RE-TUNING THE PHONETIC OBJECT

A phonetic object has internal properties. That is, much of its realisation is internally specified rather than being computed at some higher level. This object oriented approach is a major innovation in speech production theory, proposed by Fowler [8] (**Action Theory**) at the physical level and Tatham [9] (**Cognitive Phonetics**) at the cognitive level.

Tatham's model [10] allows for some dynamic adjustment of the phonetic object's internal properties. Two purposes:

1. phonological inventory enlargement;
2. dynamic contextual variation.

Dynamic contextual variation is the ability of the system to vary the precision of the realisation of a phonetic object dependent on semantic, syntactic and phonological context. The clearest example of this is when the context of a phonetic object significantly affects the probability of perception confusion — in which case its articulatory precision is enhanced. There are many examples of this kind of cognitively determined re-tuning of a co-ordinative structure [10].

In the next section of this paper, the idea of supervised, rather than automatic, execution of plans is discussed within the framework of **Articulatory Phonology** and the **Task Dynamic Model**.

SUPERVISED PLAN EXECUTION

In this revision of **Articulatory Phonology** speech production planning is concerned with specifying the **Dynamic Speech Scenario**. The variability data which formed the basis of cognitive phonetic theory was inconsistent with the notion that the gestural plan might be carried through from its abstract level to the physical articulatory level. This approach only allows for simple, non-cognitively based coarticulatory effects to explain why unexpected variants of gestures arise.

Although the Browman and Goldstein theory implies that a carry through *is* possible, it does not adequately allow for a basis to explain the observed articulatory and acoustic facts. Because **Task Dynamics** is not able to dynamically modify its procedures, the burden of explanation rests with **Articulatory Phonology** or with an additional external component. The **Task Dynamic Model** performs better if, in addition to an underlying gestural plan, it receives an input from an external component with a *supervisory* role. The supervisory component is responsible for overseeing the **Dynamic Speech Scenario** which will unfold under the control of the model.

Tatham and Morton [7] argued this point strongly in the context of modelling the causes of observed variations in articulatory precision. The phonological gestural score cannot, on its own, enable the explanation of why precision of articulation varies during the course of utterances. And a-linguistic coarticulatory phenomena offer no explanation. The *coarticulation supervisor* was introduced to allow for predictions derived from a model of perception running cotemporaneously to determine areas of an utterance requiring increased articulatory precision.

Using an example from the data presented in [5] we note that in English word-initial [p] is aspirated (as in *a pan*) whereas in French word-initial [p] is not aspirated (as in *une panne*). Waveforms of these two utterances are shown in Fig.1. **Articulatory Phonology** would account for these two utterances using the gestural scores shown in Fig.2.

But such an account resorts to explaining the long voice onset time following English

Fig.1. Waveforms of French 'une panne' and English 'a pan'. Note the aspiration in 'a pan'.

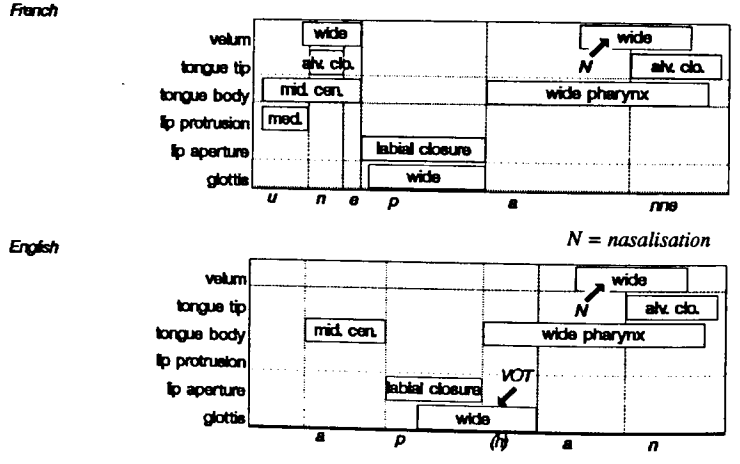


Fig.2. Unsupervised gestural scores for French 'une panne' and English 'a pan'.

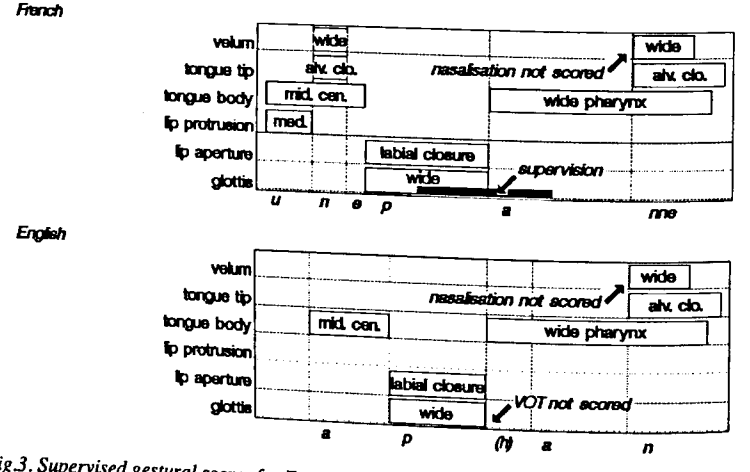


Fig.3. Supervised gestural scores for French 'une panne' and English 'a pan'.

initial [p] as a deliberate and *planned* event. Many researchers however have attributed this aspiration to an involuntary coarticulatory effect. But if the effect is involuntary it cannot be planned in or planned out at a cognitive level — unless the effect falls into our earlier optional class of low level constraint. And if this the way we might choose to model we observe then we can only say that the low level constraint exists as a *universal*, but that somehow its effect is partially negated in French.

The proposal is that the basic gestural plan for both English and French should be identical in the case of *a pan* and *une panne* in the relevant parameter (glottal constriction), but that the French plan be executed *under supervision* to allow for the optional limiting of the coarticulatory constraint (Fig.3).

Notice that it is no longer necessary to show the aspiration in the English. In the same examples we see that likewise it is no longer necessary to show nasalisation on the score. A result of introducing articulatory supervision is that we can leave phenomena such as aspiration and nasalisation to the **Task Dynamic** level — only when these phenomena are to be manipulated for the purposes discussed above do we need to deal with them in the gestural score. But because they are of a special nature (in traditional terms, not properly phonological) it is necessary to model them distinctly.

CONCLUSION

Articulatory Phonology and the **Task Dynamic Model** of speech production constitute a formidable advance in speech theory which is able to explain much data previously ignored. They do not handle well, though, subtle dynamic manipulations at the physical level during execution. This paper has argued that there is something to be gained by adding a **cognitive supervisory component** to the planning and physical components of the model.

REFERENCES

[1] Browman, C.P. and Goldstein, L. (1986), 'Towards an articulatory phonology', in C. Ewan and J. Anderson (eds.), *Phonology Yearbook 3*, Cambridge: Cambridge University Press, pp. 219–252.

[2] Browman, C.P. and Goldstein, L. (1993), 'Dynamics and articulatory phonology', *Status Reports on Speech Research*, SR-113, New Haven: Haskins Laboratories, pp. 51–62.

[3] Saltzman, E. (1986), 'Task dynamic coordination of the speech articulators: a preliminary model', in H. Heuer and C. Fromm (eds.), *Generation and Modulation of Action Patterns*, Berlin: Springer-Verlag, pp. 129–144.

[4] Fowler, C.A., Rubin, P. Remez, R.E. and Turvey, M.T. (1980), 'Implications for speech production of a general theory of action', in B. Butterworth (ed.), *Language Production*, New York NY: Academic Press, pp. 373–420.

[5] Tatham, M.A.A. (1995), 'The supervision of speech production', in C. Sorin, J. Mariani, H. Meloni and J. Schoentgen (eds.), *Levels in Speech Communication — Relations and Interactions*, Amsterdam: Elsevier, pp. 115–125

[6] Tatham, M.A.A. (1971), 'Classifying allophones', *Language and Speech*, Vol. 14, pp. 140–145.

[7] Tatham, M.A.A. and Morton, K. (1980), 'Precision', *Occasional Papers*, Vol. 23, University of Essex: Linguistics Dept., pp. 107–116.

[8] Fowler, C.A. (1977), *Timing Control in Speech Production*, Bloomington: Indiana University Linguistics Club.

[9] Tatham, M.A.A. (1979) 'Some problems in phonetic theory', in H. and P. Hollien (eds.), Amsterdam Studies in the Theory and History of Linguistic Science IV: Current Issues in Linguistic Theory, Vol. 9 — *Current Issues in the Phonetic Sciences*, Amsterdam: John Benjamins B.V., pp. 93–106.

[10] Morton, K. (1987), 'Cognitive phonetics — some of the evidence', in R. Channon and L. Shockey (eds.), *In Honor of Ilse Lehiste*, Dordrecht: Foris, pp. 191–194.

THE PHONETICS OF REDUCED VOWELS IN CHUVASH: IMPLICATIONS FOR THE PHONOLOGY OF TURKIC

Michael Dobrovolsky
University of Calgary, Calgary, Canada

ABSTRACT

In this paper, I suggest that vowel systems with non-high rounded vowels provide a challenge to speaker-listeners in both the production and perceptual domains. This challenge seems to be particularly marked for non-high front rounded vowels. I then relate the problem presented by the existence of such vowels to morphophonemic alternation patterns in three Turkic languages.

THE CHUVASH LANGUAGE

Chuvash is a unique Turkic language spoken in the Chuvash Republic, Russia, which extends inland mostly along the south and west shores of the Volga river where it turns southward some 600 kilometers east of Moscow. There are about 1.7 million Chuvash speakers (both bi- and monolingual), a number of whom live in neighboring republics.

Certain dialect variation aside, literary Chuvash—a composite of features of the two principal dialects—shows a "Turkic" underlying eight-vowel system ranged along front-back and high-low axes, with unrounded and rounded pairs in each phonological corner. The Chuvash analogs of the non-high rounded vowels are traditionally referred to as "reduced" or "weak" vowels and are typographically represented (both in Cyrillic and Latin transcription) as unrounded vowels with superscript breves. Adapting Krueger [1], I will initially use the following symbols: /i y e ĕ ɯ u a ă/.

The two "reduced" vowels of Chuvash are set apart from the other vowels of the system in that they are shorter in connected speech, subject to deletion in rapid speech and metrics, and yield to the full vowels with respect to stress assignment, which is, broadly put:: "stress the last full vowel of a word; when there are only reduced vowels in a word, stress the first vowel". But see Dobrovolsky [2] for a sketch of some factual and theoretical problems with this view

A PHONETIC PROBLEM

Data from the vowel systems of the world's languages demonstrates the relative lack of exploitation of non-high rounded vowels, especially non-back ones ([3] and [4]). I hypothesize that the presence of non-high rounded vowels creates a phonetic challenge that must be resolved in some linguistically acceptable way. This problem is multifold and arises from the linking up of a number of variables. In what follows, I coordinate a number of facts about the nature of non-high rounded vowels, especially front ones

Rounding and spectral fitness

Lip spreading renders front vowels more spectrally fit in that it serves to reinforce the height of their second and third formants. Conversely, lip rounding in front vowels can be thought of as rendering them less spectrally fit by lowering these same formants, thus contradicting their frontness. This effect appears to be particularly strong on the non-high front rounded vowels, to judge from the vowel inventories referred to above. One way to deal acoustically with this lessened fitness is to create a more distinct acoustic effect between non-high front unrounded and non-high front unrounded vowels by moving the latter to an acoustically more central position in vowel space. The ongoing conflation of /ø/ and /œ/ in Modern French and their continuing merger with schwa is one example of this path. It follows that Chuvash /ĕ/ is a prime candidate for this kind of acoustic adjustment. The non-high back rounded vowel /ă/, however, is already acoustically fit in that the lowering of the upper formants by lip rounded serves to emphasize its backness. Centralization of this vowel might be expected to result from other factors.

Articulation

Acoustic data and articulatory data suggests that the term "reduced", is appropriately used to mean "raised and/or centralized". The well-known facts of vowel neutralization in Russian illustrate this claim; its five vowel system /i e a o u/ manifests as /i a u/ in unstressed position. Russian unstressed /e/ neutralizes upwards to /i/ while unstressed /o/ appears to neutralize downward to /a/. But this phonemic description is misleading, as unstressed /a/ is manifested phonetically as central [ʌ] and [ə] depending on the segment's nearness to a stressed vowel.

Wood and Pettersson [5] have made a convincing case that reduction of open vowels in Bulgarian is related to three articulatory factors, (i) a lessening of jaw lowering, (ii) a lessening of lip rounding or spreading, and (iii) a lessening of pharyngeal narrowing. I suggest that the reduction of Chuvash /ă/ falls out of Wood and Petterson's factor (i), a lessening of jaw lowering. There is no compelling acoustic reason to deround/centralize the non-high back rounded vowel if there is no other non-high rounded back vowel in the system. However, it may well be that in contexts that deliberately contrast non-high rounded and unrounded vowels the acoustic centralizing effect will be more pronounced. This appears to be the case with some preliminary analysis of a word list containing such contrasts that I have recently made but will not report on here.

Stress/non-stress

Lack of stress may be equated with less precision in vowel articulation. This lack of precision in articulation is characterized by, among other things, a general reduction of articulator movement. A general centralizing tendency for vowels in the outcome.

Thus, articulatory and acoustic variables conspire to have an inevitable perceptual effect, namely, a lack of distinctiveness within respective sets of non-high vowels. I also speculate that the combination of greater jaw lowering and rounding requires articulatory effort that is non-optimal. If jaw lowering is compromised, a reduced vowel results. If rounding is compromised, a merger with the unrounded non-high vowels is threatened. Centralization of the merging vowel maintains its distinctiveness. The synergy of effects demands a phonological resolution.

SOME CHUVASH DATA

I report now on the spectrographic analysis of Chuvash vowels reported by Kotleev [6] for some 300 tokens of vowels "in various combinations and positions" from four speakers of the literary dialect and on other material collected by me during a two week stay in Chuvashia in July 1994. Four speakers—two females (ages: mid-twenties and early forties) and two males (ages: early thirties and mid-sixties)—were recorded in their homes or in a university residence using a Sony Walkman Professional WM-D6C and Realistic Electret tie-pin microphone 33-1063. None of the speakers had the literary dialect as their childhood speech, though all were to varying extents influenced by it. Each speaker came from a different area of Chuvashia and showed slightly different base dialect features: NW (Jadrenskij Rajon), N. Central (Cheboksarskij Rajon), E. Central (Marbosatskij Rajon) and S. (Batyrevskij Rajon). The data was elicited from a prepared word list in question-answer sessions using Chuvash and Russian. (In some cases the speaker used a different word in his or her dialect, so there are some gaps in the lists). For the purposes of this paper, some five tokens of each of the eight vowels were analyzed for each speaker (there are gaps in the number of vowel tokens reported here, notably, and inexplicably, /a/!). Spectral analysis was carried out using the LPC method on GW Instruments SoundScope/8 1.31 one-channel analyzer. All attempts were made to record formant frequencies from those areas of the vowels that appeared to be least affected by consonant transitions on either side.

VOWEL FORMANTS

Figure 1 plots the F1 and F2 formant frequency averages reported by Kotleev *op cit.* It appears that the reduction of /ĕ/ for his speaker is manifested as raising and a slight centralization of the vowel paralleling the centralization of /y/. The non-high back rounded vowel appears to be somewhat centralized as well.

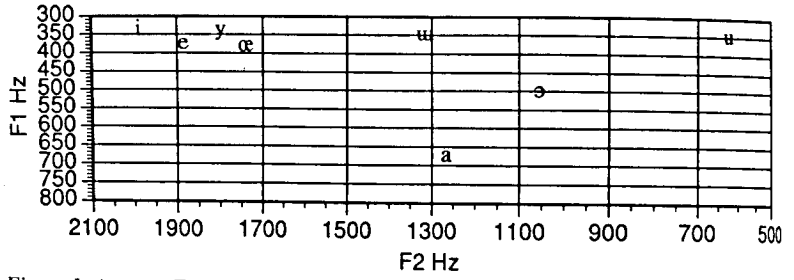


Figure 1. Average F1 /F2 for Chuvash speakers reported by Kotleev 1979.

For purposes of comparison with Kotleev, Figure 2 plots the F1 and F2 formant frequency averages for the four speakers recorded by me. A stronger trend towards centralization of the non-high front rounded vowel is evident. The non-high back rounded vowel appears to occupy the acoustic space transcribable as [ɔ], which was certainly the auditory impression it frequently gave.

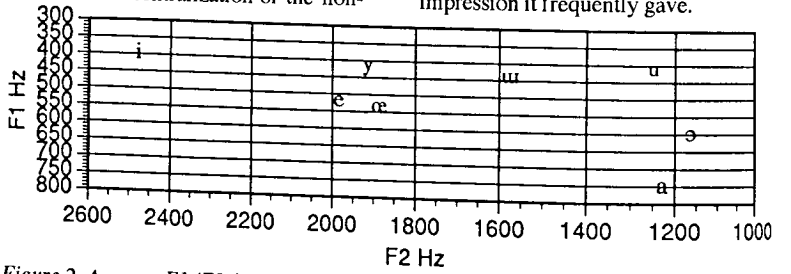


Figure 2. Average F1 /F2 for Chuvash speakers gathered for this study.

RESOLUTIONS IN TURKIC

Given that the presence of non-high rounded vowels provide the germ of a perceptual/articulatory problem that will seek to find a level of resolution that makes its way into the linguistic system as a whole (a.k.a. "language change"), I present three "solutions" from three languages along the vast continuum of the Turkic family.

Chuvash: reduction

The data from Chuvash presented above has already demonstrated what the historical resolution of the non-high rounded vowel problem in Chuvash (Possibly under the influence of neighboring Uralic languages like Mari). Non-high front rounded /ɛ̆/—more appropriately symbolized as underlying /œ/, which in my experience is its primary phonetic manifestation when stressed—in particular has become more centralized (or raised) and shorter. Non-high back rounded /a/, better represented as /ɔ/, shows less centralization. But these developments are far from complete. In stressed position, especially in contrastive monosyllables, the two vowels are still distinctive phonetically from each other and from all other vowels.

The well-know predictions of Liljencrants and Lindblom [7] regarding the shape of vowel systems are borne out in Chuvash: they predict that six-vowel systems exploit the high central area and seven-vowel systems add the high front rounded area. Their predictions for eight-vowel systems do not include a low front unrounded vowel. The proposed revision to Liljencrants and Lindblom in Crothers *op cit* still does not include a low front rounded vowel, but does, unlike Liljencrants and Lindblom, predict a schwa. Though the Turkic eight-vowel system in general confounds both predictions, it is worthwhile noting that Chuvash reduction appears to be moving in the direction both references expect, at least phonetically. Crothers p. 111, however, lists Chuvash as a system showing "an extreme typological deviation" in having two "interior" vowels, by which I assume he means /ɛ̆/ and /ă/. Recall, however, that the formant data reported on earlier suggests that /ă/ is not heavily centralized.

Turkish: restriction

Modern Standard Turkish shows another response to the presence of non-high rounded vowels. Here, these vowels are restricted to initial syllables in the native vocabulary. The high vowel suffixes of Turkish participate in a four-way harmonic alternation: /i ~ y ~ɯ~ u/. The non-high suffix vowel alternations are restricted to /e ~ a/, eliminating one possibility for the non-high rounded vowels to occur in non-initial syllables. The presence of an anomalous suffix like the *-ijor* progressive remains marked by the /o/'s opacity: it never alternates. This restriction of non-high rounded vowels to initial syllables may be viewed as the outcome of neutralization in non-stressed position if we accept the often stated proposal that Ur-Turkic stress was word-initial.

Yakut: expansion

As one moves eastward in Central Asia, there is increasing assimilation of both consonants and vowels irrespective of language. Yakut, spoken in the Saxa (Yakut) Republic in NE Siberia, is no exception to the areal pattern—rounding harmony is endemic. Krueger [8], p. 50, shows the high rounded vowels /y/ and /u/ followed in the next syllable by a high rounded vowel/diphthong or a low unrounded vowel, but non-high rounded /ø/ and /o/ followed only by a high rounded vowel/diphthong or low rounded vowel (front/back harmony applies as well). The suffix alternations among low vowels that are restricted to /e ~ a/ in Turkish thus show the full range of non-high rounded vowels in Yakut: /e ~ ø ~ a ~ o/, with /ø/ followed only by /ø/ and /o/ by /o/ among the low vowels. I suggest that the persistence of low vowel rounding improves the perceptual fitness of these vowels in longer forms.

CONCLUSION

There are several reasons for the evolution of Turkic vowel sequences, doubtless including language contact. I have suggest that phonetic theory provides some explanations for patterns of non-high rounded vowels in several Turkic languages. It is important to emphasize again that a phonetic "problem" like the existence of non-high rounded vowels is multidimensional and that there is not an inevitable common resolution along a teleological one-way street. Of course, how such apparently unattractive vowel systems arise constitutes another problem in its own right. Indeed, they exist, but as widely known and cited as the Turkic eight-vowel system is, Maddieson *op cit*, p. 127, notes that of the 317 languages in the UPSID database, only 24 (7.6%) have eight-vowel systems.

REFERENCES

[1] Krueger, J. R. (1960), *Chuvash Manual* (Uralic and Altaic Series, Volume 7), Bloomington: Indiana University Press.

[2] Dobrovolsky, M. (1990), "On Chuvash stress", in Brendemoen, Bernt (ed.). *Altaica Osloensia* (Proceedings from the 32nd meeting of the Permanent International Altaistic Conference), pp. 113-124. Oslo: Universitetsforlaget.

[3] Crothers, J. (1978), "Typology and universals of vowel systems", in Greenberg, J. (ed.) *Universals of Human Language*, vol. 2, pp. 93-152. Stanford: Stanford University Press.

[4] Maddieson, I. (1984), *Patterns of Sounds*, Cambridge: Cambridge University Press.

[5] Wood, S. A. J. and T. Pettersson (1988), "Vowel reduction in Bulgarian: the phonetic data and model experiments", *Folia Linguistica*, vol. 22, pp. 239-262.

[6] Kotleev, V. I. (1979), "Differentiating factors and acoustic characteristics of Chuvash vowels" [in Russian], *Sovjetskaja Tjurkologija* vol. 5, pp. 64-71.

[7] Liljencrants, J. and B. Lindblom (1972), "Numerical simulation of vowel quality systems: the role of perceptual contrast", *Language* vol. 48, pp. 839-862.

[8] Krueger, J. R. (1962), *Yakut Manual* (Uralic and Altaic Series, Volume 21), Bloomington: Indiana University Press.

THE EFFECT OF VOWEL CONTEXT ON ACOUSTIC CHARACTERISTICS OF [ç,x]

Christine H. Shadle, Sheila J.Mair and John N. Carter
Department of Electronics and Computer Science,
University of Southampton, Southampton SO17 1BJ, U.K.;
Neil Millner, DRA Malvern,
St. Andrew's Road, Malvern WR14 3PS, U.K.

ABSTRACT

In this paper we study the acoustic effects of vowel context on [ç,x] by spectral analysis of sustained and unsustained productions by two native speakers of German. Comparisons to non-German speakers of the same corpus allow inference of the acoustic mechanisms involved. [x] is more influenced by vowel context than [ç]. Evidence exists of changes in place, in the area of the constriction, and in the source localization or effectiveness, due to vowel context.

INTRODUCTION

In formulating models of fricative production, vowel context has important and sometimes unexpected effects, as shown by recent studies [1,2,3]. Acoustic spectra of [s] have been reported to display the greatest effect of vowel context in studies based on extensive analysis of two speakers (one a native of French, the other, of American English), and using spectral analysis primarily in the centre of the fricative [1,3]. Contrasting with these results is an analysis of aerodynamic data of fricatives, showing a greater effect of vowel context on area of the constriction as place of constriction moves posteriorly; it was suggested that fricative configurations independent of the tongue body are relatively immune to vowel context [2]. The work presented here therefore represents a more detailed look at the acoustic effects of vowel context on [ç,x].

METHOD

The speech corpuses consisted of the fricatives [s,ʃ,ç,x] in two environments: (1) preceded by the vowels, respectively, [a,a,i,a], sustained for 3 s, and repeated six times each, and (2) inserted into the nonsense words [pV_1FV_2] and repeated 10 times on a single breath, for V_1, V_2 chosen from [a,i,u]. Four subjects were recorded: the two subjects, CS and PB, reported on previously, for whom extensive articulatory and airflow data were available, a native German man, CD, for whom direct palatography and some airflow data were available, and a native German woman, EG. The inclusion of the German native speakers was prompted by several observations that CS and PB produced the phonemes not native to them more variably and with less place consistency.

The acoustic recordings were made under the 'High-Fidelity conditions' reported previously [3,4]. Time-averaged acoustic spectra were computed for the sustained fricatives using 8 non-overlapping 20-ms Hanning windows, positioned in the centre 160 ms of each token, and averaging the resultant Discrete Fourier Transforms (DFT's). Ensemble-averaged acoustic spectra were computed for the fricatives in vowel context by positioning one 20-ms Hanning window at the same position (i.e. beginning, middle or end of the steady-state portion of the fricative) in each of 8 tokens, and averaging the resultant DFT's. The technique is described in more detail in ref. [1]. The first and last tokens of an item were omitted.

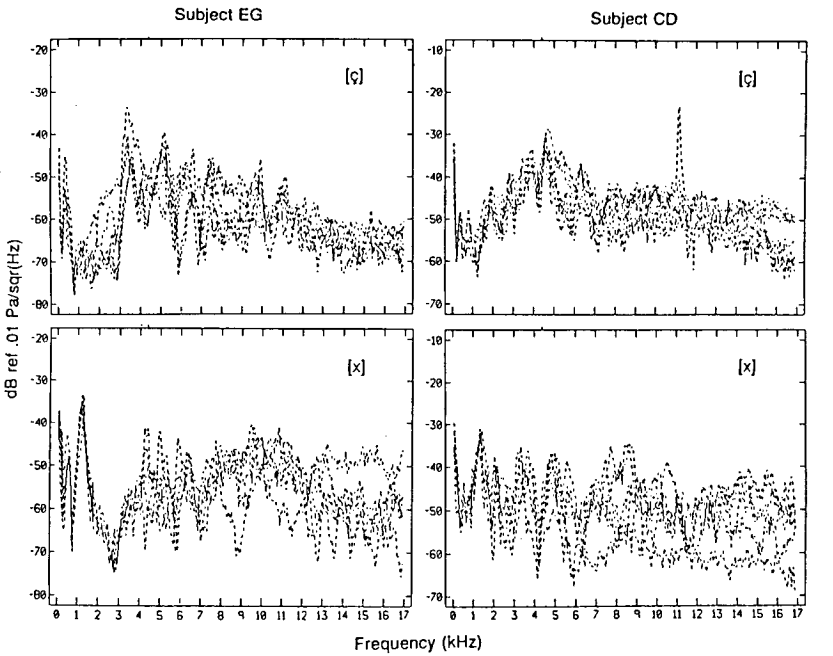


Figure 1: *Spectra of sustained fricatives produced by subjects EG and CD. Each graph shows six curves, one for each token.*

RESULTS

Figure 1 shows spectra of all six tokens of the sustained [ç, x] as produced by the native German speakers. The single peak at 11 kHz in CD's [ç] occurred in the first token produced and is indicative of a slightly whistly fricative – perhaps a response by the subject to the unnatural task of sustaining the fricative for 3 s. After comparison with the corresponding spectra of CS and PB (not shown), spectra of the sustained [ç,x] appeared to be slightly more variable token-to-token for the native Germans than for the other subjects, contrary to expectations. The overall spectral shape is fairly similar. In particular, [ç] is distinguished by low amplitude at low frequencies, extending up to 1 kHz (males) or 2-3 kHz (females), followed by a broad peak of high amplitude, made up of many smaller peaks, extending up to 6 or 7 kHz. [x] has fairly evenly spaced formant-like peaks, beginning at about 1 to 1.5 kHz, sometimes separated by deep troughs, as visible in EG at 3 kHz. Clearly there is more variation in the shape at frequencies above about 5 kHz. It is interesting to note, and

not unexpected, that the same-sex subjects show greater similarity in spectral shape than the same-language subjects.

The spectral structure (but not amplitude) of sustained and unsustained productions of a given fricative was similar for the same vowel context, for each subject. Within the unsustained productions, vowel context affected the frequencies of spectral peaks for all subjects; for example, for the most part an [u] context lowered frequencies. Figure 2 demonstrates a more extreme effect of vowel context, where 'natural' and 'unnatural' contexts are contrasted. Even though subjects commented on the 'impossibility' of items such as [paça, pixi], the fricatives did retain their distinctive spectral shapes even in such cases. They did alter, however, as shown in the figure: the high-amplitude region begins at a higher frequency for [ç] in [piçi] than [paça]; the amplitude relationship of the first two peaks in [x] is significantly reversed in the two contexts, and the trough at 2.8 kHz is much deeper in [paxa].

Subject CD shows less of a difference between [a] and [i] contexts than does EG, but more of a difference for an [u] context. Figure 3 contrasts [paxa] and [paxu], showing beginning, middle and end of the fricative steady-state for each. While the two cases begin similarly, in [paxu] the peak at 2.2 kHz has dropped 17 dB from beginning to end spectrum, contrasted with a 4 dB drop in [paxa]. Peak and trough frequencies remain the same, except for the peak at 1 kHz. This effect appears consistently in all of CD's [x] spectra with [u] context, and appears, but to a lesser degree, in EG's spectra.

DISCUSSION

The lack of articulatory and much aerodynamic data for EG and CD, and the difficulty of gathering such data for [ç,x], make it difficult to explain

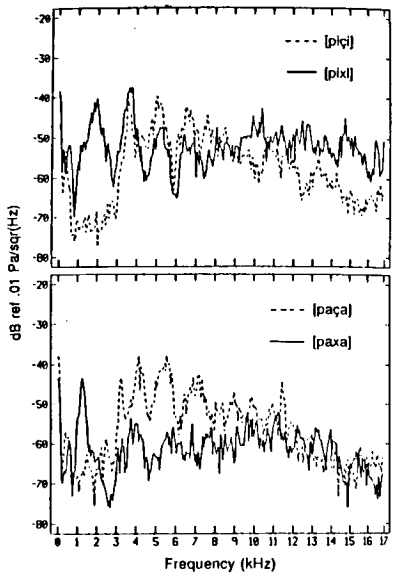


Figure 2: *Each graph contrasts ensemble-averaged spectra from the* middle *of the steady-state portion of the fricatives [ç,x] in the vowel contexts shown. Subject is EG.*

the various effects of vowel context observed above. However, from studies of CS and PB, for whom more articulatory data are available, some useful facts emerge. The cavity affiliations of the formants in [ç] were identified using a white noise source [4]; the lowest-frequency high-amplitude peak corresponds to the first front-cavity formant. In [s,ʃ] an [i] context shifts the place of constriction slightly forward relative to that for an [a] context. An [u] context for PB's [s] alters the source in a way that affects the fricative spectrum significantly [3].

The increased frequency of the peaks in EG's [piçi] compared to [paça] indicates that the place of constriction is more anterior in [piçi]. The same appears to be true for [pixi] compared to [paxa], where the high-amplitude peaks

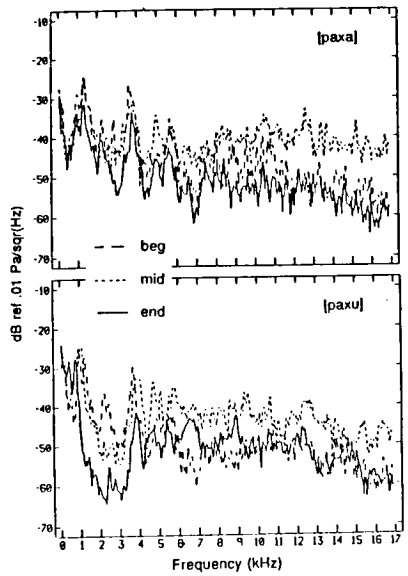


Figure 3: *Each graph contrasts ensemble-averaged spectra from the beginning, middle and end of the steady-state portion of the fricative [x] in the vowel contexts shown. Subject is CD.*

at 2 and 3.8 kHz appear to be front-cavity affiliated. However, the trough between these peaks deepens significantly for [paxa] and even more so for [puxu], which cannot be so easily explained by small shifts in place. The small peak visible in Fig. 3 at 2-2.5 kHz is a back-cavity formant, and is more prominent in [a] contexts, possibly indicating that the degree of coupling of back and front cavities is increased. The deepness of the trough in [u] contexts may be partly due to a reordering of poles and zeros due to the effect of rounding on pole frequencies, but it may also be due in part to changes in the degree of localization or effectiveness of the noise source. Since the constriction shapes and 'aims' the turbulent jet, such changes in spectral zeros may point to changes in the constriction shape. It is clear that for both subjects vowel context has a greater effect on [x] than on [ç].

CONCLUSIONS

The effect of vowel context on [ç,x] was investigated for two native speakers of German. Place of constriction moves anteriorly slightly in [i] contexts, increasing front-cavity formants, and rounding in [u] contexts decreases formant frequencies and bandwidths. However, some changes in constriction shape appear to occur as well, affecting the relative amplitude of back cavity formants and the significance of spectral zeros. These changes, and the fact that [x] exhibits more changes with vowel context than [ç], are consistent with Scully's explanation [2] of aerodynamic data on constriction area.

ACKNOWLEDGEMENTS

Work supported in part by a SERC studentship award to N. Millner, and by a collaborative EC SCIENCE award, CEC-SCI*0147C(EDB), and a European CEC-ESPRIT project Speech MAPS.

REFERENCES

[1] Shadle, C.H., Moulinier, A., Dobelke, C. & Scully, C. (1992), "Ensemble averaging applied to the analysis of fricative consonants", *Proc. of ICSLP-92*, vol.1, Banff, pp.53-56.

[2] Scully, C. (1992), "L'Importance des processus aerodynamiques dans la production de la parole", *Actes 19èmes Journées d'Etude sur la Parole*, Brussels 19-22 May, pp. 7-12.

[3] Shadle, C.H. & Scully, C. (1995), "An articulatory-acoustic-aerodynamic analysis of [s] in VCV sequences", *J. Phonetics*, vol. 23.

[4] Shadle, C.H., Badin, P. & Moulinier, A. (1991), "Towards the spectral characteristics of fricative consonants", *Proc. of the XIIth Int. Cong. of Phon. Sci.*, Aix-en-Provence, vol.3, pp.42-45.

INVERSION OF THE VOICE SOURCE FOR SOME FRICATIVES

Christophe Vescovi, Eric Castelli
Institut de la Communication Parlée U.R.A. - CNRS N° 368
I.N.P.G./E.N.S.E.R.G. - Université STENDHAL
46, Avenue Félix Viallet, 38031 GRENOBLE Cedex 1 France

ABSTRACT

In this paper inversion of a model of the vocal cords for vowel-fricative-vowel transitions is presented. The robotics approach of the inversion problem based on the forward modelling of the plant has already been successfully studied for the vocal tract and for the voice source in vocalic context. Two items (/pava/ and /paʒa/) are used here to study the validity of the inversion method with extreme source-tract interactions.

INTRODUCTION

The articulatory synthesis takes more and more importance in speech research nowadays. This kind of speech synthesis may be the only way to reflect all the human being variability, but is also very helpful to learn more about the speech production process using analysis by synthesis. Inversion of natural speech is a crucial point in those perspectives as it is the simplest way to provide appropriated commands to speech production models.

ICP's speech production model [1] can be divided in three parts : two acoustic modellings for the lungs and the vocal tract based on the Kelly & Lochbaum model [2] and a physical modelling of the glottis based on the Ishizaka & Flanagan Two-Mass model [3],[4]. Previous studies on the inversion of the voice source [5] have point out that the interaction between the voice source and its environment (mainly the vocal tract) should be taken into account in the inversion algorithm. This inversion method has already given encouraging results for Vowel-Vowel transitions, but has not be tested with more extreme source-tract interactions like in Vowel-Fricative-Vowel transitions.

METHOD

The robotics approach supporting this work is based on a forward model of the plant. The forward model can be defined as a mapping of the system under control, giving proximal to distal relations. In this work a polynomial (result of the analysis of a codebook) is used as a forward model, then a simple error backpropagation algorithm can perform speech to articulatory inversion (Figure 1).

In order to reduce the complexity of the inversion algorithm the dimension of the distal space must be as small as possible. Recent studies on the voice source [6], point out that three parameters could be enough to define the glottal flow of a speaker. Thus, the glottal flow is characterised by the fundamental frequency Fo, the energy of the speech signal E, and a wave shape parameter Rd. Using the classical LF parameters Ee and Uo (minimum derivative flow and maximum flow) [7], the declination ratio Rd can be defined by :

$$Rd = \frac{Uo}{Ee} \cdot Fo$$

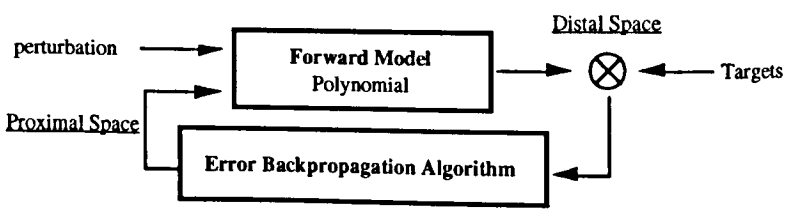


Figure 1 : general description of the inversion method

The proximal space is defined by the commands of the voice source model :
- Ps, the subglottal pressure.
- Lg, the length of the vocal cords.
- Ho, the rest aperture of the vocal cords.

Using the classical Two-Mass Model commands [3], Lg and Ho can be associated to Q and Ag0.

Perturbation of the voice source due to interactions with the sub and supraglottal cavities, must be taken into account in the inversion algorithm. Two parameters are used to define the vocal tract influence on the source, Ivt and Xvt respectively the inductance of the vocal tract and the position of the half inductance in the tract [5].

MEASUREMENTS

Two vowel-fricative-vowel items recorded by a French male speaker PB (/pava/ and /paʒa/) are analysed. The speech signal is first inverse filtered in order to evaluate the glottal flow and the three characterisation parameters are measured on the flow.

Formants trajectories used in the inverse filter are applied to ICP's inversion algorithm of the vocal tract which provides area functions and thus parameters Ivt and Xvt.

The intra-oral pressure (recorded simultaneously with the speech signal) is low-pass filtered as to give a approximation of the aerodynamic supraglottal pressure. Thus, the subglottal pressure Ps can be estimated in the consonant /p/ (the glottis is opened). ΔPg, the pressure drop at the glottis during the fricative can also be estimated by subtracting the intra-oral pressure to the estimated subglottal pressure (Table 1).

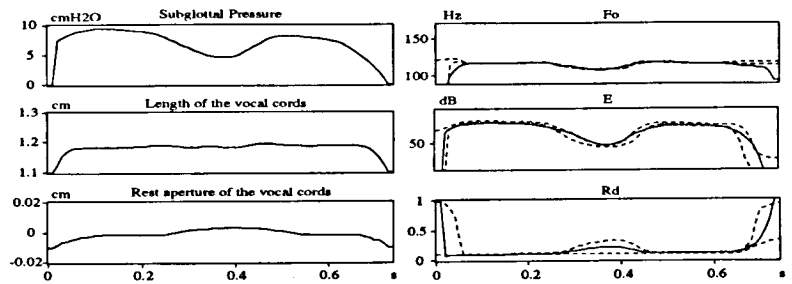


Figure 2 : results of the inversion for /pava/.
left : proximal space Ps, Lg and Ho.
right : distal space Fo, E and Rd
dashed lines = targets, solid lines = estimation of the forward model

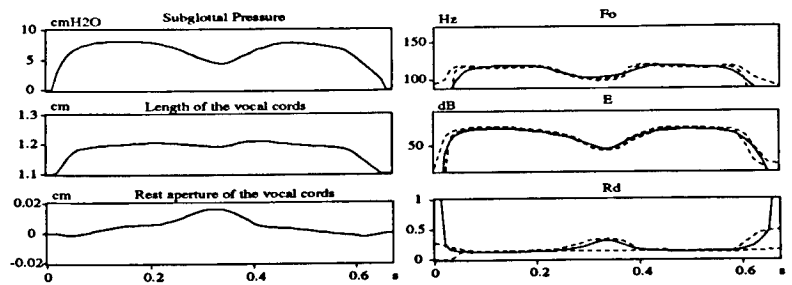


Figure 3 : results of the inversion for /paʒa/.
left : proximal space Ps, Lg and Ho.
right : distal space Fo, E and Rd
dashed lines = targets, solid lines = estimation of the forward model

Table 1 : measurements of subglottal pressure and pressure drop at the glottis in the fricative. (cmH2O)

	Ps	ΔPg
/pava/	9.2	3
/paʒa/	7.6	2

RESULTS

The measured flow characterisation parameters Fo, E and Rd are applied as targets to the inversion algorithm, whereas the parameters Ivt and Xvt are applied as perturbations. The results of the inversion for /pava/ and /paʒa/ are shown on Figures 2 & 3. For both items, the targets have been reached by the forward model without any normalisation of the distal space (especially for the wave shape parameter). this outstanding point can be explain by the specificity of the Rd parameter which does not reflect a geometrical property of the flow wave shape (like the traditional open quotien). Thus, Rd is a more "universal" wave shape parameter, unable to characterise speaker's differences, but very useful in our study to avoid speaker normalisation.

Trajectories in the proximal space are very similar for both items : one can think that glottis control strategies are the same for different voiced fricatives.

Trajectories of Lg and Ho seem to be quite realistic. There is no major actions on the control of pitch, Lg is quite constant during the utterance. Ho increase slightly in both fricatives as a way to maintain voicing [8].

On the contrary, Ps trajectory is more doubtful. In natural speech, the subglottal pressure varies very slowly, which is not the case in the result of the inversion.

The validity of the forward model can be easily checked by measuring Fo, E and Rd values on synthesised speech and comparing this values to those predicted by the forward model. Figure 4 shows that the predicted values are very close to measured ones, which means that the error might occurs in the speech production model. This is confirmed by the measurement of the simulated pressure drop at the glottis during the fricative, when the simulation is run using the inversion results. The simulated value of ΔPg (2.8 cmH_2O for /v/ and 2.2 cmH_2O for /ʒ/) are very close to the measured ones (Table 1). This result means that the two-mass model provides the wanted flow with the right glottis pressure drop, but those values are obtained for lower subglottal pressures (4.5 vs 9.2 for /v/ and 4 vs 7.6 cmH_2O for /ʒ/).

The two-mass model is not the cause of the error, but the modelling of the tract aerodynamic pressure used in the plant is not valid for small constrictions.

This problem is classical with the Kelly & Lochbaum modelling. This model is an acoustic modelling which is applied to low frequencies aerodynamic pressure by the introduction of specific losses. Those losses are subject to controversy because thy are based on very simple assumptions on the air flow through the tract that are not valid for most consonant production.

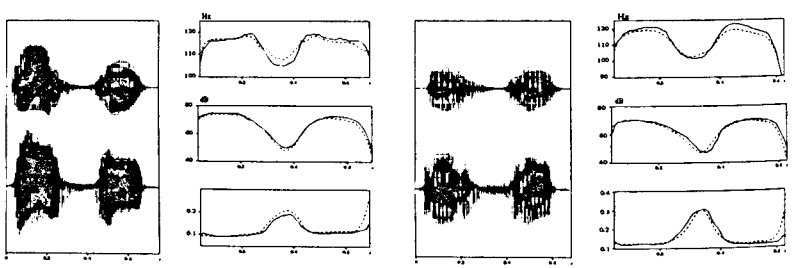


Figure 4 : comparison of estimated and measured values of the flow characterisation parameters for synthesised speech (left /pava/, right /paʒa/).
For each item (left to right and top to bottom) : synthesised speech, natural speech, Fo, E, Rd.
dashed lines : estimated values ; solid lines : measured values.

CONCLUSION

This study shows that the inversion method proposed for the voice source is able to deal with strong source-tract interactions as in voiced fricatives production. The results of the inversion are coherent with traditional knowledge on voiced fricative production and with measurements made on the speaker during the items production.

However, the aerodynamic pressure simulated in the plant, does not correspond to the measurement, and our modelling must be corrected before testing the method on other VCV, especially with unvoiced consonants, where there is a crucial contribution of the source-tract interactions to devoicing.

ACKNOWLEDGEMENT

This work has partially be funded by the European ESPRIT/BR project Speech Maps.

REFERENCES

[1] Bailly G., Castelli E., Gabioud B. (1994) *Building prototypes for articulatory speech synthesis.* Proceed. of the 2nd ESCA/IEEE Workshop on Speech Synthesis, New York, 9-12.

[2] Kelly J.L. & Lochbaum C.C. (1962) *Speech Synthesis* . in Proc. Stockholm-Speech Communications Seminar - R.I.T. 127-130. and 4th Int. Congr. Acoust., G42.

[3] Ishizaka K. & Flanagan J.L. (1972) *Synthesis of Voiced Sounds from a Two-Mass Model of the Vocal Cords.* B.S.T.J., 51, 1233-1268.

[4] Pelorson X., Hirschberg A., Van Hassel R.R., Wijnands A.P.J., Auregan Y. (1994) *Theorical and Experimental Study of Quasi-Steady Flow Separation within the Glottis during Phonation. Application to a modified two-mass model.* J. Acoust. Soc. Am., 96, 3416-3431.

[5] Vescovi C., Castelli E. (1994) *Gestural Supervisor for the Vocal Cords of a Speaking Machine.* In Proceedings of the Fifth Australian International Conference on Speech Science and Technollogy, December 1994, Perth, Vol.2, 613-618.

[6] Fant G., Kruckenberg A., Liljencrants J., Bavegard M. (1994) *Voice source parameters in continuous speech. Transformation of LF-Parameters.* Proceed. of the ICSP, September 18-22, 1994, Yokohama, Japan, Vol. 3, paper S25-4, 1451-1454.

[7] Fant G., Liljencrants J., Qi-Quang L. (1985) *A Four-parameter Model of Glottal Flow.* STL-QPSR 4/1985, 1-13.

[8] McGowan R.S., Koening L.L., Löfqvist A. (1995) *Vocal tract aerodynamics in /aCa/ utterances : Simulations.* Speech Communication, Vol 16, No 1, 67-88.

INTRINSIC VOICE SOURCE CHARACTERISTICS OF SELECTED CONSONANTS

Christer Gobl, Ailbhe Ní Chasaide and Peter Monahan
Centre for Language and Communication Studies, Trinity College, Dublin, Ireland

ABSTRACT

This paper presents data on intrinsic voice source characteristics of selected consonants for an Italian and a French speaker. The analytic method involved interactive inverse filtering of the speech pressure waveform, and measurements were obtained by matching the LF voice source model [1] to the output of the inverse filter. Results broadly bear out expectations that the degree of supraglottal constriction is the major determinant of source quality. Over and above this, differences between results for the two speakers suggest that active strategies may also come into play.

INTRODUCTION

Differences in the voice source were studied for the four consonants /l(:) m(:) v(:) b(:)/ in an intervocalic context. This is part of a more general study on the intrinsic voice source characteristics of vowels and consonants.

Our initial expectation here was that voice source effects could be modelled as a passive consequence of the differences in the supraglottal constriction associated with the different manners of articulation of these consonants.

METHODS

The main analysis technique involved inverse filtering of the speech pressure waveform. In order to obtain quantifiable results, a parametric model of differentiated glottal flow (the LF-model, [1]) was matched to the output of the inverse filter. Both the inverse filtering and the matching procedure were carried out using specially designed interactive software allowing optimisation in both the time and frequency domains [2].

From the matched model a number of parameters were subsequently measured. The ones we focused on particularly were EE, RA, RK and RG. EE is the excitation strength and is measured as the negative amplitude of the differentiated flow at the moment of maximum discontinuity. It corresponds to the overall intensity of the signal, so that an increase in EE amplifies all frequency components. RA is a measure of the return phase (dynamic leakage), which is the residual flow (from excitation to complete closure. The acoustic consequence of the return phase is a steeper spectral slope. A large RA corresponds to greater attenuation of the higher frequencies. RK is a measure of the skew of the glottal pulse: a larger value means a more symmetrical pulse shape. RG is a measure that relates to the duration of the opening branch of the glottal pulse. RK and RG together determine the open quotient, and they mainly affect the levels of the lower harmonics in the source spectrum. For a more detailed description of source parameters, see [3].

MATERIALS

The corpus used in this study was taken from recordings of two informants, one French and one Italian. The materials consisted of nonsense words read in similar carrier frames. For the Italian data, disyllabic nonsense words of the form $'C_1V_1C_2V_2$ were used in the frame *Dico ---- ancora.* As stress typically falls on the second syllable of a French disyllabic word, we used $'C_1V_1C_2$ monosyllables for the French nonsense words, set in the frame *Dis moi ---- aujourd'hui.* Note that in the French data, the final consonant of the monosyllabic word (C_2) occurred intervocalically in the carrier frame, thus providing a phonetically similar environment to that of the Italian.

C_2, the main object of our study, was each of the consonants /l(:) m(:) v(:) b(:)/. In Italian, C_2 was a long consonant. The first stressed syllable C_1V_1 was /ba/ in both languages. The unstressed V_2 was a vowel of approximately [a] quality in Italian. Five repetitions of each utterance were recorded resulting in a total of 60 utterances (2 speakers x 6 consonants x 5 repetitions).

RESULTS AND DISCUSSION

Figure 1 shows for both speakers the values for EE, RA and RK for the four consonants, and for 100 ms of V_1.

The sonorants /l(:)/ and /m(:)/

For the lateral and nasal consonants, the differences in source parameter values (compared to the surrounding vowels) were relatively small. The slight increase in RA suggests some increase in the spectral slope for both consonants. One difference between the nasal and lateral consonants seems to lie in the tendency of the nasal to have a more symmetrical glottal pulse shape (higher RK).

For the Italian speaker's nasal consonant, there are major perturbations at the V_1C transition in all parameters. There is a momentary drop in the excitation strength, along with a brief increase in dynamic leakage (RA). Concomitantly, the pulse becomes more symmetrical with a longer open quotient. There is little evidence of a similar effect for the French speaker, other than in the return phase.

All the perturbations observed in the Italian nasal would be consistent with there being a sudden reduction in the transglottal pressure drop. In the approach to the nasal consonant, the velum is likely to be lowered to some degree, resulting in some anticipatory nasal flow. However, if at the instant of oral closure, the outlet through the velar valve is insufficient for the volume of airflow, a momentary rise in oral pressure could result. The brevity of the perturbations suggests that a sudden, actively controlled, increase in velic aperture may occur soon after oral closure. It is also possible that such a sudden increased aperture could, at least partially, come about passively from the heightened oral pressure. Once the velic opening has increased, the transglottal pressure drop may resume something approaching its previous level, resulting in more efficient voicing.

This kind of explanation begs the question as to why a similar effect is not clearly found for the French speaker. We feel that the difference may have to do with the force of articulation used by the two speakers: the Italian speaker spoke with a relatively louder and more forceful voice, whereas the French speaker had a noticeably soft, lax voice. This is a point we shall expand on below.

The obstruents /v(:)/ and /b(:)/

The changes to the glottal pulse (relative to the surrounding vowels) are much greater for the obstruents than for the sonorants. This would of course be expected, given the greater degree of occlusion in the former.

For the fricative there is a gradual decrease in the excitation strength (EE) of the pulse, as well as a large rise in dynamic leakage (RA). The open quotient increases, and the pulse tends to become more symmetrical (RK), although not in every instance.

The most extreme source effects showed up in the stop. At the time of closure there is a sharp reduction in the excitation strength (EE). Although the reduction during the stop is similar for the two speakers (about 15 dB), the speed of the transition from the preceding vowel differs. For the French

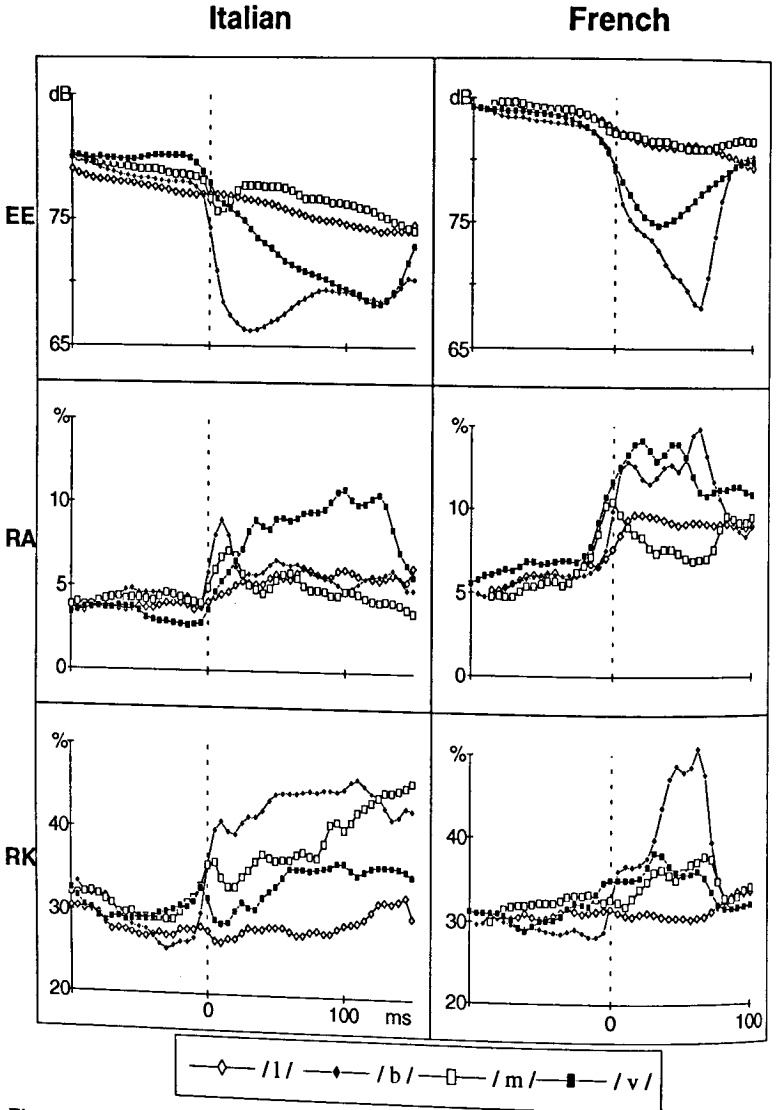


Figure 1. Source values for EE, RA and RK during the four consonants and for 100 ms of V_1. Values aligned to oral closure or onset of constriction for consonant (= 0 ms).

speaker the rate of decay was fairly similar to that of the fricative, whereas for the Italian speaker the decay is much more abrupt than for the fricative. Note also for the Italian speaker that after this sharp initial drop, EE rises somewhat. This is mirrored by a brief rise in f_0 and suggests that some active process is initiated soon after closure in the Italian geminate, which counteracts the otherwise expected decay in the pulse amplitude. This is presumably some compensatory action to maintain voicing, such as larynx lowering and/or oro-pharynx expansion. Furthermore, for the Italian speaker, there are perturbations around the time of closure, rather like those associated with the nasal.

We would suggest two (possibly complementary) factors might be responsible for these differences in the Italian and French data. First of all, it was mentioned earlier that the Italian speaker spoke in a more forceful style. One would expect a greater force of articulation to involve a more rapid oral closing gesture as well as greater respiratory effort, and a higher rate of flow through the vocal folds. These together should lead to a very rapid decrease in the transglottal pressure drop and a more extreme disruption of the vocal folds' vibratory pattern. This might explain the very rapid fall in EE at closure in the Italian stop, as well as the perturbations to the other parameters. Given this sharp initial drop in EE, the potential for devoicing is greater. Furthermore, the Italian stop here is a geminate: devoicing, due to neutralisation of the transglottal pressure drop is in any case more likely in stops of longer duration. We are therefore hypothesising that the length of the stop and the force of articulation may both conspire to make such an active compensatory adjustment necessary. We should make it clear however that these features are not postulated as necessary features of geminates: in an earlier study, [4] fully voiced geminates of Swedish were found to have a decay pattern more closely resembling the French pattern.

However, in proposing a "force of articulation" difference as being the underlying cause for a number of the differences in the French and Italian data, it remains unclear as to whether this in itself arises out of differences in reading style, possible cross-speaker differences or indeed cross-language differences.

CONCLUSIONS

Results broadly support our initial expectation that source parameters values are directly affected by the degree of supraglottal constriction. Thus, stops show more extreme effects than fricatives, which in turn are more affected than the sonorants. The latter show only minor deviations from the values of surrounding vowels. Of the two sonorants, the lateral was the least affected.

Not all the results can be modelled simply as "passive" source consequences of supralaryngeal occlusion. Some of the effects noted for the Italian speaker suggested compensatory active strategies may sometimes come into play. It is hypothesised that some of the differences between the two speakers may reflect differences in force of articulation.

ACKNOWLEDGEMENTS

This work was supported by Esprit-BRA, no. 6975, SPEECH MAPS.

REFERENCES

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

[2] Ní Chasaide, A., Gobl, C. and Monahan, P. (1992), "A technique for analysing voice quality in pathological and normal speech", *Journal of Clinical Speech & Language Studies*, Vol. 2, pp. 1-16.

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

[4] Ní Chasaide, A. and Gobl, C. (1993), "Contextual variation of the vowel voice source as a function of adjacent consonants", *Language and Speech*, 36, pp. 303-330.

Vowel-Vowel production on a Distinctive Region model. A new command strategy

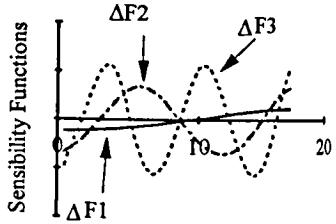
Samir Chennoukh and René Carré.
ENST, CNRS URA 820, 46 Rue Barrault,
75634 Paris cedex 13, France.

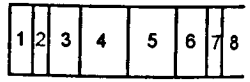
ABSTRACT

Previous studies on the distinctive region model emphasise the efficiency of the longitudinal command applied between two constriction regions to provide a quasi-rectilinear acoustic trajectory in the plane F_1-F_2 [1]. In this paper, the longitudinal command is not only applied between two constriction regions but also between a cavity and a constriction and vice versa [2]. This new command has two advantages. First, it allows to take into account the limitation of the tongue movements. Second, it allows to keep the quasi-rectilinearity acoustic property of formant trajectories.

I. DISTINCTIVE REGION MODELLING OF THE VOCAL TRACT.

The distinctive region modelling of the vocal tract has been described according to the hypothesis that the tongue realises a succession in time domain of single constriction and it articulates according to a minimum effort concept. If the articulation targets is acoustic, then the tongue must exploit regions of articulation where the formant frequencies are the most sensitive [2]. But the tongue articulates in two configurations which have different acoustic properties. If the lips are open, the configuration is a Closed-Open Tube (COT). If the lips are nearly closed, the configuration is a Closed-Closed Tube (CCT). In order to determine the regions of the model for the two configurations, the sensitivity functions [3] corresponding to each of the three first formants were calculated for an uniform tube. The regions are obtained by dividing the tube at the zero crossings of the sensitivity functions [1]. These sensitivity functions were calculated first for the configuration COT and eight regions were obtained (Figure 1). The region R8 represents the lip aperture. The closure of this region gives rise to the configuration CCT. The sensitivity functions were also calculated for this configuration and eight regions were then obtained with different boundaries (Figure 2).

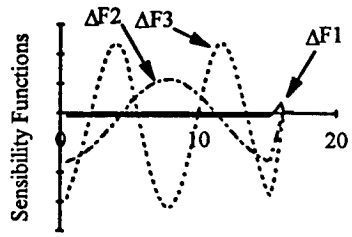


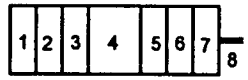


	R1	R2	R3	R4	R5	R6	R7	R8
F1	-	-	-	-	+	+	+	+
F2	-	-	+	+	-	-	+	+
F3	-	+	+	-	+	-	-	+

Figure 1. Distinctive region modelling in the configuration COT and matrix of variation of formant frequencies.

The superposition of the model with its two configurations on the vocal tract gives a physiological significance to the regions of the model (Figure 3). Four regions (R3 to R6) and five regions (R2 to R6) are considered as the tongue ones in, respectively, the configuration COT and the configuration CCT. These regions share out the tongue from the low of the pharynx up to the tongue tip.



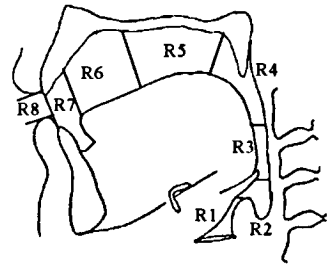


	R1	R2	R3	R4	R5	R6	R7
F1	0	0	0	0	0	0	0
F2	-	-	+	+	+	-	-
F3	-	+	+	-	+	+	-

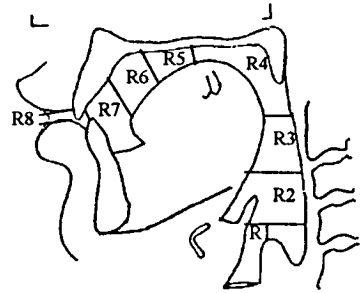
Figure 2. Distinctive region modelling in the configuration CCT and formant frequencies variation matrix.

The two configurations of the model allow to model the area function of any vowel by specifying its region and degree of constriction and its corresponding lip opening. In order to describe realistic area functions, the constraint of the constant volume of the tongue was integrated. This constraint is modelled by opposite commands on two regions linked by the acoustic synergy property. Namely, if a constriction is applied to one region among the tongue ones, a cavity is shaped on the region with which the constriction region shares the same variation of the formant frequencies by antisymetrical command.

The transition from an area function to an other is driven by a command strategy.



(a) Vocal Tract and Configuration COT



(b) Vocal Tract and Configuration CCT

Figure 3. Distinctive region modelling of the vocal tract.

II. COMMAND STRATEGY OF THE MODEL.

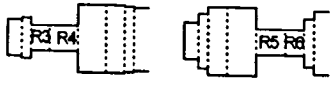
The command strategy elaborated for the model is built on a set of area function prototypes that differ in the region of constriction and the labial aperture state, i.e., open lips or nearly closed lips. The command strategy rules define a command, among possible ones, that minimises an acoustic criterion for every transition between two area function prototypes.

II.1. Area function prototypes.

The regions of constriction are chosen from the tongue region ones. Four area function prototypes are then obtained from the configuration COT (figure 4a) and one area function prototype is obtained from the configuration CCT (Figure 4b).

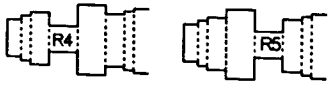
II.2. Two commands of the model.

A command of the model is described by the interpolation between two area function target parameters. If these parameters are the areas of the model regions, the interpolation gives rise to a transversal command (Figure 5a) and if these parameters describe the constriction, the interpolation gives rise to a longitudinal command (Figure 5b). This last command is obtained by the displacement of the constriction along the model.



R3 and R4 (BK)
(BK -*BacK*-)

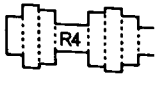
R5 and R6 (FT)
(FT -*FronT*-)



R4 (CBK)
(CBK -*Central BacK*-)

R5 (CFT)
(CFT -*Central FronT*-)

(a) Configuration COT

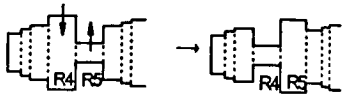


R4 (CC)

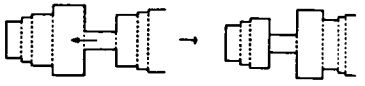
(CC - *Central Constriction*)

(b) Configuration CCT

Figure 4. Prototypes of area functions of the distinctive region model



(a) Transversal command



(b) Longitudinal command

Figure 5. Two commands of the distinctive region model.

II.3. Acoustic criterion and the primary rules of the command strategy.

The acoustic criterion is based upon an hypothesis where the best way to move from one target to another is the most rectilinear one. So the best command that is chosen to execute any transition is the one that minimises:

$$J(u) = \int_{t_0}^{t_1} (F_2(u(t), F_1) - F_{2_0}(F_1))^2 \, dt$$

where u(t) is the command versus of time, $F_2(u(t), F_1)$ represents the formant trajectory on the plane F_1-F_2 obtained with such a command and $F_{2_0}(F_1)$ is the straight line between the two acoustic targets on the plane F_1-F_2.

The optimal command choice for all possible transitions between area function prototypes gave rise to the primary rules which take into account no articulatory constraints.

II.4. Introduction of an articulatory constraint and the new command strategy rules.

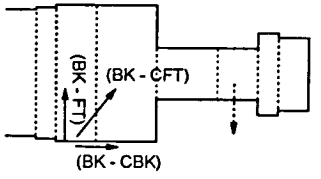
The preceding command strategy allows the longitudinal displacement of the constriction toward or from the low region of the pharynx. This behaviour of the area function is not realistic [2]. Two possibilities of representation of the area function have been considered to include this constraint. A constriction on one region of the area function involves a cavity on another region because of the constant volume of the tongue. So an area function can be represented either by the place and the degree of the constriction and the labial aperture or by the place and the degree of the opening of the corresponding cavity and the labial aperture. The exploitation of this equivalence in the configuration COT of the model allowed to replace the longitudinal command toward or from the constriction on the low region of the pharynx (R3) by the longitudinal command toward or from the corresponding cavity region (i.e., R6) (Figure 3a). This equivalent command allows to control the low region of the pharynx by only a transversal command such as it has been noticed in the last investigations of the natural articulatory data.

The use of this equivalent command in order to take into account the articulatory constraint led to the following rules:

- A target region must be different from R3 that represents the low region of the pharynx.
- A target region could be either a constriction or a cavity in the two area functions that constitute the transition. So the target regions of the command can be chosen among four possible couples. This rule derives from the use of the equivalent command.
-In order to take into account the concept of minimum of effort, the target regions of the command must be consecutive.

III. DISCUSSIONS.

Figure 6 gives an example of area function transitions on the model and their corresponding tongue movements. In this command strategy, several possibilities of the tongue deformation can be proposed and be a subject of study on the natural articulatory data. But it was showed that there is an interaction between the lips motions and the tongue mechanism. So, it would be more interesting to establish and take into account this interaction [2] before studying the natural data. However, the command strategy, as described in this paper, was used with benefits and other perspectives have been proposed on the improvement of the model [2].



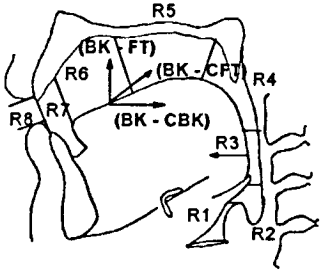


Figure 6. Examples of the tongue shaping from BK area function according to the new command strategy.

REFERENCES.

[1] Carré, R. & Mrayati, M. (1995), "Vowel transitions, vowel systems, and the Distinctive Regions Model", *Levels in Speech Communication: Relations and Interactions*, C. Sorin & al. (Eds.): Elsevier.

[2] Chennoukh, S. (1995), *Modélisation du conduit vocal en régions distinctives. Synthèse d'ensembles Voyelle-Voyelle et Voyelle-Consonne-Voyelle*, Dr. in Signal and Image thesis, ENST, Paris.

[3] Fant, G. & Pauli, S. (1974), "Spatial characteristics of vocal tract resonance modes", *Speech Communication Seminar*, Stockholm.

A REPRESENTATIONAL ACCOUNT FOR APRAXIA OF SPEECH

Jörg Mayer
Institute of Natural Language Processing, Stuttgart, Germany

ABSTRACT

The present study proposes a new interpretation of the underlying distortion in apraxia of speech. Based on the experimental investigation of coarticulation it is argued that apraxia of speech has to be seen as a defective implementation of phonological representations at the phonology-phonetics interface. The characteristic production deficits of apraxic patients are explained in terms of overspecification of phonetic representations.

INTRODUCTION

Most of the explanatory approaches in recent research refer to some higher cognitive functions within the motor system to describe the underlying pathomechanisms in apraxia of speech [1], [2]. But since pure apraxia of speech affects only verbal performance (cf. [1], [2]) and since patients suffering from apraxia of speech produce not only phonetic errors (eg. distortions) but phonemic errors as well (eg. substitutions), it is reasonable to think about a more linguistically based interpretation of this disturbance.

The aim of this paper is to propose a view on apraxia of speech which is founded on linguistic theory – in particular on nonlinear phonology and feature geometry – and which is supported by experimental evidence.

An important point of modern phonological theories is the principle of underspecification [3]. Underspecification is crucial for the description of many common processes such as vowel harmony, vowel-consonant asymmetries, reduction to unmarked sounds etc. Moreover, it has been shown that classes of sounds can easily be defined on the basis of underspecified, nonredundant representations and that the degree of underspecification can serve (a) as the structural representation of the degree of markedness within classes of sounds and (b) as a structural explanation for the sonority hierarchy [4].

Keating [5] proposes that parts of the phonologically underspecified gestures remain unspecified even in phonetic representations with the corresponding articulators resting in a neutral position or moving from one target to another without affecting the actually produced sound(s).

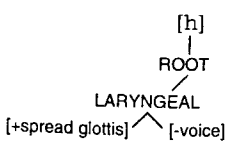
Now, our hypothesis is that apraxia of speech can be descibed as the loss of the ability of constructing underspecified phonetic representations.

COARTICULATION IN APRAXIA OF SPEECH

To illustrate and to support the overspecification hypothesis we will report some findings of experiments regarding the coarticulatory performance of patients with apraxia of speech.

VhV-sequences

The laryngeal fricative [h] is an element of the class of sounds which are maximally underspecified in phonological representation – the laryngeals (cf. [4]). For laryngeals it is only necessary to specify the features on the laryngeal tier, all features dominated by the supralaryngeal tier are underspecified:



So, assuming that the supralaryngeal underspecification is preserved in phonetic representation, [h] is expected to receive its supralaryngeal features from its phonetic context. Regarding formant structure this supralaryngeal transparency means, that the place features (which are primarily reponsible for the formant structure) of the surrounding vowels determine the position of the noise formant of [h].

Figure 1 shows a wide-band spectrogram of the first two syllables of the nonsense word *gehobe*. It is a part of the utterance *Ich habe* **gehobe** *gehört (I have heard* **gehobe**) by a male control speaker (data taken from [6]).

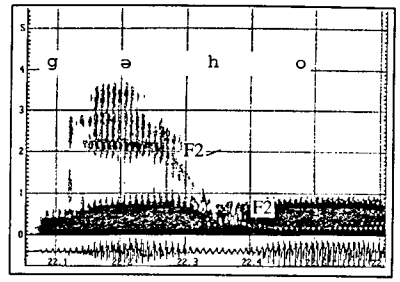


Figure 1. Wide-band spectrogram and signal display of **gehobe**, *male control speaker. Y-axis: frequency in kHz, x-axis: time in sec.*

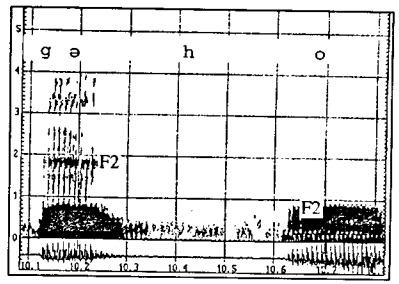


Figure 2. Wide-band spectrogram and signal display of **gehobe**, *male apraxic speaker. Y-axis: frequency in kHz, x-axis: time in sec.*

The noise formant of [h] has no target of itself. It moves, starting at the end of F2 of schwa, to the beginning of F2 of [o]. In other words, the source quality of [h] is well defined (laryngeal friction) but the characteristic of the oral filtering depends completely on the phonetic context.

In figure 2, the same part of speech is shown, uttered by a male subject suffering from severe apraxia of speech. He sustained a left-sided cerebrovascular infarct 2 years prior to testing.

There is no indication for any interaction between the vowels and the [h]. The vowels and the fricative are completely seperated. In terms of the overspecification hypothesis this can be interpreted as a full specification of each sound for all features and, as a result, in the case of [h] as the loss of the supralaryngeal transparency.

Figure 3 illustrates the two representational qualities with a underspecified representation on the left side and a maximally specified (i.e. overspecified) representation on the right side. In the non-redundant representation all segments are more or less underspecified. As mentioned above [h] is most underspecified, followed by the vowels, which are only specified for the place features (the schwa – due to its neutrality – lacks any further specification below the place node). The obstruent [g] needs the greatest amount of specification (cf. [4]).

As the dotted line indicates, formant interpolation can take place, due to the lack of any supralaryngeal specification

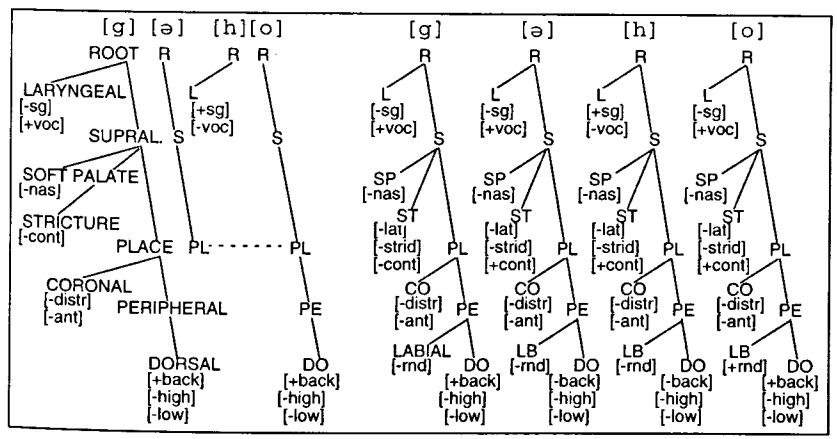


Figure 3. Underspecified, non-redundant representation (left) and overspecified representation (right) of **gehobe**. *The dotted line indicates the ability of formant interpolation.*

of the segment [h]. Regarding the overspecified representation, which is hypothesized to be the starting point of apraxic speech production, it is obvious that the propensity of features to spread is completely blocked. It is impossible for adjacent segments to interact because their (potential) transparency is eliminated.

Long-distance anticipatory coarticulation

In the second experiment (cf. [7]) anticipatory labial coarticulation at a distance has been examined. The material consists of the two test items *Spieligel* and *Spielübung* embeded in the carrier phrase *Ich sagte __ zweimal* (*I said __ twice*). The syllable structure and the morhological structure of both items are identical (# indicates a strong morheme boundary):

ˈʃpiːlˌigəl ˈʃpiːlˌybʊŋ

[CCVC]σ # [V]σ [CVC]σ

However, they differ in segmental structure: in the first item the unrounded vowel [i] serves as the nucleus of the second syllable, whereas in the second item it is the rounded vowel [y]. It is expected that these different vowel qualities in the second syllable have a different influence on the vowel in the first syllable (in both cases [i]). To check whether there is any anticipatory coarticulation we measured F2 and F3 of the [i] in the first syllable of both items. If coarticulation takes place F2 and F3 of [i] in the second item should be lowered – due to anticipated lip rounding – compared to F2 and F3 of [i] in the first item. As figure 4 illustrates this coarticulatory effect is fairly strong for control speakers (norm1 and norm2): F2 of [i] in the second item is by 30/65.5 Hz lower, F3 is lower even by 42.25/186.75 Hz (these are mean values of 4 utterances of each item). The same effect can be seen in a dysarthric speaker (dysarth) with a lowering by 49.25 Hz (F2) and 89.75 Hz (F3).

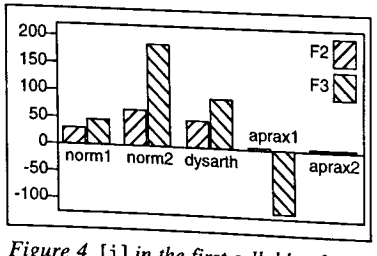


Figure 4. [i] *in the first syllable of item 1 vs.* [i] *in the first syllable of item 2: Mean differences of formant frequencies. Y-axis: difference in Hz.*

In comparison with this the apraxic subjects (aprax1 and aprax2) do not show any lowering of formants. Aprax1 is the same subject as in the first experiment, aprax2 is a 69 year old male with mild apraxia of speech as a result of a large left-sided cerebrovascular infarct 5 month prior to testing.

The lack of formant frequency lowering means that each [i] is realized identically independent of the phonetic context. The negative value for the difference of F3 in the case of aprax1 is due to a raise of the third formant in the context with the rounded vowel following. This might be interpreted as an instance of dissimilation – to get a better contrast between the target [i] and the [y] in the next syllable the formants of [i] are raised.

DISCUSSION

Both experiments have shown that apraxic speakers do not coarticulate, at least as far as macro coarticulation is concerned. Regarding micro coarticulation we have found differences between subjects dependend on the severity of the disturbance. In the speech production of aprax1, who suffers from severe apraxia, micro coarticulation is also absent. Spectrograms of his speech look like sequences of segmented steady states, with short intervalls of low energy across the whole spectrum between adjacent sounds. Transitions are missing. In comparison the milder distorted aprax2 does have transitions, at the edges of sounds his ability to coarticulate seems to be preserved.

In figures 5 to 7 the phonetic representation of the sequence [ily] is shown, which was the critical sequence in the second experiment. The representations are reduced to crucial gestures. We also abstained from indicating the hierarchical ordering to make the illustrations clearer. The most important difference to phonological representations (cf. figure 3) is that features are represented as boxes with extensions in the time domain.

The normal, partially underspecified representation is shown in figure 5. Since [i] and [l] are not specified for roundness (assuming non-roundness as neutral

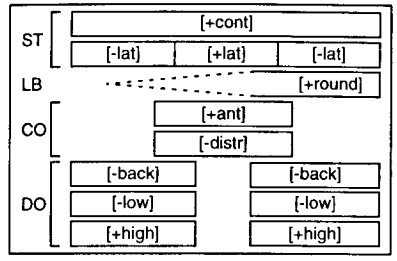


Figure 5. Reduced phonetic representation of [ily]. *Underspecification allows for the feature [+round] to spread forward.*

position for the lips) the feature [+round] in the domain of [y] is able to spread forward into the domains of the sounds ahead. This spreading process is indicated with a dotted line because the lips are supposed to achieve full rounding slowly somewhere before the [y]-domain – the [i] is still perceived as a plain [i], the correlation for anticipated lip rounding is only found with the help of acoustic analysis.

Overspecified representations as supposed for apraxic speakers are illustrated in figures 6 and 7. The representation in figure 6, however, differs from that in figure 7 in that it is additionally 'overspecified in the time domain'. All features of each segment are completely time aligned within the domain of the relevant segment. The lack of feature overlap represents the described loss of transitional phases between adjacent sounds in severe apraxia of speech. On the other side, figure 7 illustrates that preserved micro coarticulation is not a contradiction to the

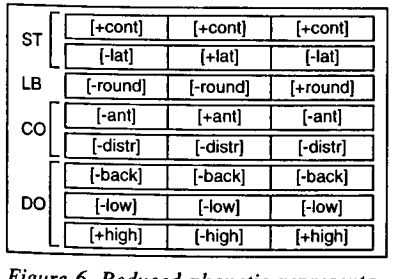


Figure 6. Reduced phonetic representation of [ily]. *Overspecification blocks spreading processes. Complete time alignment represents the loss of micro coarticulation.*

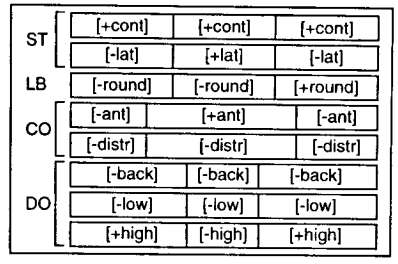


Figure 7. Reduced phonetic representation of [ily]. *Overspecification blocks spreading processes, but micro coarticulation (feature overlap) is intact.*

overspecification approach. Feature spreading at a distance is still blocked due to overspecification but the possibility of features to overlap locally is not affected.

CONCLUSION

We have explained one of the main characteristic of apraxic speech production – loss of coarticulation – in terms of overspecification of phonetic representations. Furthermore, we think that also other production deficits which are claimed to be typical in apraxia are interpretable within this framework, though we can not yet provide experimental evidence.

REFERENCES

[1] Lebrun, Y. (1990), Apraxia of speech: A critical review, *Journal of Neurolinguistics*, vol. 5(4), pp. 379-406.

[2] Ziegler, W. & von Cramon, D. (1989), Die Sprechapraxie - eine apraktische Störung?, *Fortschritte Neurologischer Psychiatrie*, 57, pp. 198-204

[3] Steriade, D. (1995), Underspecification and markedness. Goldsmith, J. (ed.), *The handbook of phonological theory*, Cambridge: Blackwell, pp. 114-174.

[4] Dogil, G. & Luschützky, H.C. (1990), Notes on sonority and segmental strength, *Rivista di Linguistica*, 2, pp. 3-54

[5] Keating, P.A. (1988), Underspecification in phonetics, *Phonology*, 5, pp. 275-292.

[6] Vollmer, K. (1993), *Untersuchung der Koartikulation bei gestörter Sprache: Experimentalstudie*, MA thesis, University of Bielefeld, January 1993.

[7] Mayer, J. (1994), Phonologisch-phonetische Überspezifizierung bei Sprechapraxie, *Phonetik-AIMS*, vol. 2(2) (Working Papers of the Chair of Experimental Phonetics, University of Stuttgart)

PHONOLOGICAL UNDERSPECIFICATION IN DEVELOPMENTAL APRAXIA OF SPEECH

I. Boers, B. Maassen and G. Thoonen
Child Neurology Center / Medical Psychology,
University Hospital Nijmegen, The Netherlands

ABSTRACT

Consonant substitutions and assimilations of two clinical groups of children (diagnosed with DAS or SLD) and a control group were analysed by means of computerized analysis (LIPP). Particular aspects of our data could be explained as the effect of phonological underspecification. Patterns of assimilation preference indicated that 'alveolar' and 'plosive' were underspecified relative to 'labial' and 'fricative' for all three groups.

1. INTRODUCTION

Developmental apraxia of speech (DAS) as a specific clinical entity has gradually gained support, but the disorder remains poorly understood. In a study bij Thoonen et al. [2], imitations of words and pseudowords, spoken by 11 carefully selected 'clear cases' of DAS and 11 age-matched control children, were phonetically transcribed. A quantitative analysis of the consonant productions showed that children in the DAS group produced significantly more substitutions and omissions than the control group, and showed a particularly low percentage of retention of place-of-articulation. Both findings were strongly related to the severity of the apraxia.

Also, the substitutions in DAS were more often assimilatory errors (anticipations, perseverations) than non-assimilatory errors, both for place and manner of articulation. This supports the finding in the literature that substitutions in DAS are related to context.

However, when assimilatory substitutions were assessed as a proportion of the total number of substitutions, no difference was found between the DAS group and the controls. Thus, quantitative but no qualitative differences were found between the groups.

In this paper we discuss the results of a qualitative analysis of the error profile in DAS, to examine if the patterns in feature assimilation of the DAS group can be explained based on the concept of phonological underspecification [1]. This concept holds that for some feature values the underlying phonological representations are not yet fully specified, but can be filled during the course of language production. It is hypothesized that assimilations of underspecified feature values to specified ones occur more frequently than the reverse. For place-of-articulation, the feature value 'alveolar' is considered to be underspecified, whereas 'labial' and 'velar' are specified.

2. METHOD

2.1 Subjects and material

More subjects were added to the DAS and control group of Thoonen et al. [2]. Also included was a small group of children diagnosed as 'speech/language-delayed' (SLD).

Our subject group consisted of 16 'clear cases' of DAS, between 5 and 10 years of age, 5 children diagnosed as speech/language delayed (SLD), ages between 4 and 10 years, and a control group of 24 age-matched children with no reported speech, language or hearing disorders. All subjects were native speakers of Dutch (for a detailed description of the subject selection see [2]).

Each child imitated 30 multisyllabic words and 36 two- and three-syllabic nonwords, spoken by the experimenter. All sessions were recorded.

2.2 Error analysis

All responses were phonetically transcribed according to standard IPA. Broad transcription was used. Based on the transcriptions, errors were classified as either a substitution, an omission, a distortion or a disfluency. A substitution was defined as a phonetically accurate production other than the intended target phoneme. To assess reliability of the transcriptions part of the material was transcribed by two other transcribers. Correlations of 0.9 and higher were obtained between transcribers.

By means of a computerized analysis [3], the transcribed substitutions were transferred into confusion matrices, indicating the relationship between consonant target and realization. The substitutions were then coded as 'correct' or 'incorrect' with respect to the features 'place' and 'manner' of articulation.

Separate confusion matrices for the features place and manner were constructed (see Table 1). The values for the feature place are 'labial' , 'alveolar/dental', 'palatal' and 'dorsal'. The feature manner had the values 'plosive', 'fricative/affricate ', 'nasal' and 'semivowel' . On the diagonals in Table 1a and 1b the 'correct substitutions' with respect to this particular feature appear, the off-diagonal numbers represent the errors.

For example, if a /t/ was replaced by an /s/, this was counted as a correct realization with respect to place of articulation (alveolar -> alveolar) so it appears on the diagonal of the place matrix. It is however incorrect for manner of articulation, and is counted as an error ('plosive to fricative') in the manner matrix. The totals of the columns of Tables 1a and b indicate the number of substitutions (including only the errors) towards this particular feature value.

In a subsequent analysis on the transcribed substitutions, the proportion of assimilatory substitutions was determined. Only regressive assimilations (anticipations) were taken into account. This analysis was done for each feature value separately, and proportions assimilatory substitutions were taken relative to the total number of substitutions towards this particular feature value. Thus, biases for particular feature values were accounted for.

3. RESULTS

In Table 1a to 1d the numbers of substitutions for place and manner are summarized.

Table 1a to 1d Confusion matrices of place and manner for the substitutions produced by the DAS and the SLD groups.

1a) PLACE (DAS, n=16)

target-	realizations L	A	P	D[2)]
L	*125*	128	1	46
A	100	*216*	11	105
P	1	9	*0*	2
D	29	137	1	*33*
tot[1)]	130	274	13	153

#place subst: 570
mean: 35.6 (sd 15.3)

1b) MANNER (DAS n=16)

target-	realizations PL	FR	NA	SV
PL	*400*	105	38	41
FR	50	*77*	7	3
NA	26	9	*64*	19
SV	13	40	31	*21*
tot[1)]	89	154	76	63

#manner subst: 382
mean: 23.9 (sd 12.8)

[1)] The column totals exclude the numbers in italics (= substitutions that are correct with respect to place resp. manner).

[2)] Dorsal combines velar and glottal

- table 1 continued -

1c) PLACE (SLD, n=5)

tar-gets	realizations L	A	P	D
L	*36*	18	0	5
A	30	*38*	1	11
P	0	0	*0*	0
D	6	9	0	9
tot[1)]	36	27	1	16

#place subst: 80
mean: 16.0 (sd 5.4)

1d) MANNER (SLD n=5)

tar-gets	realizations PL	FR	NA	SV
PL	*72*	13	7	20
FR	5	*8*	1	0
NA	6	1	*13*	9
SV	1	2	4	*1*
tot[1)]	12	16	12	29

#manner subst: 69
mean: 13.8 (sd 7.0)

All groups made more substitutions for place than for manner (means: DAS 38 vs 24, SLD 16 vs 14, controls 5 vs 4. Results for the control group are not displayed since they made so few substitutions). The numbers on the diagonal indicate the substitutions that were correct with respect to place (Table 1a and 1c) or manner (Table 1b and 1d) of articulation. Column totals are the numbers of substitutions towards each particular feature value.

In Table 2a and 2b proportions of substitutions towards a particular feature value are given.

Place. Almost half of the place substitutions in the DAS group were made towards 'alveolar' (.49) Proportions for 'labial' and 'dorsal' were approximately equal (.24 resp. .25). 'Palatal' was excluded from the table since hardly any substitutions towards this value were made. For the control group the same pattern emerged: towards alveolar the highest proportion (.43), then 'dorsal' (.31) and 'labial' (.26).

The SLD group differed in their preference: the proportion of substitutions toward labial was highest (.44), then alveolar (.34) and dorsal (.20) followed.

Table 2 Proportions substitutions towards feature values for place (2a) and manner (2b), relative to the total number of substitutions for that particular feature.

2a) feature PLACE
prop. substitutions to feature values:

group	L	A	D
DAS	.24	.49	.25
SLD	.44	.34	.20
CTRL	.26	.43	.31

2b) feature MANNER
prop. substitutions to feature values:

group	PL	FR	NA	SV
DAS	.27	.35	.20	.18
SLD	.22	.22	.21	.36
CTRL	.22	.26	.29	.23

Manner. For manner of articulation, the proportions of substitutions towards feature values were approximately equally spread, although the proportion substitutions towards 'fricative' were slightly higher than towards 'plosive', for both DAS and control group.

Assimilatory substitutions

In table 3a and 3b proportions of assimilatory substitutions are given.

Place. For both the DAS and the SLD group the majority of the substitutions to labial were assimilations (.70 resp. .76). Although both groups assimilated more often towards labial than towards alveolar (as predicted by phonological underspecification), we did not find higher proportions assimilations towards 'dorsal' than toward 'alveolar' (although predicted).

Since the control group made very few place and manner substitutions (mean 5.3 for place and 3.4 for manner) the data concerning the assimilatory substitutions (1 or 2 per feature value) should be interpreted with caution

Manner. For the DAS as well as the SLD group more assimilations were made in the direction of 'fricative' than towards 'plosive' (DAS: .72 vs .54, SLD: .78 vs .53). The proportion assimilations towards 'nasal' for the DAS group was lower than towards 'plosive' , and approximately equal for the SLD group.

Table 3 Proportions assimilatory substitutions for the feature values of place (3a) and manner (3b), relative to the total number of substitutions for that particular feature value.

3a) feature PLACE
prop. assimilation to feature value

group	L	A	D
DAS	.70	.48	.35
SLD	.76	.42	.37
CTRL	.30	.40	.25

3b) feature MANNER
prop. assimilation to feature value

group	PL	FR	NA	SV
DAS	.54	.72	.45	.30
SLD	.53	.78	.55	.43
CTRL	.32	.71	.37	.64

4. DISCUSSION

Stoel-Gammon and Stemberger [1] have used the concept of 'phonological underspecification' to explain patterns of consonant assimilation in child speech.

Two predictions can be derived from the concept of phonological underspecification. First, if the place and manner features remain underspecified until ultimate production of the segment, an alveolar plosive is produced (/t/ or /d/). Second, during the production process, underspecified feature values are vulnerable to intrusion from the context, resulting in more frequent assimilations from underspecified to specified values than the reverse.

Particular aspects of our data from children with DAS can be interpreted as the effect of underspecification. Firstly, with regard to place of articulation, DAS children produced a high proportion of substitutions towards 'alveolar' --as if in many instances the place feature remained underspecified. This pattern is in contrast with the preference for 'labial' in the children with SLD (but similar to the the control subjects). No distinctive patterns for manner were found.

Secondly, as with regard to both place and manner of articulation, the children with DAS produced --as predicted-- a higher proportion assimilations towards the specified values 'labial' and 'fricative' than towards the underspecified values 'alveolar' and 'plosive'. These patterns were quite similar for all three subject groups investigated.

These results pertain to a clinically salient characteristic of many children with DAS or SLD; they have a tendency to substitute towards /t/ (known in Dutch as "hottentottism"). Clearly, not all aspects of the presented data are explained; further study of speech production processes in DAS is needed.

5. REFERENCES

[1] Stoel-Gammon, C. & Stemberger, J.P. (1994). Consonant harmony and phonological underspecification in child speech. In: M. Yavas (Ed.), *First and Second Language Phonology*, (pp. 63-80). San Diego, California: Singular Publishing Group, Inc.

[2] Thoonen, G., Maassen, B., Gabreëls, F., & Schreuder, R. (1994). Feature analysis of singleton consonant errors in developmental verbal dyspraxia (DVD). *Journal of Speech and Hearing Research*, 137, 1424-1440.

[3] *LIPP: Logical International Phonetics Program V 1.40 (1991)*. [Computer software]. Miami, FL: Intelligent Hearing Systems.

FORMANT LOCUS EQUATIONS AND COARTICULATION IN DYSPRAXIC SPEECH

C.Chinnery, G. J. Docherty and D. Walshaw
Department of Speech, University of Newcastle upon Tyne .

ABSTRACT

Calculations of formant locus equations in the production of CV sequences were used to investigate the hypothesis that syllables produced by dyspraxic speakers could be characterised as being less coarticulated than those produced by normal speakers. The results give indications that some dyspraxic subjects can be described as having less coarticulatory cohesion between a consonant and a following vowel.

INTRODUCTION

Speech dyspraxia is an impairment in the volitional control and coordination of the muscles used in speech production. Speakers with dyspraxic speech typically have great difficulty in articulating words, even though they know exactly what they want to say. Their speech is characteristically dysfluent, marked by struggle behaviour and many false starts. It has been hypothesised that one of the central problems faced by dyspraxic speakers is precisely in the area of coarticulation. The evidence for this, however, is rather limited and is inconsistent. Ziegler & von Cramon [1, 2] report the case of a single apraxic speaker whose speech was marked by a delay in onset of anticipatory coarticulatory gestures resulting in a 'loss of segmental cohesion'. Itoh et al [3] note the presence of anticipatory coarticulation in their single apraxic subject, but note some deviations between their speaker and the pattern found for normal speakers. On the other hand, Katz [4] found no differences in coarticulatory patterns across normal speakers and those with posterior and anterior aphasia (with the anterior group being considered to be equivalent to those labelled in other studies as dyspraxic).

This paper reports a study which has investigated the hypothesis that dyspraxic speech is 'less-coarticulated' than normal speech; i.e. that speech sounds in dyspraxic speech production are produced in more discrete fashion than is found in normal speakers.

In order to measure the degree of coarticulation present in dyspraxic speech, formant locus equations have been employed. The application of locus equations to measurements of F2 was first described by Lindblom [5]. In calculating the equation, a straight line regression function is fitted to a scatter plot of F2 measured at vowel onset (F2ONSET) on the y-axis and F2 measured at the vowel midpoint (F2MID) on the x-axis. The relationship between these two quantities can be captured by the following equation,

F2ONSET = k * F2MID + c

where *k* is a coefficient relating to the slope of the regression line, and *c* is the estimated y-intercept. Sussman [6] has reported the existence of strongly linear relationships between F2ONSET and F2MID across different manners of consonant articulation. Different slopes and y-intercepts are found to correspond to different places of consonant articulation. For /g/ in English it is necessary to calculate three equations corresponding to cases with a following front, back unrounded, and back rounded vowels [6].

It has been noted [7] that locus equations can also be used as an index of CV coarticulation. A flat slope would indicate that F2 onset varies little as a function of different vowel environments suggesting relatively low articulatory cohesion between the C and the following V. Steeper slopes indicate that F2ONSET is increasingly coming under the influence of the F2MID value, indicating greater coarticulatory cohesion. In the context of the present study of dyspraxic speech, two questions arise: do dyspraxic speakers show linear relations between F2ONSET and F2MID similar to those found in normal speakers, and if so, do the slopes indicate any less articulatory cohesion than is present in normal speakers?

METHOD

Subjects

Two groups of subjects were recruited for the study; 5 dyspraxic speakers (D1 - D5) diagnosed as having verbal dyspraxia of speech, and 4 normal control speakers (N1 - N4). All subjects were native speakers of English from the North-East of England. Criteria for dyspraxic subject selection were: (a) they should be native speakers of English; (b) they should be diagnosed as having verbal dyspraxia by the speech & language therapist responsible for their case; (c) subjects should have reasonable comprehension abilities (sufficient to understand the elicitation task described below) (d) subjects should be able to read aloud real and nonsense words, or to repeat words without the aid of a visual cue. All subjects had suffered a stroke resulting in non-fluent aphasia and verbal dyspraxia to varying degrees of severity. Four normal control subjects were recruited broadly matched for age and sex with the dyspraxic subjects.

Materials

Subjects were asked to read a list of real and nonsense words with the structure C-V-/t/ formed by all possible combinations of /b,d,g/ and /i, ɪ, ɛ, a, ʌ, ɒ, ɔ, ʊ, u/ Each set of nine words was repeated six times by each subject giving a maximum of 162 single word utterances per subject. The CVC words were presented to both sets of subjects orthographically stencilled onto cards. The presentation of the cards was randomised. No data has been obtained relating to subject D3's production of words with an initial /g/, since this subject systematically produced these words with an initial /d/ (these tokens have not formed part of the analysis presented below).

Recordings and Measurements

Recordings were made in a recording studio or in a quiet room at the patient's home using a SONY Pro-Walkman D6 tape-recorder. The recordings were subsequently digitised at a sampling rate of 10Khz and analysed using a KAY Elemetrics Computer Speech Lab.

In line with [6, 8], formant measurement were carried out using two procedures (i) manual positioning of a cursor in a wide-band spectrographic representation; (b) LPC analysis of the same data, using the CSL's 'LPC formant history' routine. The average value of the two formant measurements for each vowel was taken and used in the subsequent statistical analysis.

For each word, two F2 formant measurement points were taken. (i) the value of F2 at the first identifiable glottal pulse following the release burst of the initial stop, as indicated by the first vertical striation (F2ONSET); (ii) the value of F2 at the mid-point of the vowel (the half-way point between the first and final vertical striations for the vowel (F2MID). Following [6, 8] the criteria listed were below were used to identify the measurement point for F2MID; (a) if the formant resonance was relatively 'steady state' a mid-point value of the steady-state portion was taken; (b) if the F2 resonance was diagonally rising or falling, a visually-determined mid-point was chosen; (c) if the pattern was either 'U-shaped' a measurement was taken at the point at which the curve changed direction (i.e. at the maximum or minimum frequency respectively).

RESULTS

Table 1 shows the principal locus equation parameters calculated for /b/, /d/, /g/ with a following front unrounded vowel and /g/ with a following back rounded vowel, for each of the speakers investigated. In almost every case (exceptions are discussed below) there is a strongly significant linear relationship between F2ONSET and F2MID ($p < .001$). We now consider, in turn the results from the normal and dyspraxic speakers.

Control Speakers

For normal speakers, significantly steeper slopes are found for /b/ than for /d/. For /g/, the results are less stable, but, on the whole, /g/ in the context of a back rounded vowel produces a steeper slope than /g/ in the context of a front unrounded vowel. Consistently lower y-intercepts are found for /b/ than for /d/, whilst with /g/ there is a difference depending on the following vowel environment with a lower y-intercept being found when a back rounded vowel follows. These results are entirely in line with those previously reported for normal speakers of English [6, 8]. The

Table 1. Parameters (slope, y-intercept) and R-squared) for the formant locus equations for dyspraxic (D1 - D5) and normal control subjects (N1 - N4).

Subj	Slope	(s.d)	Y-int	(s.d.)	R-sq
/b/					
D1	0.607	(0.046)	485	(72)	77.2%
D2	0.891	(0.047)	98	(72)	90.7%
D3	0.924	(0.041)	52	(72)	92.4%
D4	0.612	(0.026)	680	(45)	91.8%
D5	0.849	(0.028)	214	(38)	95.6%
N1	0.682	(0.031)	394	(48)	90.8%
N2	0.634	(0.043)	519	(78)	82.2%
N3	0.745	(0.032)	358	(50)	92.0%
N4	0.761	(0.042)	210	(56)	87.4%
/d/					
D1	0.188	(0.079)	1688	(130)	10.5%
D2	0.508	(0.058)	976	(91)	73.7%
D3	0.569	(0.049)	1088	(100)	76.9%
D4	0.309	(0.034)	1620	(65)	66.7%
D5	0.345	(0.043)	1317	(66)	61.3%
N1	0.323	(0.034)	1183	(53)	63.0%
N2	0.313	(0.036)	1432	(67)	63.2%
N3	0.367	(0.035)	1216	(57)	69.3%
N4	0.498	(0.034)	827	(47)	81.4%
/g/ before front Vs					
D1	-0.086	(0.137)	2329	(291)	0.0%
D2	0.577	(0.079)	1045	(147)	75.6%
D3	+-+-+ no data available +-+-+				
D4	0.411	(0.072)	1623	(169)	60.3%
D5	0.521	(0.058)	1127	(106)	84.4%
N1	0.200	(0.099)	1727	(191)	11.8%
N2	0.480	(0.137)	1293	(311)	34.9%
N3	0.373	(0.066)	1406	(131)	57.5%
N4	0.511	(0.145)	966	(239)	34.2%
/g/ before back rounded Vs					
D1	1.640	(0.309)	-539	(72)	69.4%
D2	0.615	(0.058)	593	(330)	22.3%
D3	+-+-+ no data available +-+-+				
D4	1.130	(0.131)	262	(160)	72.3%
D5	0.650	(0.190)	776	(212)	31.7%
N1	1.138	(0.134)	286	(154)	72.6%
N2	0.884	(0.034)	509	(185)	59.6%
N3	0.313	(0.138)	404	(147)	72.1%
N4	0.802	(0.092)	531	(104)	72.6%

steeper slope for /b/ indicates a higher degree of articulatory cohesion between /b/ and a following vowel than between /d/ and a following vowel, as might be expected given the functional independence of the bilabial and dorsal articulatory systems. The only exception to the general picture just described is with Subject N1's results for /g/ in the context of a following front vowel. The linear relationship between F2ONSET and F2MID is only borderline significant ($p = 0.056$), and the low R-squared figure suggests that only a very low percentage of variation in F2 onset can be predicted by the linear relationship with F2 midpoint.

Dyspraxic speakers

With the dyspraxic speakers, too, the general finding is that there is a strong linear relationship between F2 onset and F2 midpoint. Like the normal speakers, they show steeper slopes for /b/ than for /d/. The slopes for /g/ show considerable variability but two speakers (D1 and D4) have slopes for /g/ before back rounded vowels which are significantly steeper than for /g/ before front unrounded vowels (although see further comments on D1 below). For subjects D2 and D5 the differences in slope for /g/ as a function of vowel environment are less evident. Overall, there is no evidence that the dyspraxic speakers' slopes are any flatter than those found for normal speakers, suggesting that all speakers are showing comparable degrees of coarticulatory cohesion. The y-intercept estimates for /b/ are lower than for /d/, whilst for /g/, differences in y-intercept are found as a function of the following vowel environment. This general pattern of a linear relationship between F2ONSET and F2MID is not found uniformly across the dyspraxic group however. The clearest departure from this general trend is found in Subject D1's /g/ productions in the context of a front unrounded vowel where no linear relationship whatsoever can be found between F2ONSET and F2MID. Other subjects show instances where, whilst there is a linear relationship, its strength is considerably less than typically found for the normal speakers; for example, /d/ for speaker D1, or /g/ in the context of back rounded vowels for speakers D2 and D5.

DISCUSSION

The results show that locus equations for the normal speakers investigated conform to those previously reported in the literature showing a significant linear relationship between the onset of F2 at vowel onset and the value of F2 at the vowel midpoint, with differences in slope and y-intercept being found as a function of the identity of the pre-vocalic consonant. For the dyspraxic speakers, similar significant linear relationships are found.

However, some dyspraxic speakers occasionally show significant deviation from this normal pattern indicating, in those cases, less coarticulatory cohesion between the consonant and the following vowel. The results therefore suggest that for at least some dyspraxic speakers (such as subject D1), dyspraxia can be partially manifested in abnormal patterns of consonant-vowel cohesion as reflected in formant locus equations. We must also conclude that this impaired cohesion need not be found across every syllable produced by that speaker since subject D1's /b/ locus equation parameters are entirely within normal limits. It is noteworthy that the same speaker shows a considerably less reliable (though still significant) linear relationship between F2ONSET and F2MID in the /d/ syllables. It seems that difficulties for this speaker arise when trying to coordinate consonant and vowel articulations which involve lingual articulations and particularly when they involve the same part of the tongue dorsum (as in /g/ followed by a back rounded vowel).

This small study is the first attempt to use formant locus equations to investigate articulatory cohesion in apraxia of speech. The fact that some differences have been observed between the dyspraxic and normal samples and also within the dyspraxic group suggests the need for a follow-up study with a larger number of subjects, and looking at a broader range of pre-vocalic consonants. It would be particularly interesting to investigate whether formant locus equations permit a sub-categorisation of dyspraxic speakers by virtue of the degree of articulatory coherence which they show, and whether any such sub-categorisation corresponds to any other aspects of the subjects' speech and/or oro-motor performance.

ACKNOWLEDGEMENTS

The first author would like to acknowledge support from the University of Newcastle upon Tyne Research Committee.

REFERENCES

[1] Ziegler, W. & von Cramon, D. (1985) Anticipatory coarticulation in a patient with apraxia of speech. *Brain & Language*, 26,117-130.

[2] Ziegler, W. & von Cramon, D. (1986) Disturbed coarticulation in apraxia of speech: acoustic evidence. *Brain & Language*, 29,34-47.

[3] Itoh, M., Sasanuma, S., & Ushijima, T. (1979) Velar movement during speech in a speaker with apraxia of speech. *Brain & Language*, 7, 227-240.

[4] Katz, W. (1988) Anticipatory coarticulation in aphasia: Some methodological considerations. *Brain & Language*, 35, 340-368.

[5] Lindblom, B. (1963) Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America*, 35, 1773-1781.

[6] Sussman, H. et al (1991) An investigation of locus equations as a source of relational invariance for stop place categorisation. *Journal of the Acoustical Society of America*, 90, 1309-1325.

[7] Krull, D. (1989) Second formant locus patterns and consonant-vowel coarticulation in spontaneous speech. *Phonetic Experimental Research, Institute of Linguistics, University of Stockholm (PERILUS)*, 10: 87-108.

[8] Sussman, H. (1994) The phonological reality of locus equations across manner class distinctions: Preliminary observations. *Phonetica*, 51, 119-131.

TWO DIFFERENT SYSTEMS FOR RHYTHM PROCESSING AND THEIR HIERACHICAL RELATION

Morio Kohno
Kobe City University of Foreign Studies, Kobe, Japan

ABSTRACT

Kohno (1993) already suggested that the mechanism of rhythm processing consists of two neuropsychologically different works, by the examination of rhythm behavior of a patient with an infarction involving the corpus callosum. The present paper is to confirm this hypothesis with additional data. The latter part of this paper will clarify the neuropsychological relation between the two mechanisms of rhythm processing by the study of rhythm behavior of the patient of pure anarthria.

TWO MECHANISMS OF RHYTHM PROCESSING

Split-brain patient's rhythm processing

It is often suggested that, neurolinguistically, the processing of prosody is one thing and the processing of other linguistic elements such as syntactic structure is another. Borden and Harris [1], for example, proposed a model of speaking in which they indicated separate processors for prosody (including rhythm) and for word order. Matsubara et al. [2], proved that prosody, especially F0 control in speaking, is independent of other mechanisms involved in producing recurrent utterances in aphasic patients. But it is not known that rhythm is differently processed from intonation (F0 control). Kashiwagi et al. [3] first discovered that patients with infarction involving the forebrain commissural fibers behave very differently in fitting tempos in time with fast rhythm and slow rhythm. Kohno [3][4] ran a follow-up survey on the patient by requesting him to tap the table fitting various speeds of rhythm and found that the patient's left hand cannot follow any slow rhythms whose inter-beat intervals (IBIs) are more than 450ms, although it can manage to follow the fast rhythms whose IBIs are less than 330ms. His right hand, on the other hand, could properly tap in time to the both rhythms.

Table 1 illustrates this phenomenon.

Table 1. Tapping by a patient with infarction in the corpus callosus (male, 57 years old, right hander).

Hand Used	Target Tempo(IBI)	Observed Inter-beat Intervals N	MEAN	SD	r.v.	r
right	1000	27	1020.1	46.5	4.6	-0.52
	500	46	506.3	31.6	6.2	-0.25
	250	56	261.9	27.1	10.3	+0.19
left	1000	62	673.3	285.7	42.4	+0.96
	500	99	476.4	198.9	41.8	+0.07
	250	51	266.6	36.0	13.6	+0.13

(r.v.=relative variance, r=autocorrelations among the adjacent IBIs)

Table 2 shows the comparative data of a normal adult's behavior on the same task.

Table 2. Tapping by a normal adult(female, 55 years old, right hander).

Hand Used	Target Tempo(IBI)	Observed Inter-beat Intervals N	MEAN	SD	r.v.	r
right	1000	63	1022.7	52.1	5.1	-0.29
	500	55	512.7	22.9	4.5	-0.21
	250	99	257.5	10.6	4.1	+0.45
left	1000	57	1017.7	54.9	5.4	-0.10
	500	71	515.3	22.2	4.3	-0.12
	250	94	257.0	11.0	4.3	+0.04

The fact that the patient's left hand moves very differently not only from both hands of the normal adults, but from his own right hand, and that this feature of movements

can be seen when the tempos of stimuli swich from rapid rhythm to slow suggests that the processing of slow and rapid rhythms may be neuropsychologically different from each other. This hypothesis is supported by the fact that negative autocorrelations were detected among the adjacent IBIs in the slow response beats by the right hand of the patient and by both hands of normal adults, but never detected in any responses of the patient's left hand which produced only rapid responses even to slow stimuli and in the normal adult's response beats to the rapid stimuli (See the columns under r in the above tables.) Let us explain the mechanism of keeping time with slow and rapid rhythms.

With slow rhythm, if subjects are normal, they first get a general timing measure listening to the metronome, and then hit their first stroke on the basis of this measure. Their stroke, however, in most cases, misses the target, resulting in a stroke that is too early or too late. If the first stroke is early, they try to lengthen the next beat-interval to correct the timing. This action, however, again misses the target because of the overly-long interval. Subjects then hit their beat earlier in the second stroke by the same psychological reasoning. These reciprocal actions of earlier and later strokes produce negative autocorrelations. Therefore, we might call this processing 'analytic' 'one by one' or 'prediction-testing' processing.

With a rapid rhythm, however, there is no time for subjects to process each beat analytically. They get the configuration of the given rhythm in a flash and reproduce it in their tapping. We might call this kind of processing 'holistic' 'all-at-once' or 'Gestaltic' processing. It never produces negative correlations among adjacent IBIs.

To confirm this hypothesis, we carried out the following experiments.

Experiment 1

Subjects, materials and method: The rhythms with 250, 300, 400, 500, 750 1000ms inter-beat intervals were each aurally presented by the metronome, SEIKO Rhythm Trainer SQM-348, to the twenty university students majoring English and they were requested to reproduce those rhythms in the following two modes. Mode 1: After having listened to the stimuli for ten seconds, the subjects were requested to do multiplication of two digit numbers such as 27×48, and then to reproduce each rhythm from memory by saying ta ta ta ····. Twenty seconds were allotted for the calculation (If the calculation was finished by the end of the alotted time (signaled by a bell), the subjects had to wait). Sheets of paper were delivered on which numerical formula were described (e.g. $\frac{27}{\times 48}$) to calculate and write the answers. Mode 2: In place of calculation, the subjects drew circles(○) on the paper for twenty seconds, and then to reproduce the given rhythm by saying ta ta ta ····. All the subject's responses were tape-recorded and their IBIs were measured by the use of ONSEIKOBO NTT Advanced Technology.

Results: In order to know how diverse each response is from its target, each response beat interval was dealt with according to the following formula:

(response beat interval − target interval) ÷ target interval × 100 (absolute value).

The results about the means of differences from the targets are shown in Tables 3 and 4.

Table 3. Significance levels for comparison of the means of differences in the case of reproduction after calculation.

Target Intervals (ms)	250	300	400	500	750	1000
x	7.0	11.2	18.0	20.0	18.3	21.6
250 (7.0)		NS	0.05	0.01	0.01	0.1
300 (11.2)			NS	0.05	NS	0.05
400 (18.0)				NS	NS	NS
500 (20.0)					NS	NS
750 (18.3)						NS
1000 (21.6)						

N=20

ANOVA: $F(5, 114)=3.377$ $p<0.01$

Table 4. The means of differences in the case of reproduction of rhythms after drawing circles.

Target Intervals (ms)	250	300	400	500	750	1000
x	5.9	7.7	10.1	10.3	8.8	11.0

ANOVA: $F(5, 114)=0.959$ N.S.

The results about the SD values are shown in Tables 5 and 6.

Table 5. Significance levels for comparison of SD in the case of reproduction after calculation.

Target Intervals (ms)	250	300	400	500	750	1000
SD	6.81	11.90	23.65	21.99	22.16	28.19
250 (6.81)		0.05	0.01	0.01	0.01	0.01
300 (11.90)			0.01	0.01	0.01	0.01
400 (23.65)				NS	NS	NS
500 (21.99)					NS	NS
750 (22.16)						NS
1000 (28.19)						

N=20

Cochran's test: $F=0.316$ $p<0.05$

Table 6. SD values in the case of reproduction of rhythms after drawing circles.

Target Intervals (ms)	250	300	400	500	750	1000
SD	5.93	9.05	12.36	13.33	15.15	13.05

Cochran's test: $F=0.265$ N.S.

Tables 7 and 8 show the differences of responses between Modes 1 and 2 per each target rhythm.

All the these tables show the rapid rhythms whose IBIs are 250ms and 300ms were well-memorized and little disturbed by the tasks of both drawing circles and calculation but the memory of slow rhythms whose IBIs are more than 400ms was largely disturbed by the task of calculation while the work of drawing circles did not decrease the memory so much.

Table 7. Comparison of the means of differences between the cases of reproduction after drawing circles and after calculation.

Target (ms)	Cir. → Cal.	p
250	5.9 : 7.0	NS
300	7.7 : 11.2	NS
400	10.1 : 18.0	0.05
500	10.3 : 20.0	0.01
750	8.8 : 18.3	0.01
1000	11.0 : 21.6	0.01

ANOVA: $F(11, 228)=4.47$ $p<0.01$

Cir.: the case of reproduction after drawing circles

Cal.: the case of reproduction after calculation

p: significance levels

Table 8. Comparison of SD between the cases of reproduction after drawing circles and after calculation.

Target (ms)	Cir. → Cal.	p
250	5.83 : 6.81	NS
300	9.05 : 11.90	NS
400	12.36 : 23.65	0.01
500	13.33 : 21.99	0.05
750	15.15 : 22.16	NS
1000	13.05 : 28.19	0.01

Cochran's test: $F=0.235$ $p<0.01$

Discussion: Slow rhythms with more than 500ms IBIs will be analytically processed, as suggested by the study of split-brain patient, and this analytic processing of rhythms may be the same kind of active task as calculation, and therefore the retention of this kind was interfered with by the calculation. The rapid rhythm processing, which may be holistic, however, is neuropsychologically different from the work of multiplication, and therefore, it was never disturbed by it. The work of drawing circles is so simple that it effected nothing on memory, just like immediate recalling after having heard the target rhythms (See [4]).

HIERARCHICAL STRUCTURE OF RHYTHM PROCESSING

The above-mentioned experiment about the split brain patient's rhythm perception shows an important fact that the right hand of the patient can do analytic processing with slow rhythms, and at the same time it can do holistic processing with rapid rhythms, while his left hand can only do holistic one (only follow rapid rhythms). This suggests that the analytic processing can be carried out on the basis of holistic processing, but not vice versa. Kohno [4] explains analytic processing, and says that if some person has no ability to make up a general Gestaltic map of tempo about the given rhythm, it is impossible for the person to fit it, even if the given rhythm is a slow one. Fodor[6] says that, in his model of listening comprehension, the modules, fast and holistic processing device, constitute the preliminary processing stage and the slow and analytic, but accurate processing device, that is, so called Central Processing Mechanism, makes up a primary processing stage. All the above-mentioned investigations suggest that the analytic and holistic processing act by a hierarchical system - the holistic may act as a basic structure, and the analytic as a hyperstructure.

Kohno et al.[5], on the other hand, carried out series of experiments using a patient of pure anarthria, and found that the patient demonstrated too analytic idiosyncrasy, processing the fast rhythms with 250ms IBIs by analytic way. In spite of this expreme analytic tendency, the patient still showed the existence of the productive sense unit (PrSU), couterpart of perceptual sense unit, both of which are manifestations of human being's holistic ability (cf.[4]). The patient demonstrated the PrSU by the pitch rise at the end of each unit, a strange way of utterance which is seldom heard in the normal speech in colloquial Japanese. This abnormal way of utterance, however, automatically disappeared, as his very slow speech rate became faster on acount of rehabilitation. This phenomenon therefore shows that pure anarthria might be caused by the suppression of holistic processing by analytic processing, without destroying the former. This phenomenon also suggests the hierarchical structure of rhythm processing. (Full information of this study will be given by printed paper.)

REFERENCES

[1] Borden, G.J. et al. (1984), Speech Science Primer, Baltimore: Williams & Wilkins, pp. 9-12.

[2] Matsubara, J. et al. (1994), "Prosody of Recurrent Utterances in Aphasic Patients", Proceedings of 1994 ICSLP, Vol. 3, pp. 1211-14.

[3] Kashiwagi, A. et al. (1989), "Hemispheric asymmetry of processing temporal aspects of repetitive movement in two patients with infarction involoving the corpus callosum", Neuropsychologia, Vol. 27-6: 799-809.

[4] Kohno, M. (1993), "Perceptual sense unit and echoic memory", International Journal of Psycho-linguistics, 9-1: 13-31.

[5] Kohno, M. et al. (1994), "Rhythm processing by a patient of pure anarthria and productive sense unit", G. Mininni & S. Stame (eds), Dynamic Contexts of Language Use, University Press, Bologna, Italy.

[6] Fodor, J.A. (1983), The Modularity of Mind, Boston: MIT Press.

THE PECULIARITIES OF LATERALIZATION OF SYLLABLE PERCEPTION IN STUTTERING AND NORMAL CHILDREN.

E.S.Dmitrieva, K.A.Zaitseva.
Sechenov Institute of Evolutionary Physiology and Biochemistry, St.-Petersburg Russia.

ABSTRACT

The stuttering and normally speaking children of 4-16 years old have been found to show the similar mode of cerebral specialization age development for syllable perception, demonstrating left- hemisphere superiority beginning from 8 years old. This fact allow to suppose that one of the possible reasons of stuttering might be in some functions deficit of the right, but not the left hemisphere.

INTRODUCTION

The investigation of central mechanisms underlying such disturbance of speech as stuttering is of great importance, and it is seen from a growing body of research and clinical literature of the past two decades. Some writers have supposed that a neurological central dysfunction might be an etiological factor in stuttering or a predisposing or contributing factor to the etiology of stuttering. The theory, that account for such dysfunction by the specific features of functional brain asymmetry (FBA) in stutterers, neuropsychological theory, proposes that stuttering is caused by "aberrant interhemispheric relations" [1].

Since stuttering usually appears in childhood, data indicating which disruptions in hemispheric interrelations are present in stuttering children take on particular importance. The hypothesis exists, that stuttering may be induced by an aberrance in the formation dynamics of functional hemispheric specialization during ontogenesis [e.g. 2-4 and some others].

Though there are not very many studies of FBA peculiarities in stuttering children, they are also, as in the case with adult stutterers, rather contradictory. Some of the authors have found certain differences in cerebral laterality between groups of stutterers and fluents both for perception of words [2,4] and for perception of syllables [e.g. 3], while others have shown that stuttering children do not suffer any significant abnormality in cerebral processing [e.g. 5]. Though the results of these not numerous studies are rather controversial, a discrepancy in the data seems to be not very dramatic. Even the authors who reported the normal (left hemisphere) mode of speech lateralization for verbal perception in stutterers also marked the specific features of FBA for the latter, being of mostly quantitative character. They are: the lower magnitude of laterality degree and the fewer significant right ear advantages (REA) and more left ear advantages (LEA) as compared to nonstutterers [e.g. 3-5]. So the consideration of ontogenetic peculiarities of syllable hemispheric processing appears to be useful for further exploration of the hypothesis of "aberrant interhemispheric relations."

METHODS

Participants.

A total of 55 stutterers and 52 nonstutterers participated in the study. The age of subjects ranged from 4 to 16 years. and they were divided in 6 age subgroups. The stuttering subjects were selected from speech therapy programs of the City Children Hospital and Speech Pathology Department of St.-Petersburg Institute of Ear, Throat, Nose and Speech Diseases, where they were receiving treatment for their stuttering. Normal subjects were selected from ordinary kinder garden and ordinary school. The stuttering severity determined by a physician had moderate or severe ratings. The stuttering subjects met the following selection criteria: (1) right-handed according a brief handedness test based on Oldfield Handedness Inventory [6]; (2) without traumatic cerebral injures; (3) with normal hearing according tonal audiometry for the frequencies 0,5-4 kHz (4) of average abilities and school achievement. (5) They had no previous research experience of such a kind. All stuttering children were matched with nonstuttering of the same age and other selection criteria. There were approximately equal numbers of girls and boys within each age level.

Stimuli and Procedure

The dichotic listening test has been used to reveal the interhemispheric relations. It was composed of 60 pairs of senseless CVC syllables. The experimental program consisted of 5 blocks, comprised of 4 trials. In the first block each trial consisted of one pair of syllables; in the second block each trial consisted of two pairs and so on up to the fifth block, in which trial consisted of five pairs of syllables. An interval of 20 seconds was left between trials for subject's response. The subjects were tested individually in a sound attenuated room. The audio tape was played to subjects at 60 dB SPL. The test items were presented to subjects through lightweight earphones from a reel- to-reel stereo tape recorder. The earphones were reversed to counteract any imbalance in the channels after each 5 blocks. The task of the subject was to identify the dichotically presented pairs of syllables.

The younger children told their identifications to the experimenter who put them in the response sheet and the children of 8-16 years old wrote their answers in the response sheet by themselves.

Analysis

Analysis of laterality was carried out. A lateralization degree (LD) was measured by the coefficient of asymmetry (Cas). Cas was derived for each subject using the widespread formula $Cas = 100(R-L/R+L)$, where R (or L) is the number of stimuli identifications correctly reported from the right (or left) ear. Using this index Cas values of less than 0 indicate LEA in a given task, and Cas values of greater than 0 indicate REA. Values of 0 indicate no ear difference. The mean Cas scores for each age subgroup of subjects of both types were calculated. Then to test the statistical significance of the means T-test was used.

RESULTS

The results obtained show the greater magnitude of LD in normal children as compared to stutterers, but for the age subgroup of the 4-5-year-olds, where significant difference in LD between stutterers and nonstutterers has not been found. The LD is dependent on the age of children of both subject groups. Normal children in the age range of 4-7 years old demonstrate the decrease of absolute value of asymmetry coefficient. Beginning from 8 years old the Cas increases, achieving its maximal value in the 12-year-olds. Stutterers demonstrate the increase of absolute value of Cas in the age range of 4-7 years, then the decrease of it between 7 and 8 years and then as in normals, beginning from the 8 years one can observe the increase of LD with its maximum in 12-13-year-olds. Then in 14-16 year-olds, in both normals and stutterers, the Cas decreases until values, similar to those, obtained for adult subjects [7]. Thus the results show the similarity of LD dependence on the age both in normal and stuttering children.

The analysis of the direction of lateralization discovers the negative values of Cas both for normals and stutterers in the age range of 4-7 years. That is in this age the LEA or right hemisphere dominance is observed. The qualitative change of lateralization takes place between 7-8-9 years in both subject groups. The change of sign in the age of 8 years demonstrates the shift of perception advantage from left to right ear and the REA or left hemisphere dominance remains in all the subsequent age subgroups both for stutterers and fluents.

The advantage of the ear is not absolute because the children of all age subgroups, both normals and stutterers, are shown to be divided in two parts: with the REA and the LEA. The number of children with the REA increases in the age developmental

course. The stuttering children of the 6-7 years old is the exception: in this subgroup the number of children with the REA (13%) decreases as compared to 4-5-year-olds (30%) and the number of children with the LEA increases.

The ear preferences in the subsequent age subgroups both in normals and stutterers are rather stable: the REA is observed in 65-75% , and the LEA - in 25-35% of children.

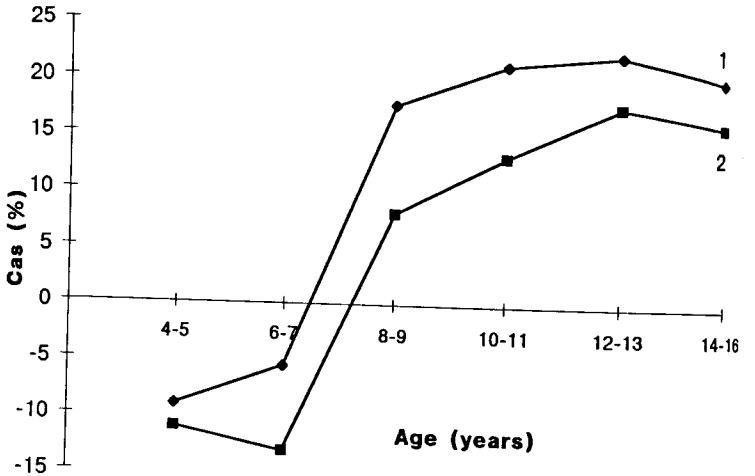


Figure. Changes in cerebral asymmetry over age in two subject groups: nonstutterers (1) and stutterers (2).

DISCUSSION.

The results of the present study showed that both normal speaking and stuttering children evidenced similar character of the FBA formation in ontogenesis for perception of syllables. However, some differences were found to exist between the two groups regarding the level of left-hemisphere dominance and the number of children having atypical processing of dichotic syllables and the development of these variables in the age course.

The results obtained for the stuttering children are consistent with the previous literature in which quantitative differences in the age development of functional specialization of hemispheres for the verbal perception have been found [2-4]. The left-hemisphere superiority has been discovered for stutterers beginning from the age of 8 years. The heterogeneity observed in the group of stutterers concerning the number of subjects demonstrating left-or right-ear preferences is congruent to some extent with the studies of some other authors [e.g. 2,3]. The younger stutterers (6-7-year-olds) manifest fewer significant right-ear preferences than the older ones and they differ significantly in this index from their nonstuttering counterparts. This fact is consistent with the above mentioned studies, but the approximately equal right-ear preferences obtained for stutterers and nonstutterers, beginning from the 8-year-olds are in some disagreement with them. These authors [2,3] reported such data only for the older subgroup.

The present study indicate the similar character of the formation in ontogenesis of interhemispheric relations for syllable perception in stutterers and normals. It allows to suggest that stuttering might be connected not only with the disturbance of the segmental analysis, that is inherent to the left-hemisphere work strategy, but also with the activity of the right hemisphere. The literature suggests that normal right-handed children exhibit the LEA, i.e. right hemisphere specialization for speech emotional information [8,9]. It has been found also, that in perception of emotional information an inversion of dominance, (i.e. left-hemisphere dominance for perception of emotions in speech), in stuttering children as compared to nonstutterers takes place and remains as the main peculiarity of the FBA in adult stutterers [4]. Thus, the stutterers have been found to demonstrate the qualitative difference in cerebral processing of emotional speech information as compared to normals while such difference has not been revealed for both word and syllable cerebral processing. This may suggest the functions deficit of right, but not left hemisphere in stutterers. The supposed left-hemisphere overload [10] might be caused by this dysfunction. Since speech is a complex performance, composed of multiple components, processed in different hemispheres [4,11] such overload might induce a disturbance in hemispheric competition [12] and as a result, the impaired speech performance - stuttering.

REFERENCES.

[1] Travis, L. (1978), " The cerebral dominance theory of stuttering. 1931-1978", *Journal of Speech and Hearing Disorders,* vol. 43, pp. 278-281.

[2] Sommers, R., Brady, W.,& Moore, W. (1975), "Dichotic ear preferences of stuttering children and adults.", *Perceptual and Motor Skills*, vol. 41, pp. 931-938.

[3] Blood, G. (1985), " Laterality differences in child stutterers: heterogeneity, severity levels, and statistical treatments.", *Journal of Speech and Hearing Disorders*, vol. 50, pp. 66-72.

[4] Zaitseva, K., Miroshnikov, D., Dmitrieva, E. (1991), " Principle of parallel processing by brain of various kinds of speech information.", *Sensory Systems,* vol.5, pp. 105-112.

[5] Gruber, L., & Powell, R. (1974), "Responses of stuttering and nonstuttering children to a dichotic listening task.", *Perceptual and Motor Skills,* vol. 35, pp. 263-264.

[6] Oldfield, R. (1971), "The assessment and analysis of handedness: The Edinburgh Inventory.", *Neuropsychologia*, vol. 9, pp. 97-113.

[7] Zaitseva, K. (1991), "The lateralization peculiarities of speech structural elements perception in stutterers.", *Human Physiology*, vol. 17, pp. 18-22.

[8] Morozov, V., Dmitrieva, E., Zaitseva, K., Suhanova, N. (1983), "Age peculiarities of human perception of emotions in speech and singing.", *Journal of Evolutionary Biochemistry and Physiology*",vol. 19, pp. 289-292.

[9] Saxby, Z. & Bryden .(1984), "Left-ear superiority in children for processing auditory emotional material.", *Developmental Psychology*, v.20, pp. 72-80.

[10] Lohov, M., (1988), "Interhemispheric asymmetry in mechanisms of nonaphasic disruptions of speech functions.", *Human Physiology*, v.14, pp. 38-42.

[11] Perkins, W., Kent, R., & Curlee, R. (1991), "A theory of neuropsycholinguistic function in stuttering.", *Journal of Speech and Hearing Research*, vol. 34, pp. 734-752.

[12] Webster, W. (1986), "Neuropsychological models of stuttering-II. Interhemispheric interference.", *Neuropsychologia*, vol. 24, pp. 737-741.

COMMUNICATIVE SUITABILITY OF STUTTERED SPEECH

R. van Bezooijen and M.C. Franken***
** University of Nijmegen, Nijmegen, The Netherlands*
*** Academic Hospital, Rotterdam, The Netherlands*

ABSTRACT

This study investigated the merits of the concept of communicative suitability, i.e. judged adequacy of speech for use in everyday communicative situations, for assessing the quality of stuttered speech. General acceptability was also judged. Stutterers, non-stutterers, and speech therapists served as judges. Communicative suitability seems a promising criterion to realistically evaluate speech quality.

INTRODUCTION

Various methods have been developed to normalize speech fluency of stutterers. Some therapies use so-called fluency enhancing techniques, which affect prosodic and temporal aspects of speech. A widely used criterion for assessing the resulting speech quality has been judged naturalness (e.g. [1,2]). It appears that fluency shaping therapy changes unnatural sounding stuttered speech into unnatural sounding stutter-free speech. However, it is difficult to evaluate this finding. How exactly should the rather abstract and global concept of naturalness be interpreted and translated to suitability of speech for use in everyday life with all its variation in communicative settings, communicative goals, and types of communicators? And to what extent do judgments from "ordinary" people, not involved in problems of stuttering, differ from those given by stutterers and speech therapists specialized in stuttering?

The main goal of our study, then, was to try and develop an alternative, more sociolinguistically based approach to the evaluation of stuttered speech and explore the merits of the concept of communicative suitability, i.e. judged adequacy of speech for use in everyday communicative situations. Three questions were asked:

(1) Do suitability judgments vary as a function of the situation?
(2) Do suitability judgments of stutterers, speech therapists, and non-stutterers differ?
(3) How do suitability judgments relate to general acceptability?

METHOD

Speakers were 10 stutterers and 10 non-stutterers. The 10 stutterers took part in the Dutch adaptation of the Precision Fluency Shaping Program [3]. They were recorded three times: pre-treatment, immediately after treatment ("post-treatment") and six months after treatment ("follow-up treatment"). All were males, of varying ages and from varying educational backgrounds. Many had a regional accent. The 10 non-stutterers, matched for sex, age, education, and accent with the 10 stutterers, served as distractors and as a reference. The stimuli for the judgment experiment consisted of 45 sec semi-spontaneous speech samples. They were presented to three groups of each 17 listeners: (1) "ordinary", non-stuttering adults, (2) speech therapists specialized in stuttering, and (3) stutterers involved in stuttering modification therapy [4]. The 51 judges rated suitability scales (1=completely unsuitable, 10=perfectly suitable) for communicative situations varying in (1) the setting (**private** versus **public** domain), (2) the number of persons spoken to (**single** versus **multiple** interlocutor), (3) the relation to the person spoken to (**known** versus **unknown** interlocutor). and (4) communicative function (**social** versus **informative**). Plausible combinations of these four factors resulted in the ten communicative situations listed below, ordered from most informal to most formal. Uneven numbers refer to situations stressing the **social** function, even numbers to situations stressing the **informative** function, except for 9 and 10, where the distinction could not be made.

+ private, + single, +known
1. talking about everyday events with a friend
2. telling a housemate about one's new job

+ private, - single, + known
3. chatting with housemates during a party game
4. giving a speech at a family celebration

- private, + single, + known
5. making conversation with a friend in the train
6. ordering bread from the baker around the corner

- private, + single, - known
7. getting into contact with a stranger on the bus
8. asking a bypasser for directions

- private, - single, - known
9. instructing a group at a dancing school
10. giving a lecture to a newly founded professional association

After judging the suitability of the speech sample for each situation, the listeners rated the general acceptability (1=completely unacceptable, 10=perfectly acceptable) of each speech sample on a separate, eleventh scale, not tied to a specific situation.

The reliability of the ratings was assessed, separately for the 11 scales and the 3 listener groups, by means of Cronbach's alpha. All alpha's exceeded .95, which shows that all three groups of listeners agreed on the relative suitability of the speech samples for use in various communicative situations and on their general acceptability.

RESULTS AND DISCUSSION

Separate analyses of variance were carried out for the suitability ratings and the acceptability ratings. The level of significance was set at 5%. We will only present and discuss significant effects directly bearing upon the three questions asked in the introduction.

Do suitability judgments vary as a function of the situation?

The factor "situation" had a significant effect on the suitability ratings, explaining as much as 27% of the variance. This means that judges strongly differentiated their judgments depending on the specific characteristics of the communicative situation in which the speech was supposed to be used. The ratings for the ten communicative situations are listed in Table 1. The data show that the order of judged suitability corresponds with degree of formality: speech was judged least suitable for the most formal situations 9 and 10 and most suitable for the least formal situation 1. The other situations, with intermediate degrees of formality, received intermediate ratings of suitability. This holds for the stuttered speech at different stages of treatment as well as for the reference speech. So, judges consistently place higher demands upon the quality of speech as the situation is more public, involves a greater number of less well-known interlocutors, and focusses more on information transmission.

We think that the variation in the height of the suitability ratings has to do both with linguistic and extra-linguistic factors. At the linguistic level, intelligibility can be assumed to play a role. That is, the typical charac-

teristics of formal communicative situations, e.g. high information density, listener(s) unfamiliar with (the speech style of) the speaker, large distance between listener(s) and speaker, require speech that is clearly enunciated, without deviant and unpredictable properties. This would be a functional reason. At the extralinguistic level, there are social conventions, which dictate, for example, a particular style of clothing (tie, suit) but also a particular style of speaking, represented by the standard variety (RP, standard Dutch), without pathological or dialectal deviations.

Do suitability judgments of stutterers, speech therapists, and non-stutterers differ?

There was a significant effect of the factor "type of judge", accounting for 9% of the variance. The mean suitability ratings, averaged over the four types of speakers, given by the stutterers, therapists, and ordinary people were 5.9, 5.6, and 4.6, respectively. So, the data reveal that overall ordinary people are considerably less tolerant in their judgments than therapists, who in turn are somewhat stricter than stutterers. This

Table 1. Mean judged suitability (1=completely unsuitable, 10=perfectly suitable) of stuttered speech (pre-, post-, follow-up) and reference speech for ten communicative situations. In the last column overall means, which have served as the basis for the ordering from lowest to highest suitability.

No	Context	Pre	Post	Fol.	Ref.	All
9	group/instructions	2.4	3.0	3.5	5.7	3.6
10	association/lecture	2.4	3.1	3.6	5.8	3.7
4	family/speech	3.2	4.1	4.4	6.7	4.6
7	stranger/bus	3.9	4.9	5.2	7.3	5.3
8	bypasser/directions	4.2	5.4	5.6	7.5	5.7
6	baker/bread	4.5	5.7	5.8	7.6	5.9
5	friend/train	4.8	5.7	6.0	7.7	6.0
3	housemates/party game	5.0	5.7	6.1	7.7	6.1
2	housemate/job	5.1	5.8	6.1	7.7	6.2
1	friend/everyday events	5.4	6.0	6.4	8.0	6.4
All		4.1	4.9	5.2	7.2	

pattern emerged for each speaker group separately as well. Perhaps therapists and stutterers are less sensitive to deviations in speech as a result of repeated exposure to deviant speech. Apparently and quite remarkably, the differential sensitivity would hold not only for pathological deviations such as stutters, but for dialectal aspects of speech as well (as mentioned under Method, many reference speakers had regional, non-standard accents). Also, therapists and stutterers may be milder because they know from experience how difficult it is get rid of deviant speech characteristics. The difference between ordinary people on the one hand and therapists and stutterers on the other holds particularly for the less formal situations (the interaction between "situation" and "type of listener" accounts for 2% of the variance).

How do suitability judgments relate to general acceptability?

The mean general acceptability ratings closely resemble the results for the suitability data averaged over ten communicative situations. The acceptability ratings of 3.6, 4.5, 4.9, and 6.9 for the pre-treatment, post-treatment, follow-up treatment, and reference speakers can be compared to the suitability ratings of 4.1, 4.9, 5.2, and 7.2. The acceptability ratings of 5.6, 5.0, and 4.3 for the stutterers, therapists, and ordinary people can be compared to the suitability ratings of 5.9, 5.6, and 4.6. Also, for both types of judgments similar patterns of significant effects were found. The grand mean of the general acceptability ratings is 5.0, which constitutes the exact midpoint of the suitability continuum as used by the judges, with the extremes 3.6 and 6.4.

CONCLUSION

Communicative suitability appears to be a useful approach to assessing speech quality since it does justice to everyday reality where different demands are placed upon speech depending on communicative settings, interlocutors, and goals. It is further shown that it is dangerous to generalize judgments from persons used to stuttering, such as speech therapists and stutterers, to the type of people stutterers will usually interact with in everyday life. The norms of the latter appear to be stricter. These findings should be taken into account when evaluating the communicative consequences of stuttering and the effects of stuttering therapy. Finally, general acceptability appears a useful scale to measure "average" suitability. Further research is needed to examine the relationship between general acceptability and naturalness.

ACKNOWLEDGMENT

The contribution by the first author has been made possible by a fellowship from the Royal Netherlands Academy of Arts and Sciences. The authors wish to thank Anneke Olierook en Margot Spanhoff for collecting the data.

REFERENCES

[1] Runyan, C.M., Bell, J.N., & Prosek, R.A. (1990), "Speech naturalness ratings of treated stutterers", *Journal of Speech and Hearing Disorders*, vol.55, pp.434-438.

[2] Franken, M.C., Boves, L., Peters, H.F.M., & Webster, R.L. (1992), "Perceptual evaluation of the speech before and after fluency shaping stuttering therapy", *Journal of Fluency Disorders*, vol.17, pp.223-242.

[3] Webster R.L. (1980), "Evolution of a target-based behavioral therapy for stuttering", *Journal of Fluency Disorders*, vol.5, pp.303-320.

[4] Peters, T.J., & Guitar, B. (1991), *Stuttering. An integrated approach to its nature and treatment*, Baltimore: Williams & Wilkins.

PHONOLOGICAL SIMILARITY EFFECTS IN CANTONESE WORD RECOGNITION

Anne Cutler (MPI for Psycholinguistics, Nijmegen)
and Hsuan-Chih Chen (Dept. of Psychology, Chinese University of Hong Kong)

ABSTRACT

Two lexical decision experiments in Cantonese are described in which the recognition of spoken target words as a function of phonological similarity to a preceding prime is investigated. Phonological similarity in first syllables produced inhibition, while similarity in second syllables led to facilitation. Differences between syllables in tonal and segmental structure had generally similar effects.

INTRODUCTION

The vocabulary of a language contains hundreds of thousands of words, all made up of a very small number of building blocks: phonemes. The phonemic inventory of a language is reckoned not in the hundreds or thousands of items but in the tens (Maddieson [1] lists phoneme inventory ranges from 11 to 141, with a median of 28-29). Thus most words have close phonological neighbours - other words whose sound pattern is only minimally different. The process of spoken-word recognition involves distinguishing a heard word from other words it might possibly be, and recent research in this area has supported the proposal that recognition involves a process of active competition between phonologically similar words (see McQueen, Cutler, Briscoe and Norris [2] for a review). Thus recognition is clearly affected by the presence of phonologically similar words in the vocabulary. However, whether spoken-word recognition can be affected by prior processing of phonologically similar words is as yet uncertain.

Only few studies have examined the recognition of spoken words as a function of immediate prior auditory presentation of another word similar in sound; from these differing, in part apparently incompatible, results have emerged (in fact, cross-modal studies have also produced conflicting results, but this large literature is beyond the scope of the present report). For example, Slowiaczek and Hamburger [3], using a word repetition task in English, found that overlap of initial phoneme between prime and target (e.g. *smoke-still*) facilitated response latency, but overlap of three phonemes (*stiff-still*) inhibited it. Radeau, Morais and Dewier [4], also using word repetition, but in French, found only inhibition effects regardless of amount of overlap. They found similar interference also with the lexical decision task. Emmorey [5] used lexical decision in English and found facilitation for certain pairs sharing the final syllable (e.g. *tango-cargo*). Zhou [6] (to our knowledge the only study of this kind in a Chinese language) presented listeners with pairs of bisyllabic Mandarin words, and found (like [5]) facilitation if the pairs shared the final syllable, but (similarly to [3] and [4]) inhibition if they shared the first syllable.

Mandarin is a tone language, with a four-tone system; its phonological inventory is small and there is very extensive homophony. We here report two experiments, in many respects similar to Zhou's, examining spoken-word recognition in Cantonese as a function of phonological similarity between the response target and a preceding word. Cantonese is also a tone language, but has a much more complex tonal inventory (a nine-tone system, of which three "glottalised" tones occur only on certain syllable types), as well as a more varied phonological structure than Mandarin.

EXPERIMENT 1

Materials

96 pairs of bisyllables were constructed; each pair comprised prime plus target (response item). Half of the 96 targets were nonwords. Of the 48 real-word targets, one quarter had a phonologically and semantically unrelated prime (a baseline control condition; e.g. the words for *scarf-tomato*). A further quarter had a semantically but not phonologically related prime (e.g. *piano-guitar*; a further control to ensure that conditions for inter-word effects had been met). The remaining 24 items had a phonologically related (but semantically unrelated) prime; prime and target shared initial syllables but differed in second syllables. In 12 items, the second syllable differed in tone (e.g. *ji6liu4* "treatment" - *ji6liu5* "feed"); in 12, it differed in rime (e.g. *to4fa1* "peach flower" - *to4fool* "butcher"). In all cases the overlapping syllables were morphologically different.

Subjects and Procedure

32 students at the Chinese University of Hong Kong were tested individually in the experiment; all were native speakers of Cantonese with no reported hearing impairment. 16 subjects in the priming group were instructed to listen to the pairs of bisyllables and to decide, as quickly as possible, whether or not the second was a real word of Cantonese, and to signify their response by pressing one of two response keys (labelled YES and NO) in front of them. For the remaining 16 subjects the targets occurred without preceding primes and subjects were instructed to make a lexical decision for each word. The stimuli, which had been spoken by a female native speaker of Cantonese and digitised (at a sampling rate of 22 kHz), were presented over Sound MD-802A headphones at a comfortable listening level. Prime-target ISI was 400 ms. Stimulus presentation and response timing were controlled by a Macintosh IIsi computer.

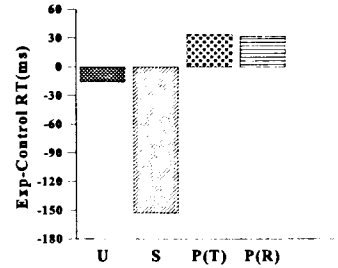


Figure 1. Mean differences between experimental (priming) and control (no-priming) group RT (measured from word offset) for the four prime conditions of Experiment 1. U: Unrelated prime; S: Semantically related prime; P[T]: Phonologically related prime (tone difference in second syllable); P[R]: Phonologically related prime (rime difference in second syllable).

Results and Discussion

Each missing data point was replaced by the mean response time (RT) for the same subject in the same condition. Analyses of variance across subjects and items revealed significant differences between the four conditions and an interaction of the condition factor with the priming/no-priming group factor; this interaction was examined via *t*-tests (again across subjects and items separately) comparing groups in each condition. Figure 1 shows the between-group differences. In the unrelated-prime condition, the groups differed by just 16 ms, a statistically insignificant difference ($t1$ and $t2 < 1$). Thus there are no inter-group differences *per se*. With semantically related primes, there was a highly significant facilitation for the priming group ($t1$ [15] = 3.62, $p < 0.003$; $t2$ [22] = 3.16, $p < 0.005$). Thus the experiment is sufficiently sensitive to exhibit priming where this occurs. However, neither phonologically related-prime condition showed facilitation effects; in contrast, in both groups there was, rather, inhibition, i.e.

RTs were slower than in the baseline condition, although this effect did not reach significance in either condition.

Thus recognition of a spoken Cantonese word is not facilitated, and may indeed be inhibited, by having just heard another word beginning in the same way. Alterations of rime and of tone between prime and target have exactly parallel effects. The pattern of findings parallels reported results from English and French, when the overlap encompasses several phonemes, and data from Mandarin. It is an effect which is consistent with competition-based models of spoken-word recognition (see [2]) in which simultaneously activated words may inhibit one another's recognition.

EXPERIMENT 2

Materials, Subjects and Procedure

The materials were constructed exactly as in Experiment 1, except that in the 24 phonologically related pairs the difference between prime and target occurred in the first rather than the second syllable. Again, the difference involved tone in 12 pairs (e.g. *to4wa6* "picture" versus *to2wa6* "dialect") and rime in the other 12 pairs (*si6yip6* "career" versus *sue6yip6* "leaf"); the second syllables of prime and target were always phonologically identical and morphologically different. 32 subjects from the same population participated in the experiment; none had taken part in Experiment 1. The procedure was as in Experiment 1.

Results and Discussion

The data were analysed as for Experiment 1; Figure 2 shows the between-group differences for each condition. Again there was a significant main effect of condition and a significant interaction between conditions and groups in the analysis of variance; again, the between-group difference in the unrelated-prime condition (6 ms) was not significant (both $t1$ and $t2 < 1$), but there was a significant facilitation effect in the semantically related-prime condition ($t1$ [15] = 3.27, $p < 0.005$; $t2$ [22] = 4.52, $p < 0.001$). For the two phonologically related conditions, the results of Experiment 2 differed from those of Experiment 1; there was facilitation instead of the inhibition observed previously. When the first syllable differed in rime, the difference between groups was significant ($t1$ [15] = 3.12, $p < 0.007$; $t2$ [22] = 3.67, $p < 0.001$). For the condition in which the first syllable differed in tone, the difference between groups was only half as large, and did not reach our criterion of significance ($t1$ [15] = 1.12; $t2$ [22] = 1.47).

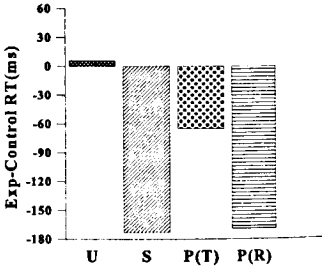


Figure 2. Mean differences between experimental (priming) and control (no-priming) group RT (measured from word offset) for the four prime conditions of Experiment 2. U: Unrelated prime; S: Semantically related prime; P[T]: Phonologically related prime (tone difference in first syllable); P[R]: Phonologically related prime (rime difference in first syllable).

Thus recognition of a spoken Cantonese word appears to be facilitated by having just heard another word ending in the same way. Again, the results from Cantonese parallel those from English [5] and Mandarin [6]. Again, alterations of rime and of tone between prime and target appear to pattern similarly, although when the first syllable differed in tone less robust facilitation was observed than when the first syllable differed in rime.

GENERAL DISCUSSION

Our two experiments on phonological similarity in spoken-word recognition in Cantonese motivate two general conclusions. Firstly, effects of phonological overlap between two bisyllabic words differ as a function of whether the overlap is located in the words' first or second syllable. Secondly, differences between syllables in tonal and in segmental structure have in broad outline similar effects.

Apparent contradictions between previous studies may therefore reflect differences between the types of phonological overlap manipulated. In English [3] and French [4], inhibition occurs when successively presented words overlap in the first few phonemes; the same result occurs with an initial syllable overlap in Mandarin [6] and, as the present study shows, in Cantonese. In this respect there may be little cross-linguistic difference. The findings so far do not allow us to decide whether the overlap must involve integral syllables or merely any word-initial portion. We suggest that the latter is the simpler option. Certainly it is fully compatible with our preferred explanation of the inhibition effect, namely that competition between simultaneously activated word candidates inhibits recognition of a target word.

Facilitation effects of phonological overlap in word-final position have also now been observed in more than one language: English [5], Mandarin [6] and, in the present study, Cantonese. In none of these experiments was this facilitatory effect due to morphological priming (recall that, like Zhou, we exploited the unique properties of Chinese languages to use prime/target overlaps which were syllabic but not morphological); however, it would be interesting to examine whether it might have its origin in a processing strategy designed to exploit the overwhelming tendency across languages for affixes to occur predominantly in suffix position [7].

In Zhou's experiments, and in the studies in European languages, the non-overlapping portions of the prime and target pairs were completely different; in our experiments they differed only in tone or in rime. The similarity between the present results and those from other languages thus suggests that either a tonal or a segmental difference is sufficient fully to distinguish between words, i.e. that tone functions analogously to segmental structure in spoken-word recognition. The parallel effects of the tone and rime difference manipulations (inhibition in Experiment 1, facilitation in Experiment 2) further support this conclusion.

REFERENCES

[1] Maddieson, I. (1984), *Patterns of sounds*, Cambridge: Cambridge University Press.

[2] McQueen, J.M., Cutler, A., Briscoe, T. & Norris, D. (1995), "Models of continuous speech recognition and the contents of the vocabulary.", *Language and Cognitive Processes*, vol.10.

[3] Slowiaczek, L.M. & Hamburger, M. (1992), "Prelexical facilitation and lexical interference in auditory word recognition." *Journal of Experimental Psychology: Learning, Memory and Cognition*, vol. 18, no. 6, pp. 1239-1250.

[4] Radeau, M., Morais, J. & Dewier, A. (1989), "Phonological priming in spoken word recognition: Task effects.", *Memory & Cognition*, vol. 17, no. 5, pp. 525-535.

[5] Emmorey, K.D. (1989), "Auditory morphological priming in the lexicon.", *Language and Cognitive Processes*, vol.4, no. 2, pp. 73-157.

[6] Zhou, X. (1992), *The mental representation of Chinese disyllabic words*, Ph.D. Dissertation, University of Cambridge.

[7] Cutler, A., Hawkins, J.A. & Gilligan, G. (1985), "The suffixing preference: A processing explanation.", *Linguistics*, vol. 23, pp. 723-758.

DIFFERENTIAL USE OF TONAL AND SEGMENTAL INFORMATION IN LEXICAL DECODING DECISIONS IN CANTONESE

Rosemary Varley, Sandra P. Whiteside & Yuet Yee Yim
Speech Science, University of Sheffield, Sheffield S10 2TA, United Kingdom

ABSTRACT

The study investigated the perceptual processing of tonal and segmental information in Cantonese. The hypothesis that the processing of tonal information was more robust was tested by presenting 40 young adult subjects with stimuli masked by white noise. Two experimental conditions were developed: free-choice and forced-choice. In both conditions, the results showed a significant primacy of tonal information over segmental information.

INTRODUCTION

The project investigated whether speakers of a tone language (Cantonese) show a preferential use of tonal over segmental information in lexical decoding. Acquisitional research supports the early acquisition of tonal information over segmental [1], and anecdotal evidence from L2 learners of Cantonese suggests that whereas segmental errors can be assimilated by native Cantonese speakers, tonal errors have more profound effects on comprehension. Specifically, where the L2 learner makes a tonal error, the listener appears to make an assumption that the tonal information is as given, but that the segmental form is at fault. Evidence from brain-damaged speakers also suggests that segmental information is more vulnerable to disruption than tonal [2].

These observations suggest a primacy in the processing of tonal information over segmental. This study subjected this hypothesis to experimental test by placing listeners in marginal listening conditions and observing whether there was preferential use of tonal or segmental information.

The tonal system of Cantonese consists of six contrastive tones [3]. There are three level tones, differing in pitch height; high-level (tone 1), mid-level (tone 3), and low-level (tone 6), and three contour tones; high-rise (tone 2), low-rise (tone 5), and low-fall (tone 4). There are also three clipped or entering tones, but these are regarded as allophonic variants of the high, mid, and level tones as they occur only in CVC syllables where the final consonant is /p, t, k/. Tones which share similar contours or starting heights have been shown to be confusable [3, 4, 5].

METHOD

In the first stage of the study, 47 triplets of words were identified. Each triplet included a stimulus word, a tonal minimal pair (same segmental form as the target, but different tone) and a segmental minimal pair (same tone, different segmental form). For example, for target /kun3/, tonal pair /kun1/, segmental pair /pun3/. The minimal pairs were all highly confusable with the target item. The tonal pairs involved tones 1-3, 2-4, 3-4, and 2-6. (There is also a further sub-set of highly confusable tone pairs which include tones 4-6, 3-6, and 2-5, but we were unable to identify triplets including these pairs). Segmental confusability norms are not available for Cantonese and so extrapolations were made from English norms [6]. Segmental pairs involved the following contrasts: /p-k, t-p, ph-th, ph-kh, th-kh, s-f, j-w, j-l, w-l/.

Stimulus Preparation

An adult female native speaker of Cantonese (PY) read the list of target stimuli in a sound proof room. This list also included an introductory phrase (IP) and a prompting phrase (PP). The recording was made on a SONY TCD-D3 DAT recorder. The word stimuli, IP, and PP were digitised using a KAY Computerised Speech Lab (CSL) model 4300 at a sampling rate of 20 kHz. The amplitudes of the word stimuli were then scaled to a mean of 53 dB using the CSL.

These lists were used to record the test stimuli with white noise from the CSL onto a SONY DAT machine. The white noise was sampled at 20kHz and had an amplitude of 62 dB, to give a mean signal to noise ratio value of -9dB. The speech signal was output to the left channel and the white noise to the right channel. The two signals were then output through a mono channel which was then converted into a stereo channel so that both the white noise and the speech stimuli would be played back in both ears. This was done to control for any bias effects which could have resulted from selective attention listening strategies.

Each set of stimuli included two practice items. An IP was included at the start of both the practice items and the listening items. Each stimulus was preceded either by an IP (at the start) or a PP. This was done to draw the attention of the listener to the stimuli and to familiarise the listeners with the speaker's voice. This is particularly important in tonal perception, where a decision regarding the identity of a tone is relative to the speaker's indexical pitch.

The stimuli were arranged into four random orders. As subjects completed two experimental conditions, each subject received the stimuli in different random orders across conditions, and within conditions, half of the subjects received stimuli in one random order, and the other half in a second.

Subjects

Forty young adult subjects (20 male, 20 female) completed the task. Ages ranged from 19 to 26 (mean 21.8, *sd.* 1.68). All subjects had Cantonese as their L1 and their home language. They reported no hearing problems. All subjects were students in tertiary education.

Procedure

Two experimental designs were developed to test the hypothesis: free-choice and forced-choice. In the free-choice design subjects were required to listen to a stimulus word and to write down what they thought they had heard. Responses were examined for the relationship between the stimulus and the character written down and were analysed into the following categories: stimulus; tonal-response (same tone, different segmental form); segmental response (same segmental form, different tone); and 'other' (for example, both the tone and the segmental form differed from that of the stimulus). In the forced-choice design, subjects were played a distorted syllable and were required to select from two choices the syllable that they thought matched the one they had heard. The two response choices were the segmental minimal pair and the tonal minimal pair, but the actual stimulus word was not given as a response choice. Preferential use of tonal or segmental information was noted.

All subjects were tested individually in sound-damped booths. Test tapes were played on a Sony Stereo Cassette-Corder (TC-D5M), and subjects listened through AKG Dynamic System (K135) headphones. The test procedure took approximately 30 minutes for each subject. The order of the free and forced-choice tasks was not counterbalanced. The free-choice task was presented first so that given responses for the forced-choice task did not influence response in the free-choice condition.

RESULTS

Due to occasional errors in the data collection phase, not all subjects made decisions on all 47 stimuli. Data were therefore converted into percentage scores to permit comparisons.

Free-choice condition

Table 1 shows the percentages of responses classified into each category.

Table 1. Percentage of free-choice responses by category and sex.

	Stim.	Tonal	Segm.	Other
Male mean	41.48	42.98	6.52	8.97
sd.	7.74	7.56	4.07	3.20
Fem. mean	42.20	42.87	5.39	9.51
sd.	10.12	6.56	3.21	6.91

The pattern of performance between male and female subjects was very similar and therefore data from the two groups were combined for subsequent

analysis. The data show that despite the distortion of the stimulus word, subjects were able to decode the tonal information of the syllable on approximately 84 percent of trials. Segmental information was more vulnerable to disruption and was accurately decoded on approximately 47 percent of trials. It was rare for segmental information to be decoded but for the tone not to be correctly perceived.

Differences between response categories were compared with Wilcoxon signed-rank tests. Comparisons were non-significant between target and tonal responses. All other comparisons were significant (target score > segmental and other scores $p>0.01$ (T = 0 and 1 respectively), tonal scores > segmental and other scores $p>0.01$ (T = 0), and other score >segmental score $p>0.01$ (T = 160).

Forced-choice condition

Table 2 shows the percentages of tonal or segmental responses made by subjects on the forced choice condition.

Table 2. Percentage of tonal and segmental responses on the forced-choice task, by sex.

		Tonal	Seg.tal
Male	mean	62.65	37.35
	sd.	21.76	
Female	mean	68.26	31.75
	sd.	20.74	

Inspection of the data suggests that again there is no obvious difference between male and female subjects and therefore their results were combined. Subjects preferentially used tonal information when placed in the forced-choice situation. Comparison with a Wilcoxon-signed ranks test revealed the difference to be significant (T = 126.5, $p<0.01$). It is noticeable, however, that subjects utilised a segmental-decoding strategy more often than one might predict from the free-choice situation. Certain individuals showed a marked preference for a particular response type. For example, 13 subjects chose the tonal response on over 80% of trials; strong segmental response choices were less evident, with only two subjects producing 70% or more segmental responses. Spearman correlation coefficients were calculated between tonal responses on the free and forced choice paradigms, and also on segmental responses across the two paradigms. Both figures for *rho* were very low and non-significant (tonal choice *rho* = 0.196, segmental *rho* = 0.001). The higher segmental responses may therefore reflect the effects of chance, and also in the more constrained forced-choice task, the results of selective attention strategies.

DISCUSSION

The results of the perceptual experiments revealed findings consistent with the initial hypothesis that the decoding of tonal information showed primacy over segmental decoding. Subjects were able to perceive the correct tonal information on over 80 percent of trials on a free-choice task, whilst the segmental information was correct on less than half of the trials. Similarly, when placed in a situation where the listener had to choose a character which matched a distorted stimulus, subjects showed a preference for selecting the response choice which matched the target in tone, but not in segmental form.

There are a number of possibilities in accounting for this result. The first is, in line with our experimental hypothesis, that the perception of tone is in some way primary in the recognition systems of tone language users. But primacy here does not mean temporal primacy. Segmental and tonal perceptual processes are likely to happen in parallel: one could not envisage a perceptual system which 'held' on segmental information until a tonal decision had been made. It may be that the processing of tonal information is a more rapid process. On a mathematical basis, there are only six contrastive tones in Cantonese, but many more possible segments and their combinations. As the tone paradigm is smaller, then decisions will be made more quickly. This might then suggest that the phonological lexicon will be organised around tone - the rapid tonal decision permitting a narrowing down of the possibilities in matching an input to a stored representation.

A second possibility is that our results are more simply an artefact of our experimental design. There are two possibilities here. The first is that our segmental minimal pairs were more confusable than the tonal ones. Whereas it was possible to select or extrapolate from English norms highly confusable consonants, we were unable to identify triplets including the very highly confusable tone pairs. The absence of such triplets may be an example of avoidance of contrasts in a single language which pose heavy demands on perceptual systems.

A second explanation of the results is that the white noise masking is more likely to distort the acoustic information necessary for segmental (especially consonantal) perception than the lower frequency information involved in tonal discrimination. Preliminary analysis of free-choice errors supported the idea that some errors were an artefact of the experimental design. Consonant errors for example were more frequent than vowel errors: suggesting that the lower frequency information of vowels and diphthongs (and also therefore tones) was more robust in the white noise masking used in this study. It was felt however that all the results could not be totally dismissed as artefacts. The experiment used blanket masking of the signal, but the speech signals varied temporally in amplitude. This meant that the vowel/ diphthong nuclei and formant transitions in particular were most robust to the masking than consonantal features like frication and plosion. Given this we would have expected listeners to use the formant transition patterns to opt for more segmental judgements over tonal judgements. However as was noted above this was not the case and listeners tended to opt for tonal judgements. Furthermore some of the distinctions between the tone pairs used in this study were signalled at the tail-end of the vowel [3], and therefore were subject to the same level of masking as consonant cueing transitional information. Despite this, listeners tended to opt for the tonal decisions. A study on Mandarin Chinese tones [7] found that the perception of tone was robust to various filter conditions. Here listeners were still able to make tone-phoneme identifications with missing acoustic information. This lends further support to our findings which illustrate the primacy of tone in the perceptual systems of tone language users. Further research is planned in this area.

ACKNOWLEDGEMENTS

The authors would like to thank the Department of Speech and Hearing Sciences at the University of Hong Kong for use of their facilities during the data collection phase of this project, and P. Yip for her help in recording the stimuli.

REFERENCES

[1] Tse, J.K.P. (1978), Tone acquisition in Cantonese: a longitudinal case study. *Journal of Child Language* vol. 5, pp. 191-204.

[2] Yiu, E. M. L. (1989), *Tonal Disruption in Chinese (Cantonese) Aphasics*. Unpublished M. Phil. Thesis. University of Hong Kong.

[3] Fok, C.Y.Y. (1974), *A Perceptual Study of Tones in Cantonese*. Hong Kong: University of Hong Kong.

[4] Gandour, J. (1983), Tone perception in Far Eastern languages. *Journal of Phonetics* vol. 11, pp. 149-175.

[5] Varley, R. & So, L. (1995), Age effects in tonal comprehension in Cantonese. In press. *Journal of Chinese Linguistics*.

[6] Miller, G. & Nicely, P. (1955), An analysis of perceptual confusion among some English consonants. *Journal of the Acoustical Society of America*, vol. 27, pp. 338-352.

[7[Stagray, J. R. , Downs, D. & Sommers, R. K. (1992), Contributions of the Fundamental, resolved harmonics, and unresolved harmonics in tone-phoneme identification. *Journal of Speech and Hearing Research*, vol. 35, pp. 1406-1409.

UPWARD F0 TRANSITION IN FALLING-FALLING TONES AND RISING F0 PART IN FALLING-CONVEX TONES

Maocan Lin
Institute of Linguistics, CASS, Beijing, China

ABSTRACT

The results of acoustic analysis and perceptual experiment indicated that the information of tones is mainly carried by the syllabic vowel and its adjacent transition. The upward F0 transition in VCV and VV with falling-falling tones of Standard Chinese is not perceived, because the durations of the upward F0 transition only have 89ms and 60ms in average and it occurs durning the non-voiceless initial and its adjacent transition. The durations of the rising F0 part of F0 in VCV and VV with falling-convex tones of the Chinese dialect of Fuzhou have 167ms and 140ms and it occurs durning the syllabic vowel and its adjacent transition, therefore, it can be perceived.

I. INTRODUCTION

The F0 transition in the intersyllable that the second syllable is with non-voiceless initial was discussed in our paper [1]. Acoustic data from two tone languages were presented to demonstrated that the perceived segmental structure is an important factor in the interpretation of F0 as pitch [2].

In this paper, acoustic analysis and perceptual experiment were done on falling-falling tones in VCV(c=/m, n, l/) and VV of Standard Chinese and falling-convex tones in VCV and VV of the Chinses dialect of Fuzhou to discover why the upward F0 transition is not perceived and the rising F0 part in convex tone is perceived.

II. FALLING-FALLING TONES IN VCV AND VV OF S.C.

In disyllabic utterances with falling-falling tones and a voiced intervocalic segment, the F0 must change from low-ending on the first syllable to high(falling) on the second syllable, the upward F0 transition in the intersyllable being formed.

2.1 F0 and amplitude(Am)

15 disyllabic utterances with falling-falling tones in VCV and VV were uttered by a native male speaker of Beijing Mandarin. A formant transition are formed in the intersyllable. The perceptual boundary of the first syllable and the second syllable with non-voiceless initial was determined with the trancation method[2]. In the Fig. 1.1, the perceptual boundaries were indicated with "a-b". The second syllable, therefore, started with "6".

It can be seen in Fig. 1.1 that the starting-point of the upward F0 transition occurred within "a-b". The duration of the upward F0 transition in the second syllable, However, was counted from the point "b". The magnitude of F0 in the upward transition was about 25Hz. The duration of the upward transition of F0 was 89ms and 60ms in average in VCV and VV, amounting to 38% and 30% of the whole duration of the second syllable, respectively. A Am curve in the second syllable being with flat-topped.

2.2 Carrier of the information of tones in S.C.

In this experiment, the duration of 120ms in each stimulus was selected, because a vowel duration greater than 100ms was required to optimize movement feature perception[3].

It can be seen in Fig. 1.2.1 that the highest sensitivity to falling pitch in the first syllable was stimulus 8(140-260ms), and the stimulus was made from the syllabic vowel and its adjacent transition; the highest sensitivity to the falling-pitch in the second syllable was stimulus 25(480-600), and the stimulus was made from the syllabic vowel and its adjacent transition. However, the sensitivity of the stimulus covering the vocalic-ending in"调"[tiao ||] and the nasal coda in "任"[n ||] were lower than that covering the syllabic vowel; the stimulus covering the vocied fricative initial didn't be identified as falling pitch.

It can be seen in Fig. 1.2.2. that the highest sensitivity to falling pitch in the first syllable was stimulus 4(60-180 ms), and the stimulus was made from the syllabic vowel and its adjacent transition; the highest sensitivity to falling pitch in the second syllable was stimulus 17(320-440ms), and the stimulus was mainly made from the syllabic vowel and its adjacent transition. However, the sensitivities of the stimuli covering the vocalic-ending in "概"[kai ||] and"要"[iao ||]were lower than that covering the syllabic vowel, and the stimulus covering the zero-initial didn't be identified as falling pitch.

III. FALLING-CONVEX TONES IN VCV AND VV OF THE CHINESE DIALEXT OF FUZHOU

3.1 F0 and Am

13 disyllabic utterances with falling-convex tones in VCV and VV were uttered by a native speaker of the Chinese dialect of Fuzhou. It can be seen in Fig. 2.1 that in falling-convex tones in VCV and VV, the starting-point of the rising F0 part in convex tone was synchronized with the second syllable. A Am curve in the second syllable being with pinnacl. The magnitude of the F0 rise in convex tone was about 15Hz, and the durations of the F0 rise in VCV and VV were 167ms and 140ms, amounting to 52% and 49% of the whole duration of the second

syllable, respectively.

3.2 Carrier of the information of tones in the Chinese dialect of Fuzhou

Here, the duration of 140ms in each stimulus was selected. It can be seen in Fig. 2. 2. 1 that the highest sensitivity to falling pitch in the first syllable was stimulus 3(40-180 ms), and the stimulus was mainly made from the syllabic vowel and its adjacent transition; Those that was identified as level pitch covering the turning-point in convex tone was timulus 17(320-460ms), and the stimulus was made from the syllabic vowel and its adjacent transition, too. However, sensitivities of the stimuli covering the vocalic-ending in the first and second syllables were lower, and the stimulus covering the nasal consonant initial didn't be identifed as rising pitch.

It can be seen in Fig. 2. 2. 2 that the highest sensitivity to falling pitch in the first syllable was stimulus 4(60-200 ms), and the stimulus was made from the syllabic vowel and its adjacent transition; Those that was identified as level pitch covering the turning-point in convex tone was stimu-lus 19(360- 500ms), and the stimulus was made from the syllabic vowel and its adjacent transition, too; However, sensitivity of the stimuli covering the last part of the final in the first syllable and the nasal code in the second syllable were lower, sensitivity of the stimulus covering the zero-initial was lower, too.

IV. Conclusion and discusion

1. The information of tones is carried by the syllabic vowel and its adjacent transition, but the formants in the area of the syllabic vowel and its adjacent transition in [|], [iao |], [mau |] and [i |] rapidly change. The Am curves in the area of the syllabic vowel and its adjacent transition in the second syllable rapidly change, too.

2. In VCV with falling-falling tones, the duration of the upward F0 transition in the second syllable was 89ms in average in which the duration of the initial /m, n, l/ was about 60ms. The upward F0 transition in VCV that is not perceived can be interpreted by D. House' theory [3]. The upward F0 transition in VV, However, can't be interpreted by D. House' theory, This is because the duration of the upward F0 transition was about 60ms and the complexty of the spectrogram in the area of the upward F0 transition is not more than that in the area of the syllabic vowel and its adjacent transition.

3. The upward F0 transitions in VCV and VV with falling-falling tones of S.C. are not perceived, because they don't occur durning the syllabic vowel and its adjacent transition, and their durations just have 89ms and 60ms amounting to 38% and 30% of the whole

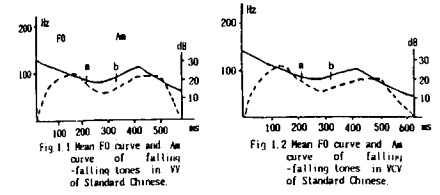


Fig. 1.1 Mean F0 curve and Am curve of falling-falling tones in VV of Standard Chinese.

Fig. 1.2 Mean F0 curve and Am curve of falling-falling tones in VCV of Standard Chinese.

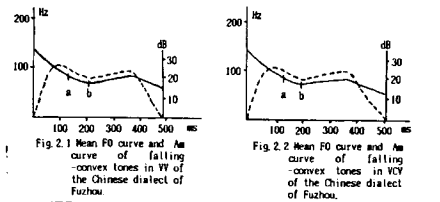


Fig. 2.1 Mean F0 curve and Am curve of falling-convex tones in VV of the Chinese dialect of Fuzhou

Fig. 2.2 Mean F0 curve and Am curve of falling-convex tones in VCV of the Chinese dialect of Fuzhou.

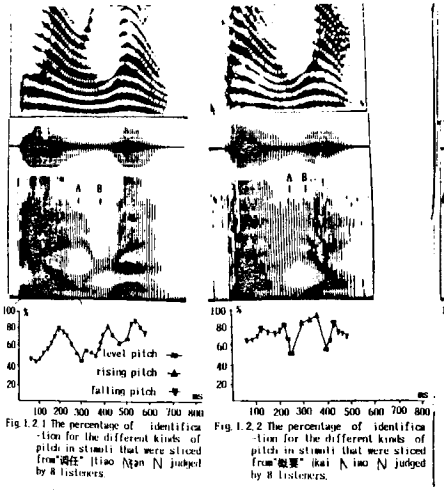


Fig. 1.2.1 The percentage of identification for the different kinds of pitch in stimuli that were sliced from "调往" [tiao Man N] judged by 8 listeners.

Fig. 1.2.2 The percentage of identification for the different kinds of pitch in stimuli that were sliced from "做事" [kai N tao N] judged by 8 listeners.

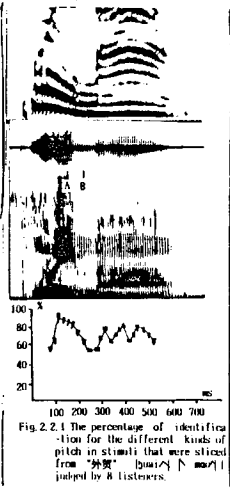


Fig. 2.2.1 The percentage of identification for the different kinds of pitch in stimuli that were sliced from "外贸" [tsy N mau] judged by 8 listeners.

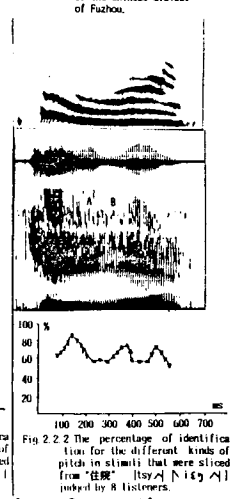


Fig. 2.2.2 The percentage of identification for the different kinds of pitch in stimuli that were sliced from "住院" [tsy N ɦɛŋ N] judged by 8 listeners.

duration of the second syllable, respectively, but the rising F0 parts in VCV and VV with falling-convex tones of the Chinese dialect of Fuzhou occur durning not only the non-voiceless initial, but also the syllabic vowel and its adjacent transition, and their durations have 167ms and 140ms amounting 52% and 49% of the whole duration of the syllable, respectively. In VCV, the dura-tion of the rising F0 part minus that of the voiced consonant initial is about 100ms, the remaining rising F0 part occuring durning the syllabic vowel and its adjacent transition; In VV, the most part of the rising F0 occur during the syllabic vowel and its adjacent transition. Therefore, the rising F0 part in convex tone in VCV and VV is perceived.

REFERENCE

[1] Jingzhu Yan & Maocan Lin (1988), "The acoustic manifestation of stress in three syllable groups of Beijing Mandarin", FangYan, No. 3(in Chinese).

[2] Rose, P. J. (1988), "On the non-equivalence of F0 and pitch in tonal discription", In prosodic Analysis and Asian Linguistics: to honour R. K. Spring, eds. by D. Bradley, Eugenie, J. A. Henderson and Martine Mazandon.

[3] House, P. J. (1990), Tonal perception in speech, Lund Univ. Press, Sweden.

TONAL MOVEMENTS IN THAI

A. Tumtavitikul
Rangsit University and Kasetsart University, Thailand

ABSTRACT

This paper investigates the acoustics of the five Thai tones with respect to the constraints for contour tone perception proposed by House [5]. A comparison with the production model which compares a tonal contour to the response of a step-input of a second order linear system [2], [7], [8] is given. The phonological representation of Thai contour tones is suggested. Finally, 'the optimal range' of tonal movements for contour tone perception is discussed.

INTRODUCTION

House [5] finds three perception constraints for contour tones as movement contour features-- a minimal vowel duration of 100 ms., contour movement onset in synchrony with vowel onset, and contour movement occurring during spectral stability. When these criteria are not met, the tone is perceived as tonal level. These perception constraints do not apply to tonal excursions which do not fall within a certain 'optimal range' which is yet to be defined. House notes that the tonal contours in his studies range between 3-8 semitones per 100 ms. Such an optimal rate of tonal movement serves to distinguish the perception of contour tone features, e.g. Rise, Fall, from level tone features, e.g., High, Low. He further invites studies on both production and perception of various tonal languages for verification.

This paper investigates the five tones in Thai traditionally described as three levels, High, Mid, Low, and two contours, Fall and Rise. The perception and representation of the Thai contour tones is controversial [1], [3], [4], [10], [11] while High, Mid, and Low are well agreed upon to be single level tones [9]. Phonetically, modern Mid and Low tones fall slightly whereas High tone rises. The acoustic inspection in this study focuses on the direction and rate of F0 change, and duration and onset of F0 change. All were examined in synchrony with an observation of the spectral pattern.

MEASUREMENTS

Subjects are ten native Thai speakers, 5 males and 5 females ages 19-33 years. The recordings were made on isolated [aa] syllables in all five tones, 5 tokens per tone, 25 tokens per subject. The F0 and time measurements were made on a PC computer using Kay Elemetrics CSL 4300 programs.

F0 end-points for each contour movement were measured at the beginning and end of the rise or fall. Such measurements define the total interval of the tone. The velocity of F0 change was calculated from the mid 75% portion of the entire contour for each slope [6]. This is the 'response' slope where the maximal rate of pitch change occurs [6]. The contour movement onset was measured beginning at the voicing onset of the vowel to the time where the F0 begins to change direction, either rising or falling. The same measurements were applied to all tones. A few tokens with discontinuous F0 pattern were disregarded. Both the total and the response intervals were normalized to semitones and the velocity of the response slope ($\Delta F0/\Delta time$) to semitones per second.

RESULTS

For both males (tbl.1) and females (tbl.2), only Low tone has the beginning of the falling F0 within 50 ms. after the vowel onset, i.e., having the beginning of the tonal contour synchronized with the vowel onset [5]. For the vowel duration, all tones have the tonal slope span over a period of ± 200-300 ms. during which there is spectral stability, with an average of ± 200 ms. for Fall, Rise, and ± 300 ms. for High, Mid, Low (tbls. 1 & 2). Combining the contour movement onset and interval time (tbls. 1 & 2), the approximate vowel duration is 350-400 ms. for all tones.

For the tonal interval (fig. 1), High, Mid and Low have an average interval of ≤ 3 semitones whereas Fall and Rise span an interval of 5-7 semitones.

The tones were grouped in two groupings according to the direction of the tonal excursions; Falling pattern with Fall, Low and Mid, and Rising pattern with Rise and High. A correlation was calculated between the response interval and response velocity of pitch change, and between the response interval and response duration for each tone separately, and for each tonal group. The velocity and the interval are found to be highly correlated ($p < .005$) for each tone, and for each tonal group (tbl. 4). The duration and interval, however, do not correlate for most tones for both males and females, with an exception of Fall (females) and Mid (males) (tbl. 3). For the tonal groups, the duration inversely correlates with the interval size (tbl. 3).

DISCUSSIONS

The only tone that meets all the three criteria for the contour feature perception [5] is Low. However, Low is categorized as a level tone in Thai. Two possible determining factors seem to be the rate of F0 change and the interval size.

The correlation between the velocity of F0 change and the interval size for all tones ($p < .005$) for both males and females, with no correlation found for the interval and the duration for most tones except for Fall in females and Mid in males (tbl. 3) seem to indicate a time constant in the pitch control mechanism, [2], [7], [8], [10], [11]. Such data advocates for a directional, interval dependent default rate of F0 transition which is automatically generated, with sequences of level tones as the representations (comparable to the step-input of a linear system). Moreover, the correlation between the velocity and the interval ($p < .005$) for both tonal groups (tbl. 4) seems to indicate that the same mechanism for tonal movements is being applied to all tones with the same direction of tonal change; Fall, Low and Mid, and Rise and High.

Since velocity is not distinctive for Low tone, the factor for level tone categorization seems to be narrowed down to the interval size. Interestingly, High, Mid and Low have an average interval of 3 semitones (fig. 1).

For Mid, the contour movement onset criteria is not met. Neither is High. In all, it seems that the level tone category is attributed to its interval threshold of 3 semitones when the vowel duration is 350-400 ms. regardless of the contour movement onset.

For both Fall and Rise, the contour movement onset criteria is not met. However, based on the production model derived from the correlations discussed above [10], [11], the tones are suggested to be represented as sequences of levels, High-Low for Fall, and Low-High for Rise.

Finally, the correlations found for the tonal groups (tbls. 3 & 4), especially with regard to duration, seem to indicate time adjustments made between Thai 'level' and 'contour' tone productions.

Table 1. Average Contour Movement Onset, Total Interval Size and Total Interval Time for the five Thai tones from combined male speakers.

Tones	Contour Movement Onset (ms.)	Total Interval (s)	Interval Time (ms.)
Mid (n=21)	82 (s.d.=78.54)	2.53 (s.d.=1.12)	269 (s.d.=88.26)
Fall (n=23)	99 (s.d.=41.38)	7.38 (s.d.=2.01)	237 (s.d.=45.23)
High (n=24)	70 (s.d=75.83)	2.99 (s.d.=1.21)	263 (s.d.=70.87)
Rise (n=16)	142 (s.d.=51.04)	6.05 (s.d.=1.17)	203 (s.d.=48.18)
Low (n=17)	48 (s.d.=24.19)	1.51 (s.d.=0.81)	316 (s.d.=50.18)

Table 2. Average Contour Movement Onset, Total Interval Size and Total Interval Time for the five Thai tones from combined female speakers.

Tones	Contour Movement Onset (ms.)	Total Interval (s)	Interval Time (ms.)
Mid (n=21)	114 (s.d.=135.46)	2.46 (s.d.=0.74)	310 (s.d.=103.51)
Fall (n=25)	208 (s.d.=66.21)	6.88 (s.d.=1.43)	200 (s.d.=41.81)
High (n=25)	102 (s.d.=97.06)	2.99 (s.d.=0.87)	286 (s.d.=111.23)
Rise (n=25)	229 (s.d.=54.47)	5.22 (s.d.=1.32)	190 (s.d.=36.16)
Low (n=21)	25 (s.d.=20.84)	2.71 (s.d.=0.81)	309 (s.d.=83.15)

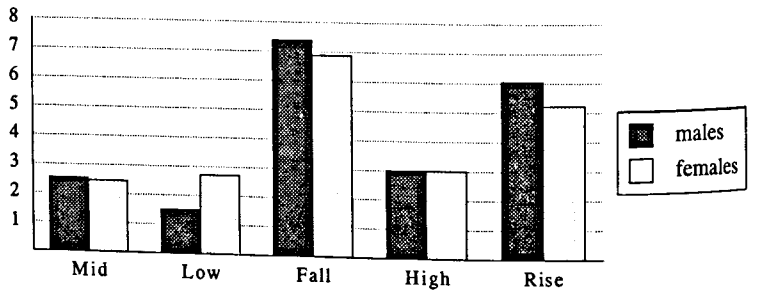


Figure 1. Average Total Interval for the five Thai tones (in Semitones).

Table 3. Correlation between Response Interval vs. Duration for Thai tones (p = .05)

	M	L	F	H	R	F, L &M	H &R
females	r = .23 n.s.	r = .27 n.s.	r = .52 p < .005	r = -.20 n.s.	r = -.04 n.s.	r = -.42 p < .005	r = -.44 p < .005
males	r = -.59 p < .005	r = -.31 n.s.	r = .33 n.s.	r = -.04 n.s.	r = -.11 n.s.	r = -.35 p < .01	r = -.39 p < .01

Table 4. Regression & Correlation between Response Interval and Response Velocity of Thai tones in two groupings; Fall, Low & Mid, and Rise & High.

Tones	Y = a + bX	Correlation (r)	Significance (p)
Falling Pattern: Fall, Low & Mid			
males	Y = -.06 +5.72X	0.9531	< .005
females	Y = -2.37 +6.89X	0.9426	< .005
Rising Pattern: Rise & High			
males	Y = -2.58 +7.36X	0.8849	< .005
females	Y = -5.72 +8.56X	0.8976	< .005

SUMMARY

Implications from this study are: first, the rate of F0 transition is the same, interval dependent default rate for all tones with the same direction of F0 movement. Second, the difference between Thai 'level' and 'contour' tones is in the interval size, ± 3 semitones is the threshold for level tones. Contour tones span an interval of 3-8 semitones. Also, there seems to be some time adjustments between 'level' and 'contour' tone productions. Third, Thai Rise and Fall do not meet House's constraints on contour feature perception. Rather, the correlation between the velocity and the interval, and the time constant for each tone favors the representation as a sequence of levels. Finally, the 'optimal range' of contour tone perception in House's studies (3-8 s/ ≥100 ms.) includes a rate which is faster than the 'optimal' or 'default' production rate defined within this study. Whether the contour perception 'optimal range' contains the production rate awaits further verification.

ACKNOWLEDGMENT

The author gratefully acknowledges Donna Neleson and Siriphan Sriwanyong for their kind contributions to this work.

REFERENCES

[1] Abramson, A. S., and Svastikul, K. (1982), "Intersection of Tone and Intonation in Thai", in H. Fujisaki, and E. Gårding (eds.), *Working Group on Intonation: Preprints for the XIIIth ICL.*

[2] Fujisaki, H. (1983), "Dynamic Characteristics of Voice Fundamental Frequency in Speech and Singing", in P. MacNeilage (ed.), *Speech Productions,* New York: Springer-Verlag.

[3] Gandour, J. T. (1983), "Tone Perception in Far Eastern Languages", *J. Phonetics 11.149-175.*

[4] Gandour, J. T. (1975), "The Representation of Tones in Siamese", in J.G. Harris, et al (eds.), *Studies in Tai Linguistics in honor of William J. Gedney,* Bangkok: Central Institute of English.

[5] House, D. (1990), *Tonal Perception in Speech,* Lund: Lund University.

[6] Ohala, J. J., and Ewan, W.G. (1973) "Speed of Pitch Change", *JASA 53.345.*

[7] Öhman, S.E.G. (1968) "A Model of Word and Sentence Intonation", *Speech Transmission Laboratory Quarterly Progress and Status Report 2-3.1-11.*

[8] Sundberg, J. (1979), "Maximum Speed of Pitch Changes in Singers and Untrained Singers",*J. Phonetics 7.71-79.*

[9]Tingsabadh, M.R. K., and Abramson, A.S. (1993), "Thai", *JIPA 23.1.24-28.*

[10] Tumtavitikul, A. (1994), "Thai Contour Tones", in H. Kitamura, et al (eds.), *Current Issues in Sino-Tibetan Linguistics,* Osaka: The Organization Committee of the 26th ICSTLL.

[11] Tumtavitikul, A. (1989), *Rate of Fundamental Frequency Change in Tones,* MA Report, UT Austin.

THE INFLUENCE OF SILENCE ON PERCEIVING THE PRECEDING TONAL CONTOUR

David House
Dept. of Linguistics and Phonetics, Lund University,
Helgonabacken 12, S-223 62, Lund, Sweden.

ABSTRACT

Interactive adjustment tests were carried out to test if a silent interval influences the perception of the preceding tonal contour. Results from 16 subjects show a strong influence of silence on tonal perception indicating that silence increases sensitivity for the preceding tonal endpoint with subjects showing greatest response consistency for the stimuli with the longest pause where adjustment is based on endpoint frequency before the pause.

INTRODUCTION

In both read and spontaneous speech, a prosodic phrase boundary is often accompanied by a silent pause which is preceded by a tonal contour marking the boundary. Considerable attention has been directed to the respective roles of silent pauses and boundary tones as markers of prosodic phrase and syntactic boundaries, see e.g. [1], [2], [3], [4], [5], [6], [7], [8] and [9]. The central question approached by this investigation is whether a silent interval influences the perception of the preceding tonal contour.

A number of general questions concerning boundary tones relate to the central issue of this investigation. Boundary tones may have several functions in addition to boundary signalling, e.g. signalling feedback seeking or turn regulation in spontaneous dialogue [10]. Are the tones and functions perceived categorically and, if so, does a silent interval facilitate perception?

In a previous study [11] it was shown that in synthesized VCVCV sequences where V= [a] and C= [m], the tonal configuration in vowels is perceptually more salient than in consonants. It can be conjectured, however, that if a silent interval is inserted before a vowel, tonal perception in the preceding consonant may be sharpened. This could give the final tonal level in the consonant greater perceptual significance than when immediately followed by a vowel. Thus, the following specific questions are addressed by this investigation: 1. Does a silent interval influence tonal perception? 2. Are final sonorant consonants important tone carriers? 3. Is perception of the tonal endpoint before a pause sharpened by the pause, and if so, does this sharpening increase with increased pause duration?

METHOD

Stimuli and task

To answer these questions, a set of adjustment tests was designed. Stimuli consisted of synthesized [amamam] sequences in three temporal conditions: 1) no pause between segments, 2) a 100 msec pause between the fourth and fifth segment [amam.am] and 3) a 1000 msec pause between the fourth and fifth segment [amam.....am]. Formant synthesis was used to generate the stimuli [12].

The subjects' task was to match different tonal configurations within each temporal condition. Matching was done interactively using a mouse pointer on a computer screen (Sun workstation, ESPS-Waves+ environment). The tonal configurations were 1) a falling F0 contour where the fall occurred through both the second vowel and second consonant and 2) a falling F0 contour through the second vowel only with a constant F0 on the second consonant. 10 Hz steps between 140 and 60 Hz were used to create 9 different stimuli in each tonal configuration making a total of 18 different stimuli for each temporal condition, i.e. a total of 54 stimuli for all three temporal conditions and both tonal configurations. See Figure 1 for stylized samples of stimuli.

Where endpoint frequency is most salient, subjects would be expected to match tonal configurations having the same endpoint frequency regardless of whether the contour falls in the vowel only or in both the vowel and consonant.

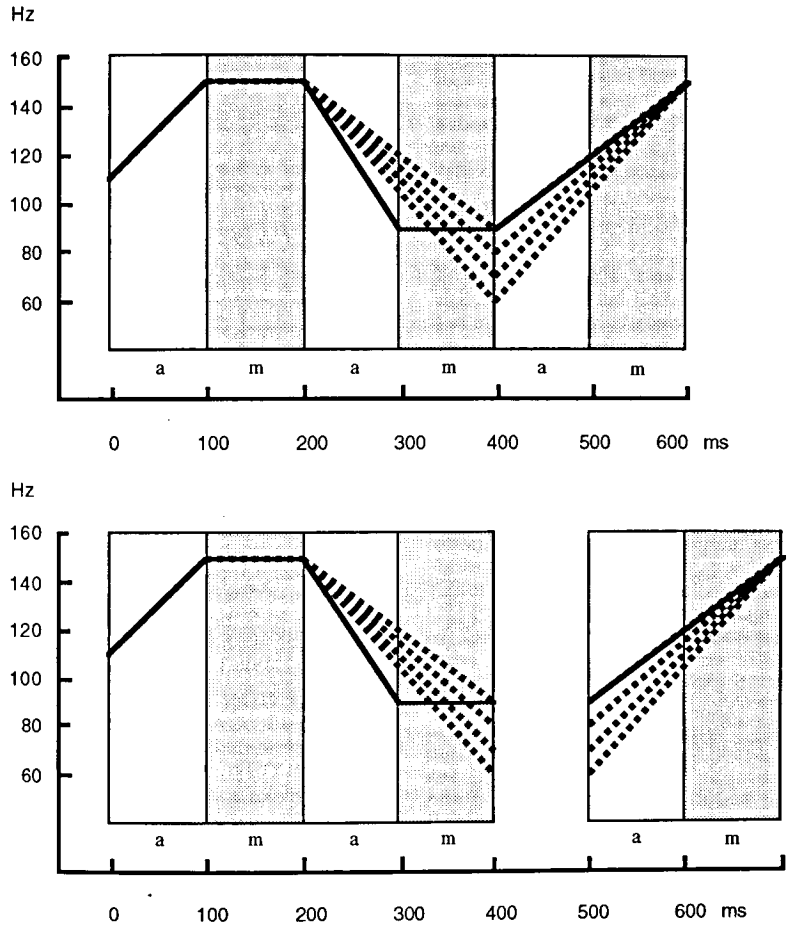


Figure 1. Stylized contours of some example stimuli. The upper panel represents the temporal condition "no pause" while the lower panel represents "short pause". The dotted lines represent tonal contours falling in both vowel and consonant (VC-fall) while the solid lines represent tonal contours falling in the vowel only (V-fall).

If endpoint frequency is of less perceptual importance, subjects would be expected to match contour shapes. This would result in subjects matching a lower endpoint for a vowel-consonant fall with a higher endpoint for a vowel only fall.

Test configuration

The middle five stimuli in the continuum of nine stimuli in each tonal configuration were presented as original stimuli. This resulted in six blocks of five stimuli each for a total of 30 presented stimuli. Stimuli were randomized in each block and block order was randomized between listeners. Subjects were asked to match each original stimulus to one of the nine stimuli having the same temporal conditions but the different tonal configuration. The test was presented as an adjustment procedure as in [13] with the nine choices presented in frequency order on the computer screen. All screen input was logged to a file.

Each subject began the test with a practice/calibration block in which the original stimulus was identical to one of the nine choices. The entire test took an average of 33 minutes with a minimum individual time of 13 minutes and a maximum of 55 minutes.

Subjects

16 subjects participated in the experiment. Subjects were mostly students and staff at the Dept. of Linguistics and Phonetics, Lund University, and all but two were native speakers of Swedish. Subjects were not paid, but were rewarded with chocolate and coffee after their participation in the test.

RESULTS

Results were very consistent between subjects: one factor ANOVA df=15, F=0.69, p>0.05, and within subjects df=2, F=43.17, p<0.0001. Figure 2 shows the percentage of same endpoint responses for the three temporal conditions and for the two tonal configurations within each condition.

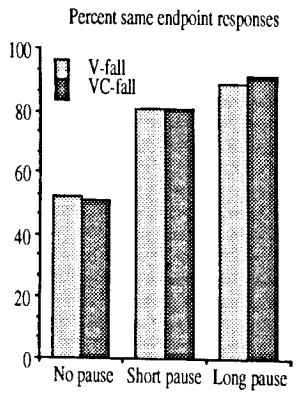


Figure 2. Graph showing percentage same endpoint responses as original stimulus when original is V-fall (falling contour in vowel only) and VC-fall (falling contour in vowel and consonant).

A strong effect of pause on endpoint perceptual salience can be seen. In the no-pause condition, only half the responses were same endpoint, while the other half were in the direction of a lower endpoint for vowel-consonant fall (VC-fall) being matched with a higher endpoint for the vowel fall (V-fall). Table 1 shows the response distribution where the direction is from the vowel fall. Endpoint salience also seems to increase somewhat with pause duration.

Table 1. Endpoint response distribution for the three temporal conditions. Direction is frequency of endpoint related to endpoint of tonal contour falling in vowel only.

	Lower	Same	Higher
No pause	76	83	1
Short pause	28	130	2
Long pause	12	144	4

A chi square test of independence on the above distribution results in $\chi 2=76.54$, df=4, p<0.001. One way ANOVA shows a significant difference comparing no pause with short pause F(2,45)=21.13, p<0.001 and comparing no pause with a long pause F(2,45)=40.5, p<0.001, but not when comparing a short pause with a long pause F(2,45)=3.13, p>0.05.

DISCUSSION

The results demonstrate a strong effect of silence on the perception of the tonal contour. They also demonstrate the importance of a sonorant consonant as a tone carrier particularly when followed by silence.

In the pause stimuli, matching seems to be based primarily on endpoint frequency before the pause, while in non-pause stimuli, listeners seem to be attending more to fall gradients or to average frequency through the fall. An interpretation concerning auditory memory may serve to help explain the results. If short-term auditory memory for frequency is sharpened by the presence of a pause, then subjects should find it more salient to match endpoint frequency even if the final segment is not a vowel. In the no-pause condition, auditory memory relies more on the tonal contour since endpoint frequency is rendered less salient by the following vowel.

This interpretation may be modified by the fact that, due to test construction constraints, there was also some information after the pause which listeners could have used as well as the endpoint information before the pause. The fact remains, however, that the presence of the pause significantly influenced perception of the tonal contour.

This can have implications for perception of such tonal phenomena as boundary tones and discourse markers. The presence of a pause may therefore sharpen perception of a boundary tone or discourse marker. More precision may be called for in intonation modelling and automatic stylization of intonation, especially concerning tonal contours before pauses. This would be in line with data in [14] where automatic stylization for recognition tended to fail most often in prepausal positions.

ACKNOWLEDGMENT

Many thanks are due to Marcus Filipsson for writing the interactive program used for the perception test.

REFERENCES

[1] Beckman, M., and Pierrehumbert, J. (1986), "Intonation structure in Japanese and English", in J. Ohala (ed.), *Phonology Yearbook 3,* pp. 255-309.

[2] Gårding, E., and House, D. (1987), "Production and Perception of Phrases in some Nordic Dialects", In P. Lilius and M. Saari (eds.). *The Nordic Languages and Modern Linguistics 6,* pp. 163-175, Helsinki University Press.

[3] Wightman, C.W., Shattuck-Hufnagel, S., Ostendorf, M. and Price, P.J. (1992), "Segmental durations in the vicinity of prosodic phrase boundaries", *Journal of the Acoustical Society of America,* vol. 91, pp. 1707-1717.

[4] Bruce, G., Granström, B., Gustafson, K. and House, D. (1993), "Phrasing strategies in prosodic parsing and speech synthesis", *Proceedings Eurospeech '93,* pp. 1205-1208, Berlin, Germany.

[5] Bruce, G., Granström, B., Gustafson, K. and House, D. (1993), "Interaction of F0 and duration in the perception of prosodic phrasing in Swedish", In B. Granström and L. Nord (eds.). *Nordic Prosody VI,* pp. 7-21. Stockholm: Almqvist & Wiksell International.

[6] Pijper, J.R. de, and Sanderman, A. (1993), "Prosodic cues to the perception of constituent boundaries", *Proceedings Eurospeech '93,* pp. 1211-1214, Berlin, Germany.

[7] Strangert, E. (1993), "Speaking style and pausing", *Reports from the Department of Phonetics, University of Umeå, PHONUM 2,* pp. 121-137.

[8] Strangert, E. and Strangert, B. (1993), "Prosody in the perception of syntactic boundaries", *Proceedings Eurospeech '93,* pp. 1209-1210, Berlin, Germany.

[9] Swerts, M. and Geluykens, R. (1994), "Prosody as a marker of information flow in spoken discourse", *Language and Speech,* vol. 37, pp. 21-43.

[10] Bruce, G., Granström, B., Gustafson, K., House, D. and Touati, P. (1994), "Modelling Swedish prosody in a dialogue framework", *Proceedings of the 1994 International Conference on Spoken Language Processing,* pp. 1099-1102, Yokohama.

[11] House, D. (1990), *Tonal Perception in Speech,* Lund: Lund University Press.

[12] Carlson, R., Granström, B. and Hunnicutt, S. (1991), "Multilingual text-to-speech development and applications", in W. Ainsworth (ed.), *Advances in speech, hearing and language processing,* pp. 269-296, London: JAI Press.

[13] d'Alessandro, C. and Castellengo, M. (1993), "The pitch of short-duration vibrato tones", *Journal of the Acoustical Society of America,* vol. 93, pp. 1617-1630.

[14] House, D. and Bruce, G. (1990), "Word and focal accents in Swedish from a recognition perspective", In K. Wiik and I. Raimo (eds.), *Nordic Prosody V.* pp. 156-173. Turku University.

TWO KINDS OF STRESS PERCEPTION

Thomas Portele, Barbara Heuft
Institut für Kommunikationsforschung und Phonetik, Universität Bonn

ABSTRACT

This paper describes some preliminary results concerning the perception of syllable stress as either a binary feature or a nearly continuous parametric value. Two experiments were set up: a perception test where the subjects were forced to assign stress to one of two syllables, and a labelling experiment. Here, the subjects had to rate the degree of stress carried by a syllable using values between 0 and 32.

MOTIVATION

Word accent may serve as a distinctive cue in discriminating two words that are identical on the segmental level (in English, for instance, <pro'cess> and <'process> ; in German <'umlaufen> (to run over something) vs. <um'laufen> (to run around something)). An accented syllable may also serve as a first guess in segmenting the speech signal into words [1]. To perform these functions a syllable must be perceived as either stressed or unstressed. This implies that some kind of categorical perception takes place.

On the other hand, listeners are able to distinguish between syllables regarding the amount of stress they carry. The perception of focus accents is an indication for this ability. Fant and Kruckenberg [2] used a 30 point scale for subjective judgements of syllable stress and obtained reliable results.

To obtain an impression about how strong these abilities are developed in German listeners two experiments were set up. The first experiment was designed to assess the listeners' abilities in assigning stress to one of two syllables when one parameter, i.e. the position of the F_0 peak, is gradually changed. A similar experiment was carried out by Kohler [3] but he explored not lexical but semantic categories. In the second experiment three listeners judged more than 8500 syllables regarding their amount of stress. They used a scale from 0 to 31. The correlations between their ratings were evaluated as well as possible factors guiding their judgements.

EXPERIMENT 1

Method

The pairs <voll Milch> (full of milk) - <Vollmilch> (whole milk) and <zwei Räder> (two wheels) - <Zweiräder> (bicycles) were embedded in short texts and read by a male (<Vollmilch - voll Milch>) and a female speaker (<Zweiräder - zwei Räder>). A previous experiment [4] showed that the position of an F_0 peak serves as a cue to syllable stress for these stimuli, and that a movement of the peak to the other syllable often leads to a different stress assignment by listeners.

For each sentence containing one of the syllable pairs the F_0 contour was parametrized using a method by Portele et al. [5]. This method describes an accent by four numerical values, the position of the F_0 maximum being one of them. This parameter was modified and the position of the F_0 maximum was shifted towards the other syllable in 6 steps with a stepwidth of 50 ms (Figure 1). In the vicinity of the position where the stress shift was supposed to take place 10 steps with a stepwidth of 10 ms were used. These 15 intonation contours were imposed on the original sentence using the PSOLA method. Altogether, 60 stimuli were used. They were presented to the subjects (n=13) via headphones in a quiet room. The subjects' task was to decide which orthographic representation was correct for a given stimulus, e.g. whether <Vollmilch> or <voll Milch> was spoken. The linguistic content of the stimulus sentences was neutral to both interpretations. Each stimulus was presented

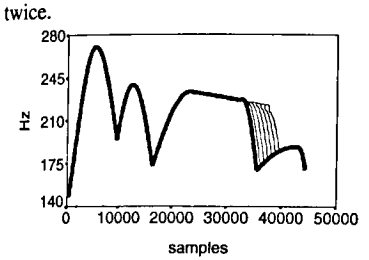


Figure 1. Parametrized intonation contour for the utterance "In diesem Keller wurden nachweislich Zweiräder montiert." (Provable, bicycles were assembled in this cellar). The thick line indicates the original contour, the thin lines show the shift of the F_0 peak in steps of 50 ms to the right.

Results

The results are displayed in Figure 2. Each picture displays in the upper half the individual scores of those subjects where the difference between the ratings to the left and the right from the vertical line (point of stress shift) is significant (t-test, $p < 0.075$). In the lower half (between 0 and 1) the pooled results from all subjects are shown. Rating 2 in the upper half and 1 in the lower half stands for perceiving the first syllable as stressed, rating 1 in the upper half and 0 in the lower half for preceiving the second syllable as stressed. The vertical line indicates the position of the F_0 peak where the change in perception occurs for most subjects (there are some individual differences). The subjects performed differently; some subjects obtained highly significant results for all four stimulus groups (<Vollmilch>: 10 subjects, <voll Milch>: 6 subjects, <Zweiräder>: 7 subjects, <zwei Räder>: 10 subjects) while others behaved completely erratic.

Figure 3 shows the results from one subject. Here, the position of the F_0 peak where the stress shift takes place, is clearly recognizable.

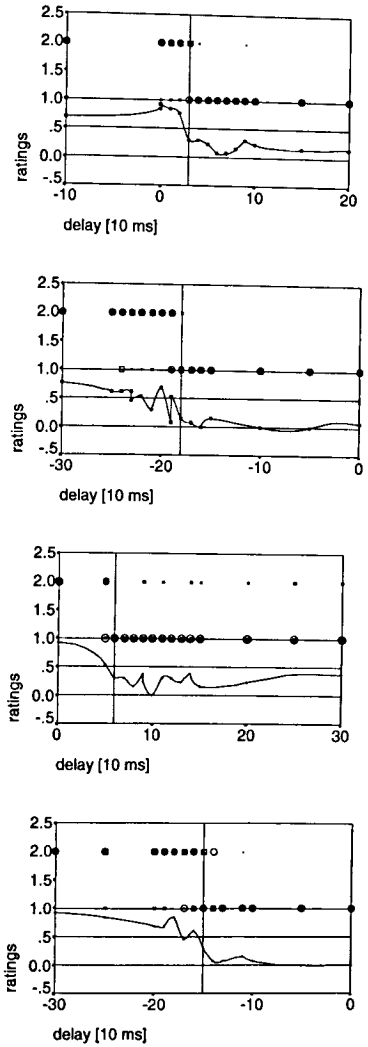


Figure 2. Results of the stress assignment experiment. The original stimulus was <Vollmilch>,<voll Milch>,<Zweiräder>, <zwei Räder> (from above). Further explanation is given in the text.

Discussion

The results show that the identification of the stressed syllable in suitable syllable pairs depends on the position of the F_0 peak. When this peak is moved towards the originally unstressed syllable there is a

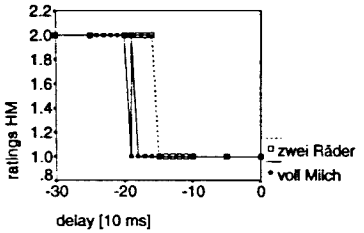


Figure 3. Judgements by one subject for the stimuli <zwei Räder> and <voll Milch>. Rating 2 stands for "first syllable stressed", rating 1 for "second syllable stressed".

certain point where the perception of the stressed syllable switches from one syllable to the other. However, not all subjects were consistent in their judgements. As shown in [4], for the material used in this experiment syllable duration is a stress cue as strong as F_0. The apparent mismatch between these two cues might be responsible for the erratic behaviour shown by some subjects. For other subjects (Figure 3) the position of stress switch can be located within 10 ms. This interval is a little longer than one pitch period.

These results are ony preliminary. A lot remains to be done, i.e. identifying the causes for the different behaviour of the subjects, looking for a connection between the position of the switch point and the properties of the speech signal, checking the discriminative ability of the subjects as the second requirement for categorical perception etc. However, the results show that, in German, the exact placement of the F_0 peak is crucial for correct perception, and, therefore, for intonation modelling in speech synthesis [3]. Our description of F_0 contours [5] was designed specifically to meet these requirements.

EXPERIMENT 2

Method

The corpus used in this experiment is part of a prosodic database [6]. More than 300 sentences were read by one male and two female speakers. Altogether, 8646 syllables were used. Their amount of stress was rated by three listeners on a scale between 0 and 31 [2]. The speech signal and the transcription were presented. They used a graphical scale to indicate the prominence level. They were allowed to listen to an utterance as often as they liked.

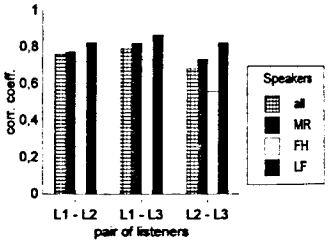


Figure 4. Correlation coefficients between the listeners' judgements.

Results

The correlation coefficients between the listeners' judgements are diplayed in Figure 4. The agreement between the subjects is slightly lower for speaker FH. An average correlation coefficient of 0.75 indicates the high similarity between the ratings. The agreement between the listeners was higher for prosodically marked utterances (orders or yes/no questions) than for simple statements.

The relation between acoustic properties of a syllable and its perceived prominence was also investigated. Figure 5 displays the connection between syllable duration and prominence rating.

There is a marked dependency between the existence of an F_0 peak asscociated with a syllable and the syllable's perceived prominence (U-test, $p < 0.001$). The average difference between syllables with and without F_0 peak is 12 prominence grades.

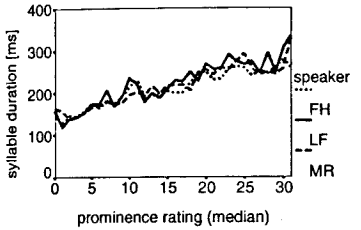


Figure 5. Relation between prominence rating and syllable duration displayed for each speaker.

The relation between the height of an F_0 peak and the perceived prominence is not very strong (but significant; Kruskal-Wallis-test, $p < 0.001$).

It was found that the offset between the start of a stressed vowel and the asscociated F_0 peak is significant for offset values greater than zero. Larger offsets (late peaks) induced higher prominence values [3].

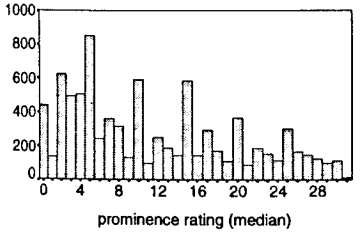


Figure 6. Histogram of the prominence ratings by all listeners.

Discussion

The results show that the listeners are able to differentiate between more than two or four levels of syllable prominence. How many prominence levels are sufficient, however, can not be deduced from the data, because the distribution is quite even (Figure 6). However, one can assume that judgements were made relative to other syllables in the utterance. It is unlikely that an isolated presentation of the syllables would have led to similar results. But this procedure would be far away from a real communication situation.

The acoustic properties of the speech signal have some measurable influence on the perceived prominence. However, linguistic factors that could be deduced from the transcription by the listeners seem to play a more important role [7].

CONCLUSION

The two experiments show that stress can be perceived as a binary feature and as a multi-level parameter of a syllable. Both kinds of perception are necessary for efficient communication.

ACKNOWLEDGEMENT

This research was partly funded by the *Deutsche Forschungsgemeinschaft* and by the *Verbmobil* project sponsored by the *Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie.* We thank Bernd Möbius for the texts used in the first experiment, Monika and Florian for painfully judging 8646 syllables, and all other subjects and speakers.

REFERENCES

[1] Cutler, A.; Norris, D. (1988), "Te role of strong syllables in segmentation for lexical access." J. Exp. Psych. Human Perception and Performance **14**, 113-121

[2] Fant, G.; Kruckenberg, A. (1989), "Preliminaries to the study of Swedish prose reading and reading style." STL-QPSR **2/1989**, 42-45

[3] Kohler, K.J. (1987), "Categorical pitch perception." Proc. XIth ICPhS, Tallin, 91.2

[4] Heuft, B.; Portele, T. (1994), "Zur akustischen Realisierung des Wortakzents." Proc. Elektron. Sprachsignalverarbeitung V, Berlin, 197-204

[5] Portele, T.; Krämer, J.; Heuft, B. (1995), "Automatische Parametrisierung von Grundfrequenzkonturen." (to appear in Proc. DAGA-95, Saarbrücken).

[6] Heuft, B., et al. (1995), "Parametric description of F_0 contours in a prosodic database." Proc. ICPhS'95, Stockholm

[7] Heuft, B., et al. (1995), "Betonungsstufen von Silben und ihre Beziehung zum Sprachsignal." (to appear in Proc. DAGA-95, Saarbrücken).

WHAT DEFINES VOWEL IDENTITY IN PRELINGUAL INFANTS?

Ocke-Schwen Bohn and Linda Polka***
**English Department, Kiel University, Germany*
*** School of Human Communication Disorders, McGill University, Montreal, Canada*

ABSTRACT

This is the first report on experiments examining which acoustic properties of coarticulated vowels (target spectral, dynamic spectral, temporal) define vowel identity in prelingual infants. German-learning infants were tested for discrimination of German vowel contrasts in the Silent Center paradigm. Results indicate that infants derive vowel identity from dynamic spectral information, and that target information is not needed for perceived vowel identity.

INTRODUCTION

Previous studies of vowel perception by adult American English (AE) listeners and by adult German listeners have shown that three types of acoustic information contained in consonant-vowel-consonant (CVC) syllables contribute to vowel identity: target spectral information, dynamic spectral information of the syllable onsets and offsets, and temporal information. Strange and her collaborators established the relative importance of these three types of information [1], [2], [3]. The most important finding was that adult AE listeners and adult German listeners identify coarticulated native vowels highly accurately if target spectral information has been electronically removed from the CVC stimuli. This indicates that vowel identity does not depend on acoustic information on spectral targets. Instead, the perceptually relevant information for vowel identity seems to reside in the changing spectral structure of coarticulated syllables.

Several studies by Strange and her collaborators examined in detail the sources of dynamic information in the perception of native coarticulated vowels by adult AE listeners ([1], [2]) and adult German listeners ([3], [4]). These studies found that Vowel Centers (VCs), which consist only of the syllabic nuclei with target information, are not perceived more accurately than Silent Center (SC) syllables, which consist only of the dynamic portions of the syllable onsets and offsets in their appropriate temporal relationship. Vowel identity is maintained very well in SCs even though the vocalic nucleus with information on formant targets is silenced in SCs. These studies also showed that syllable onsets alone (INIs) or syllable offsets alone (FINs) do not provide sufficient information on vowel identity for adult listeners.

These findings and others on the insufficiency of target information for vowel perception are accounted for by Strange's Dynamic Specification Theory (DST), which states that vowels are specified by dynamic information defined over syllable onsets and offsets [1]. The dynamic information reflects each vowels' characteristic opening and closing phases in their appropriate temporal relationship and style of movement of the vocal tract. DST provides an elegant solution to an important problem in vowel perception research, viz., perceptual constancy. Unlike DST, Target Models of vowel perception have to account for the target undershoot problem (formant targets are often not reached in coarticulated vowels) and for speaker normalization (formant targets for the same vowel category differ greatly across men, women, and children). Context- and speaker-dependent variation pose problems only for those theories of vowel perception that view formant (or gestural) targets as objects of perception. Research motivated by the DST, however, strongly suggests that coarticulated vowels are specified by styles of movement that are invariant across consonant contexts [2] and across different speakers [5].

The experiments reported here are the first to examine the role of the three types of acoustic information of CVC syllables in infant vowel perception. This was done by testing German-learning infants' discrimination of naturally produced German /dVt/-syllables which were modified to manipulate the availability of the three types of acoustic information. No study has ever examined how target spectral information, dynamic spectral information, and temporal information contribute to perceived vowel identity in prelingual infants. This is somewhat surprising given the fact that many studies of infant vowel perception implicitly assume that spectral targets alone specify vowel identity (e.g. [6]).

METHOD

Stimuli

Six tokens each of the German vowels /i/, /I/, /e/, /ɛ/, /U/, /o/, /ɑ/ were produced in /dVt/-syllables by a male native German speaker and recorded onto DAT. The vowels were selected to be presented in the contrasts /i/-/e/, /e/-/I/, /I/-/ɛ/, and /o/-/U/ because these contrasts were confusable in the identification experiments reported by Strange & Bohn [3]. The maximal /i/-/ɑ/ contrast was selected as a control contrast. Measurements of syllable duration, voice onset time and fundamental frequency of multiple tokens of the seven vowels were used to make the final selection of four instances each of the six vowels.

To test the role of target vs. dynamic spectral information, the original syllables were modified as follows. SCs were generated by attenuating to silence the center portion of each of the original syllables, leaving onset and offset portions in their original temporal position. The onset and offset portions included the major part of the transitions. VCs were generated by silencing the onset and offset portions. INIs were generated by silencing both center and offset portions, and FINs were generated by silencing both onset and center portions.

To test the role of temporal information, all eight tokens for a given contrast were electronically edited so that they had the same duration. This was done by iterating or deleting full pitch periods (for full syllables with neutral duration - FNDs, and for VCs with neutral durations - CNDs), or by adding or deleting silence (for SCs with neutral durations - SCNDs).

Subjects

80 infants served as subjects. All were healthy, full term infants with no history of ear infections (by parental report). The infants aged between 7 and 11 months were being raised in monolingual German-speaking families in Kiel, Germany.

Procedure

Infants were tested using the headturn procedure (for details of our implementation, s. [7]). In this procedure a syllable is played from a loudspeaker every 1.5 sec and at random intervals this background syllable changes to a target syllable for a brief interval. Discrimination is assessed by first conditioning the infant to turn his/her head in the direction of a visual reinforcer above the loudspeaker when they detect a change in the background syllable. Correct headturns are reinforced by the activation of a visual reinforcer (an electronic animal that moves) accompanied by verbal praise. We implemented this procedure as a category change paradigm in which the background and the target consist of multiple tokens of each syllable type.

Discrimination of a vowel contrast in a given condition (e.g., /i/-/e/ as SC) was tested in a single session. The infant was seated on a parent's lap across a small table from an experimenter (E1). The loudspeaker and an array of visual reinforcers, located behind a smoked plexiglass panel, were arranged to one side of the parent and infant. The parent and E1 listened to music over headphones to prevent them from hearing the stimuli and influencing the infant. A second experimenter (E2), located ouside the test room, observed the infant through a one-way window and operated the computer.

The session begins with a conditioning stage in which the infant is given an opportunity to learn the contingency between the vowel change and availablity of the visual reinforcer (s. [7]). During the testing stage, E2 initiates trials when the infant is in a "state of readiness" (not fussing, facing E1 etc.). E2 is blind to the trial type and pushes a button when she observes a headturn during the trial interval. The visual reinforcer is activated automatically for a change trial when E2 records a headturn by pushing a response button. Twenty-five trials were presented during testing stage.

Infant testing was conducted in an sound-treated chamber. Custom software controlled stimulus delivery, activation of the reinforcers, and trial selection (i.e. presentation of change vs. no-change trial), and also recorded the number of trials, hits, misses, correct rejections and false alarms.

Design

Groups of 10 subjects each were assigned to one of the eight listening conditions, which were defined by the 4 vowel contrasts (/i/-/e/, /e/-/I/, /I/-/ɛ/, /o/-/U/) and by availability of temporal information (in unmodified syllables, SCs, and VCs) vs. its neutralization (in FNDs, SCNDs, and CNDs). Each subject was first tested for discrimination of the unmodified test contrast. Only infants who discriminated full syllables (criterion: 7/8 consecutive correct trials and > 60 % correct responses) were then tested on separate days for discrimination of the contrast tested initially in the edited conditions. Infants were randomly assigned to the two series of experiments. In the first series, infants were tested for discrimination of SCs, VCs, and INIs or FINs. In the second series, infants were tested for discrimination of FNDs, SCNDs, and CNDs. The vowel category which served as the background was counterbalanced within each group.

RESULTS

Figure 1 gives the overall results for the eight stimulus conditions for two vowel contrasts (/i/-/e/, /e/-/I/), expressed as percentage of correct responses averaged across subjects. Overall discrimination levels for unmodified syllables (mean % correct: 69.3), SCs (mean % correct: 67.0), and VCs (mean % correct: 63.7) did not differ significantly. Compared to the Full condition, vowel identity was well maintained in the SC condition, even though the vocalic nucleus with information on formant targets was not presented in that condition. The mean per cent correct values for INIs (47.5) and FINs (51.9) suggest that vowel onsets or vowel offsets alone do not preserve vowel identity.

The two vowel contrasts did not differ significantly in discriminability in the Full condition (/i/-/e/: 69.7 % correct, /e/-/I/: 68.9 % correct) and in the SC condition (/i/-/e/: 66.8 % correct, /e/-/I/: 67.1 % correct). In the VC condition, the /i/-/e/ contrast was significantly less discriminable (52.9 % correct) than the /e/-/I/ contrast (74.3% correct).

Neutralization of the temporal contrast reduced the discriminability of both contrasts in the Full conditions (Full syllables: 69.3 % correct; FNDs: 55.2 % correct), in the SC conditions (SC: 67.0 % correct; SCND: 56.3 % correct), and in the VC conditions for the /e/-/I/ contrast (VC: 74.3 % correct; CND: 53.0 % correct), but it did not affect the discriminability of the /i/-/e/ contrast in the VC conditions (VC: 52.9 correct; CND: 53.0 % correct).

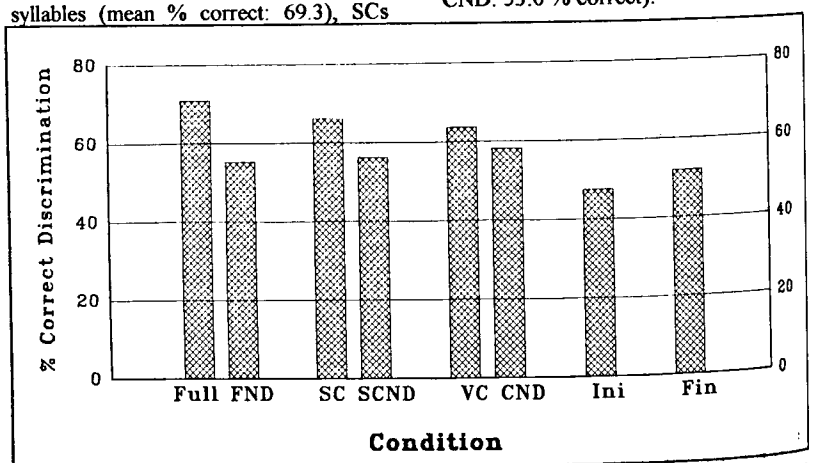


Figure 1: Overall per cent correct responses in vowel discrimination in unmodified syllables (Full), full syllables with neutral duration (FND) silent-center syllables (SC), silent-center syllables with neutral duration (SCND), vowel centers (VC), vowel centers with neutral duration (CND), initials (INI), and finals (FIN) conditions for the vowel contrasts /i/-/e/ and /e/-/II.

CONCLUSIONS

The most important finding was that German-learning infants discriminated two German vowel contrasts equally well when these contrasts were presented either as unmodified full syllables or as SCs, which preserve only the dynamic spectral information of the syllable onsets and offsets in their appropriate temporal relationships. This suggests that infants do not need target spectral information to differentiate the spectrally similar high front vowel contrasts /i/-/e/ and /e/-/I/. Rather, trajectory information specified over syllable onsets and offsets is a good source of information for vowel identity in prelingual infants, as it is in adult native speakers of AE and of German.

The overall pattern of results for the German infants is quite similar to that for German adults, who discriminated the same contrasts in a related study [4]. Within each age group of infants and adults, discriminabilty of Full, SC, and VC syllables did not differ significantly, but discrimination levels for INIs and FINs were lower than for SCs in both the adult and the infant study.

The first results from our experiments on the acoustic specification of vowels in infants support Strange's Dynamic Specification Theory, which states that vowels are specified by dynamic information defined over syllable onsets and offsets together. German infants discriminated two German vowel contrasts by making use of the dynamic sources of information associated with the opening and closing gestures at the margins of the CVC syllables. This indicates that infants perceive coarticulated vowels in terms of their characteristic styles of movement. We suggest that perceptual representations of these styles, which seem to be invariant across consonant contexts [2] and across different speakers [5], contribute importantly to perceptual constancy for vowel categories in infants. Further research is underway to establish the generality of our first results by examining how accurately infants discriminate vowels produced in varying contexts and by multiple speakers when presented only with dynamic information specified over syllable onsets and offsets.

One interesting aspect of our study is that infants' discrimination abilities suffer considerably if contrastive temporal information is not available. Unlike adult German listeners, for whom the neutralization of duration contrasts had only a very selective effect for individual vowel contrasts in specific experimental conditions (unpublished data from the study reported in [4]), German infants discriminated both contrasts at lower levels of performance when the contrasts were temporally neutralized. Further research will have to show whether German infants' sensitivity to temporal manipulations reflects L1 experience with the German vowel system, or whether duration differences have a universally important function in learning to differentiate vowel contrasts.

ACKNOWLEDGMENTS

Research supported by a grant from the *Deutsche Forschungsgemeinschaft* (DFG grant Bo-1055/3-1) to O. Bohn. We thank Desiderio Saludes and Sonja Trent for assistance in generating stimulus materials, and Kirsten Schriever, Tatjana Soldat, Anja Steinlen, and Amira Yassine for assistance in testing infants. Special thanks go to Winifred Strange for her essential support, advice, and encouragement.

REFERENCES

[1] Strange, W. (1989), "Evolving theories of vowel perception." *J. Acoust. Soc. America* 89, 2081-2087.

[2] Jenkins J. J. (1987), "A selective history of issues in vowel perception." *J. Memory and Language* 26, 542-549.

[3] Strange, W. & Bohn, O.-S. (1995), "Dynamic specification of coarticulated German vowels: I. Perceptual studies." To appear in *J. Acoust. Soc. America.*

[4] Bohn, O.-S. & Strange, W. (1995) "Discrimination of coarticulated German vowels in the Silent-Center Paradigm". *Proc. 13th Int. Congr. Phon. Sciences.*

[5] Jenkins, J. J. et al. (1994), "Vowel identification in mixed-speaker silent-center syllables" *J. Acoust. Soc. America* 95, 1030-1043.

[6] Kuhl, P. K. et al. (1992), "Linguistic experience alters phonetic perception in infants by 6 months of age." *Science* 255, 606-608.

[7] Polka, L. & Werker, J. F. (1994), "Developmental changes in the perception of nonnative vowel contrasts". *J. Experim. Psychology: Human Perception & Performance* 20, 421-435.

PRODUCTION-PERCEPTION RELATIONSHIP IN THE VOICING CONTRAST FOR MEDIAL STOPS IN CHILDREN AND ADULTS.

Cecile T. L. Kuijpers
Max-Planck-Institut für Psycholinguistik, Nijmegen, The Netherlands

ABSTRACT

A phoneme identification experiment was carried out to investigate perception of the word-medial voicing contrast by Dutch four-year-old children, six-year-old children, twelve-year-old children, and adults. The data of this experiment are related to an earlier production study in which we studied word-medial closure durations. Production and perception of the voicing contrast are considered to develop synchronously in young children.

INTRODUCTION

The perception and production of the voicing contrast by children and adults has been investigated in numerous phonetic studies. Several acoustic cues play an important role depending on the position of the phoneme in the word. This paper will be concerned with one of the durational features related to the distinction, namely closure duration. In English, the Voice Onset Time (VOT) is the major acoustic cue to distinguish initial voiced and voiceless stops. In final position vowel duration differences cue the distinction. In Dutch, the intial contrast is characterized by a negative VOT (of approximately -70 ms) for voiced and a positive VOT (of 10-20 ms) for voiceless stops [1]. In final position, the contrast is neutralized because of a devoicing rule.

In many languages the word-medial voicing contrast is characterized by a difference in closure duration; a short closure duration for voiced and a long closure duration for voiceless stops. In some classical studies on perception of the medial voicing contrast the closure duration was manipulated (that is, a silent interval). Here, the same tendency showed up: short intervals were identified as a voiced stop, and long intervals as a voiceless stop [1], [2].

Experimental study of production and perception of the medial voicing contrast in young children is scarce. In [3],[4] we discuss a number of English production studies, and we report on our own production data of Dutch 4-, 6-, 12-year-old children and adults. In our production study the young children displayed a large variability and, consequently, voiced and voiceless stops were less clearly differentiated in the younger age groups than in the older age groups [4]. Those results will be integrated in the discussion section of this paper where we will describe shortly the relationship between the perception and production data of the Dutch children.

Nearly all developmental studies investigated perception of the *initial* and/or *final* voicing contrast. Perception of *intervocalic* stops by 6-year-olds and adults was examined by [5]. They used the naturally produced words 'petal' vs. 'pedal' and examined the interaction of VOT, closure duration, and preceding vowel duration. With respect to the role of closure duration *per se* the data indicated a difference between children and adults: in the children's data the phoneme boundary (50% crossover) was situated at approximately 110 ms and in the adult data at 130 ms.

The identification experiment presented in this paper deals with the perception of intervocalic stops by Dutch children and adults. The silent interval of naturally produced words was manipulated. On the basis of our production data, and on the basis of the literature on initial and final stops, we expect that the perceptual differentation of voiced and voiceless stops develops only gradually with age.

METHOD

Participants

Four age groups participated in the experiment: 4-year-olds (mean age 4;6), 6-year-olds (mean age 6;4), twelve-year-olds (12;3), and adults (age range 22-55). In each age group 15 listeners were tested, male and female, and all were monolingual speakers of Dutch. There was no known hearing deficiency in any of the listeners.

Material

Two minimal pairs of bisyllabic nonsense words were used, viz. the names 'Táppi'-'Tábbi' and 'Pátto'-'Páddo'. The sound structure of these names did not evoke any association with existing words or names. The initial vowel was always /A/, and the final vowels were /i/ and /o/ since Dutch names are often characterized by these endings. The four words were used to construct four different continua, henceforth the P-, B-, T-, and D-set of words.

Task

An identification task was set up as a game with two pairs of large puppets. The resemblence of the names was reflected in the resemblence of the puppets. The stimuli were put into carrier phrases in which the stimulus name was repeated at the end, such as 'Give the ball to Tappi..to Tappi'. By actually giving a ball to a puppet the child manifested its response in a two-alternatives-forced-choice labelling paradigm.

Manipulation

The four tokens were digitized on a Digital microVAX II computer using a 20 kHz sampling frequency. The initial vowel was set at a neutral value so that listeners would not be biased by its duration. The formant transitions were always unimpaired. The stop closure was replaced by different silent intervals with endpoints at 10 ms and 130 ms. We used 20 ms increments and created 7 different versions of each test word. The prominence of the voiceless burst was modified in order to yield consistent voiced and voiceless responses at the end-points; duration and intensity were modified on the basis of independent perceptual judgments.

Procedure

First, an auditory picture discrimination pre-test (ADIT) was carried out to ensure that the young children could discriminate the selected minimal pairs. Next, all children were carefully trained to learn the test words and to relate the names to the puppets. The final training set comprised 8 randomized trials ('Now show me..Patto) in which the children had to give at least 7 correct responses out of 8. In the main experiment each stimulus occurred twice and the children were tested in two sessions within two weeks; in each session 28 stimuli were presented. The stimuli were randomized and counter-balanced across subjects. The sentenced were separated by an inter-stimulus interval of 6.5 s. The 12-year-olds and adults were tested individually and they listened to the same sentences as the young children.

RESULTS

Four identification functions were calculated for each age group. The data were processed by a statistical analysis which considers the identification function as a cumulative probability distribution on Z-scores. Next, a linear regression analysis was carried out to fit the original data. Phoneme boundaries between groups were compared by means of a t-test. The slope of the functions (phoneme boundary width) were compared by means of a t-test on the coefficients of regression. We refer to [4] for a detailed description of the analyses.

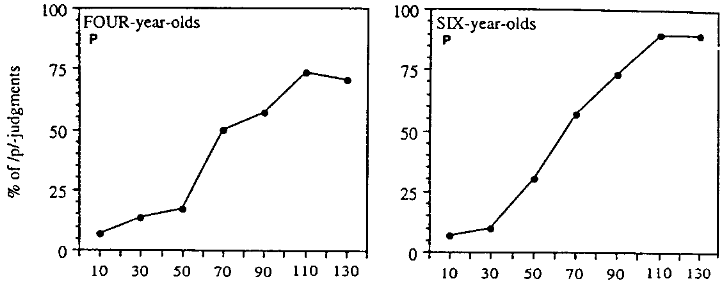


Figure 1. Identification functions of the /b/-/p/ continuum from the P-set of words for the 4-year-olds and the adults.

Phoneme boundary

The group identification functions were calculated for the four separate continua. Phoneme boundary corresponds to the 50% crossover, and to Z=0 in the transformed data. The percentage of voiceless jugments are given as a function of the silent interval. In Figure 1 we illustrate the identification functions of the 4-year-olds and the adults (P-set of words). In the P- and B-set of words no significant differences were found between any age group. In the T-set of words a significant difference was found between the 4-year-olds and 12-year-olds ($t=3.09$; $p<0.008$). No significant differences were found in the D-set of words. The data do show that all listeners need relatively long silent interval durations in the B/D-set of words in order to perceive a voiceless stop (see Figure 2). This results from the differences in formant transition information. Moreover, the young children's judgments mostly range between 10%-80% voiceless responses probably due to their inaccurate awareness of the durational characteristics of medial voiced and voiceless stops.

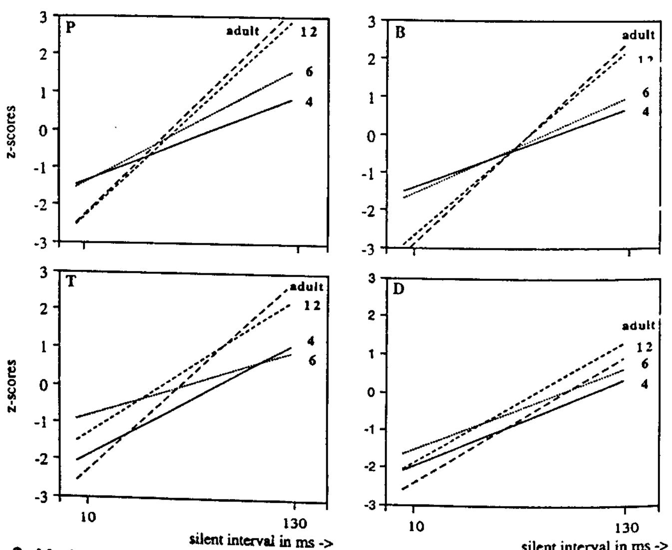


Figure 2. Modelled identification functions of Z-scores on stimuli with different silent intervals for the four age groups.

Phoneme boundary width

The phoneme boundary width corresponds to the 25%-75% interval on the fitted regression line. The comparison of the slopes of the regression lines in the P- and B-set of words indicated significant differences at $p=0.008$ between the groups 4-12, 4-adults and between the groups 6-12, 6-adults. In the T-set of words significant differences were found between the groups 4-6, 4-adults, 6-12, 6-adults. In the D-set of words, although not significant, there is a remarkable parallellism between the functions of the age groups 4 and 6 on the one hand, and 6 and 12 on the other hand.

We also examined the individual variability within each group. In short, the older age groups mainly displayed a continuous response mode, that is an upward ordering of voiceless judgments to stimuli with an ever increasing silent interval. The younger age groups displayed numerous discontinuous response modes, that is a less consistent perceptual behaviour. The upward ordering of voiceless judgments was frequently interrupted by voiced judgments.

DISCUSSION

We have indicated that the perception of the intervocalic stop voicing contrast by 4- and 6-year-olds deviates from the perception of that contrasts by 12-year-olds and adults. In comparison with the two latter groups, the young children need a relatively large difference in silent interval to perceive the distinction, and they have difficulty in determining which allophones come to be grouped in the same phonemic category. They do not categorize as consistently as the older listeners.

These findings concur with those reported by [6] who likewise concluded that the categorical perception of 6-year-olds is still unlike the adult norm. Surely, we presented stmuli in which the medial stop contrast was induced by only one acoustic parameter (silent interval) neglecting other parameters such as voicing. It is possible that younger children rely more than adults on multiple acoustic cues.

The findings reported in this paper are in agreement with our production data on closure duration differences in the medial voicing contrast. As in perception, the distinction becomes more and more attuned with age [4]. Both developments (production and perception) can be expressed in terms of a 'distincivity' which gradually increases with age. We assume that a parallel can be drawn between children's perception and production of the phonemic contrast.

ACKNOWLEDGEMENTS

This research was part of my Ph.D study financially supported by the University of Amsterdam and carried out at the Institute of Phonetic Sciences Amsterdam. I would like to thank L. Pols and F. Koopmans-van Beinum for their many helpful suggestions and discussions on the dissertation.

REFERENCES

[1] Slis, I. & Cohen, A. (1969a,b). "On the complex regulating the voiced-voiceless distinction I, II". *Language & Speech*, vol. 12, pp. 80-102, pp. 137-155.

[2] Lisker. L. (1978)."Rapid vs. Rabid. A catalogue of acoustic features that may cue the distinction". *Haskins Laboratory Status Report on Speech Reserach*, vol 54, pp. 127-132.

[3] Kuijpers, C. (1993a). "Temporal aspects of the voiced-voiceless distinction in speech development of young Dutch children". *Journal of Phonetics, vol.* 21, pp. 313-327.

[4] Kuijpers, C. (1993b). *Temporal co-ordination in speech development*. Un-published Doctoral dissertation, University of Amsterdam, The Netherlands.

[5] Allen, G. & Norwood, J. (1988). "Cues for the intervocalic /t/ and /d/ in children and adults. *Journal of the Acoustical Society of America*, vol. 84, pp. 868-875.

[6] Burnham, D., Earnshaw, L. & Clark, J. (1991). "Development of categorical identification of native and non-native stops: infants, children and adults". *Journal of Child Language*, vol. 18, 231-260.

THE DEVELOPMENT OF EARLY VOCALIZATIONS OF DEAF AND NORMALLY HEARING INFANTS IN THE FIRST EIGHT MONTHS OF LIFE

Chris J. Clement, Els A. den Os, and Florien J. Koopmans - van Beinum
Institute of Phonetic Sciences, Institute for Language and Language Use, University of Amsterdam, Institute for the Deaf St.Michielsgestel, The Netherlands

ABSTRACT

To establish how and from which age onwards, speech perception influences the development of vocalizations in the first year of life, we studied the speech production of deaf and normally hearing infants longitudinally from 2.5 months until 7.5 months of age. Several differences between deaf and normally hearing infants were observed indicating that lack of auditory feedback influences speech production already at this early stage of speech development.

INTRODUCTION

Some recent studies suggest a deviant speech production of hearing impaired compared to normally hearing infants in the first year of age [e.g.1]. No canonical babbling was found in deaf infants before the age of eleven months while most hearing infants start babbling before that age [2]. In several studies differences were observed in consonantal features and phonetic repertoire size [e.g.3].

Until now - to our knowledge - no systematic study has been performed on the vocalizations of deaf infants starting within the first half year of life. The present study reports on longitudinal data of 6 deaf and 6 normally hearing infants between 2.5 and 7.5 months of age. The main question we address is: do hearing impaired infants differ from normally hearing infants with respect to number and type of vocalizations?

METHOD

Subjects

Twelve mother-infant pairs participated in this study; six infants profoundly hearing impaired (group HI) and six infants with normal hearing (group NH). All infants have normally hearing parents. By means of developmental tests performed when the infants were 12 and 18 months of age no clear cognitive or motor delays were found. The HI infants had an average hearing loss over 90 dB at the best ear, established by Auditory Brain-stem Response audiometry (ABR) in the first six months of life. The profound hearing loss was confirmed by several pure-tone audiometric tests at later ages. Hearing aids were used by three subjects within the studied period. Two hearing impaired infants were raised with TC (Total Communication)/Dutch Sign Language, two infants by TC and two mainly by the Oral method.

The NH infants were matched with the HI infants on the following criteria: sex, birth order, duration of pregnancy, mother age, socio-economical status of the parents, and dialect of the parents. All NH infants were recorded from the age of 2.5 months onwards, two HI infants from the age of 2.5 months, two from 3.5 months and three from the age of 5.5 months onwards.

Data collection

Audio recordings, lasting about half an hour each, were made every two weeks. The mothers of the infants themselves made the recordings at home. The mothers were asked to talk with their children in a face-to-face situation while the children were sitting in an upright position.

Procedure of analyses

Of every monthly audio recording, the first 10 minutes were transcribed. Two trained phoneticians performed and verified the transcriptions. The inter-judge agreement based on all material (62 recordings) was 93% for the infant utterances. An infant utterance was defined as a sound production during one breath cycle starting with inspiration. Laughing, crying and vegetative sounds were not taken into account. The number of infant utterances during the first 10 minutes were counted.

Fifty infant utterances per recording were selected evenly out of the transcribed ten minutes. The total of 3100 utterances were digitized into a computer with a sample frequency of 48 kHz and stored for further analysis.

Each utterance was classified in one of three possible types of articulation 1) no articulatory movement; 2) one articulatory movement (e.g. gooing); 3) two or more articulatory movements during two or more syllables, i.e. babbling. Furthermore, each utterance was classified according to one of five possible types of phonation: 1) uninterrupted phonation; 2) interrupted phonation 3) variegated phonation (variation in the intonation, pitch or loudness e.g. screaming and growling) 4) a combination of interrupted and variegated phonation 5) no phonation. Types 4 and 5 were rarely found and therefore left out of consideration in this paper. The classification was derived from an earlier study on infant speech development [4].

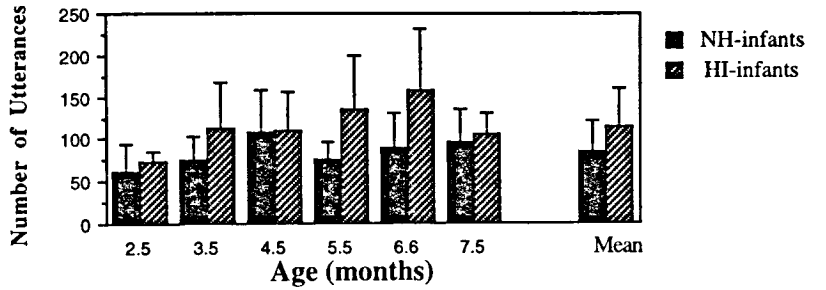


Figure 1. Mean number of utterances and standard deviations during ten minutes of interaction presented for the two groups of infants per month, as well as the mean number for the 6 months combined. (N is 6 in case of the NH infants at each age. N is 2, 3 and 3 at 2.5, 3.5 and 4.5 months resp. and 6 at 5.5, 6.5 and 7.5 months in case of the HI infants.)

RESULTS

Number of utterances

Figure 1 represents the mean number of utterances in 10 minutes and their standard deviations per age as well as the average number over 6 months for both groups.

It can be observed that the mean number of utterances for the combined 6 months is higher in case of the HI infants (115, sd=45) compared to the NH infants (85, sd=37). A t-test on the data of the combined 6 months indicates a significant difference between the groups ($t(60)=2.95$, $p \leq .005$, one-tailed). By separating the ages in two different age groups (from 2.5 to 4.5 months and from 5.5 to 7.5 months) we can get an indication of a developmental effect. By means of a Mann-Whitney U test no significant differences between the HI and NH infants are found at the early age. At the later age, a Mann-Whitney U test shows that HI infants produce significantly more utterances than their hearing peers ($U(18,18)=79$, $p \leq .005$).

Utterance duration

In figure 2 the mean utterance duration of the 50 selected utterances is presented

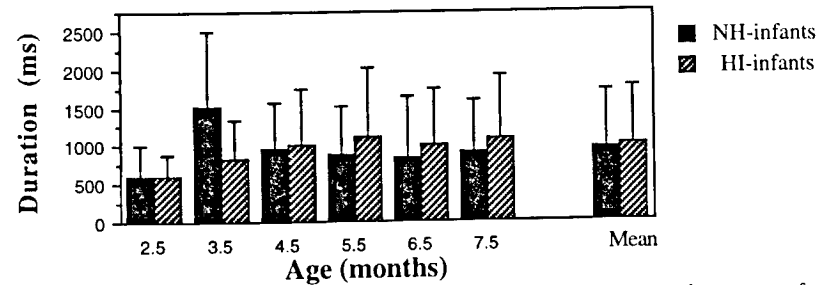


Figure 2. Mean utterance duration and standard deviations of the 50 selected utterances for the HI and the NH group per month, as well as the mean duration for the 6 months combined. (N is 300 at each age of the NH infants. N is 100, 150 and 150 at 2.5, 3.5 and 4.5 months resp. and 300 at 5.5, 6.5 and 7.5 months in case of the HI infants.)

per age, as well as the average duration over the 6 months for both groups. It can be observed that the mean utterance duration for the 6 months combined is somewhat longer for the HI infants (997 ms, sd=761) than for the NH infants (948 ms, sd=761). A z-test on the months combined shows a low significant difference ($z=1.77$, $p\leq.05$). A z-test performed on the utterance duration per month indicates that from 5.5 months to 7.5 months the HI infants produce longer utterances (5.5: $z=3.72$, $p\leq.0005$; 6.5: $z=2.65$, $p\leq.005$; 7.5: $z=2.63$, $p\leq.005$). At the age of 2.5 and 4.5 months no significant differences between the two groups are found. At the age of 3.5 months, however, the mean duration of the NH infants is longer than at any other age in the studied period. A z-test indicates that the NH infants produce significantly longer utterances at the age of 3.5 months than the HI infants ($z=9.45$, $p\leq.0005$).

Type of utterances

In figure 3 the "articulation types" are shown per age for both groups. A tendency can be observed that the HI infants produce fewer utterances with articulation movements than NH infants in the first months although this turned out to be not significant according to a Mann-Whitney U test, nor in the later age period. Utterances with 2 or more articulation movements are produced more often by NH infants than by HI infants in the later 3 months ($U(18,18)=100.5$, $p\leq.05$). The total amount of this utterance type in the HI group is due to only one subject who started to babble at 7.5 months of age.

The "phonation types" are presented in figure 4. It can be seen that NH infants produce more interruptions in the airstream specially in the later months ($U(18,18)=85.5$, $p\leq.01$ for the data at 5.5, 6.5 and 7.5 months combined). Although a tendency for more variegated phonation can be observed by the HI infants, no significant differences are found.

DISCUSSION

In the present study it could be observed that, as a group, the HI infants produced more utterances than their hearing peers in the period between 2.5 and 7.5 months. This was found as well in a previous study on HI and NH infants between 5.5 and 9.5 months of age [5] These studies support the suggestion of Locke [6] that deaf infants vocalize more than normally hearing infants, possibly due to extra effort HI infants expend to get auditory stimulation. It seems that the often reported reduction in number of utterances takes place after the period we studied, namely towards the end of the first year [e.g.7].

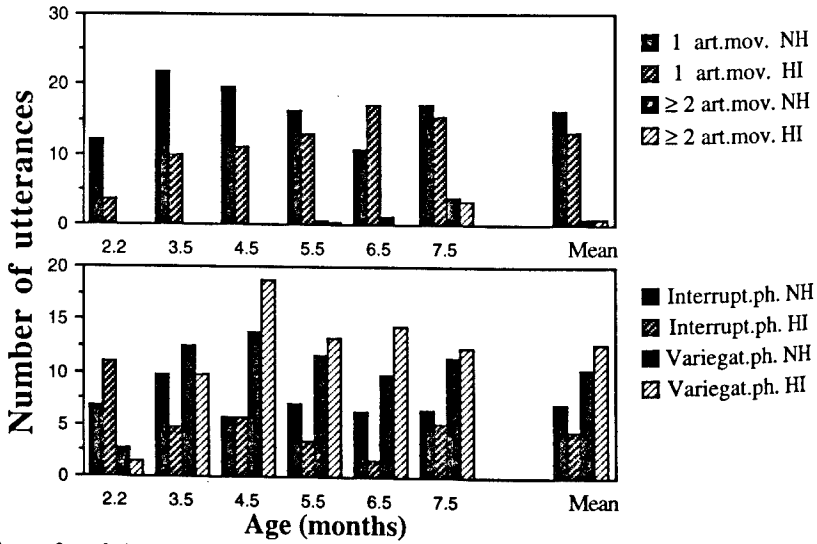


Figure 3 and 4. Mean number of utterances with 1 articulation movement or 2 (or mor articulation movements (fig.3) and interrupted or variegated phonation (fig. 4) for the NH an HI group per month , as well as the mean for the 6 months combined. (N: see fig. 2)

Furthermore, we found a longer utterance duration of the NH infants at age 3.5 months compared to the HI infants. After the 5th months the profile seems to be reversed; the HI produce longer utterances than the NH infants. The longer utterance duration might indicate - already in this early age - a tendency of HI children to prolong syllable duration as was found in a study on 6-to-10-year-old children [8].

In the phonation domain we found differences between the two groups in number of utterances with interrupted phonation, particularly in the later ages of the studied period. We did not find evidence for the finding of Stark [9] that the sound types which are characteristic of the "vocal play stage" (experimentation with squealing, growling, friction and other noises) are produced by HI infants to a limited extent only. A possible explanation for this difference in results might be that Stark studied the utterances of HI infants from 15 months onwards. Furthermore, the HI infants produced fewer babbling utterances within the studied age period than their NH peers.

In summary, it seems that already within the investigated period, i.e., between 2.5 and 7.5 months of age, several differences in the speech production between HI and NH infants can be observed. The differences become more clear from about 5.5 months onwards, with respect to number of utterances, utterance duration, interrupted phonation, and babbling. This may be due to lack of auditory feedback on the speech production from that age. In the first months fewer differences between the two groups can be observed. This may suggest a stronger influence of biologically determined factors (e.g. anatomical growth) on vocalizations in these first months compared to a later period.

CONCLUSION

Since the results of the present study are based on a small sample size, specially in the early months of age, caution should be taken when making any conclusion. In the period between 2.5 and 7.5 months, described in this paper, we observed a number of differences in the vocalizations between 6 deaf and 6 hearing infants. These differences can be found both in a quantitative and in a qualitative sense. Our preliminary results suggest that a lack of auditory feedback influences the speech production already in this very early stage of development.

REFERENCES

[1] Kent, R.D., Osberger, M.J., Netsell, R., & Goldschmidt Hustedde, C. (1987), "Phonetic development in identical twins differing in auditory function", *JSHD* 52, pp. 64-75.

[2] Oller, D.K. & Eilers, R.E. (1988), "The role of audition in infant babbling", *Child development* 59, pp. 441-449.

[3] Stoel-Gammon, C. (1988), "Prelinguistic vocalizations of hearing impaired and normally hearing subjects: a comparison of consonantal inventories", *JSHD* 53, pp. 302-315.

[4] Koopmans-van Beinum, F.J. & Van der Stelt, J.M. (1986), "Early stages in the development of speech movements", In: B. Lindblom and R. Zetterström (Eds), *Precursors of early speech.* Wenner-Gren Int. Symp. Series 44, New York: Stockton Press, pp. 37-50.

[5] Clement, C.J., Den Os, E.A. & Koopmans-van Beinum, F.J., (1994), "The development of vocalizations of hearing impaired infants". *Proceedings of the Institute of Phonetic Sciences Amsterdam*, 18, pp. 65-76.

[6] Locke, J.L. & Pearson, D.M. (1992), "Vocal learning and the emergence of phonological capacity", In: C.A. Ferguson, L. Menn & C. Stoel-Gammon (Eds), *Phonological development: models, research, implications*, Timonium: York Press, pp. 91-129.

[7] Maskarinec, A.S., Cairns, G.F., Butterfield, E.C. & Weamer, D.K. (1981), "Longitudinal observations of individual infant's vocalizations", *JSHD* 46, pp. 267-273.

[8] Ryalls, J. (1993), "An acoustic study of speech production in normal, moderate-to-severe and profoundly hearing-impaired French-speaking children", *Proceedings of the third Congress of the International Clinical Phonetics and Linguistics Association*, University of Helsinki 39, pp. 167-174.

[9] Stark, R.E. (1983), "Phonatory development in young normally hearing and hearing impaired children", In: I. Hochberg, H. Levitt & M.J. Osberger (Eds), *Speech of the hearing impaired: Research, training and personal preparation*, Baltimore: Universtiy Park Press, pp. 251-266.

ADULT JUDGEMENTS OF INFANT VOCALISATIONS

Francisco Lacerda[1] and Tamiko Ichijima[2]

[1]Institute of Linguistics, Stockholm University, S-106 91 Stockholm, Sweden
[2]Sofia University, Tokyo, Japan

ABSTRACT

In normal adult-infant interaction, adult listeners make spontaneous on-line judgements of the infants' vocalisations, providing the infant with immediate feedback. Thus, the adults' spontaneous judgements may influence the adult-infant communication because, if consistent, they implicitly assign "meaning" to the infants' vocalisations.

Although normal interaction is a two-way procedure — infant and the adult influence each other — the goal of this study was to assess how adults interpret infant vocalizations, per se. Adult listeners were requested to judge vocalizations that had been produced by infants living in the same or in a foreign ambient language. The results suggest that opening degree is the most consistent phonetic dimension conveyed by infant babbling.

INTRODUCTION

Adults do not expect pre-linguistic infants to produce accurate and intentional speech sounds. However, as the infant babbles, adults tend to provide interpretations for the vocalisations produced by the infant. This behaviour may generate a mutual feedback process that, in addition to contributing to the development of bounding between infant and adult, can provide relevant phonetic information to the infant. Since the infant's attention is directed to the adult's utterances, this adult-infant interaction offers potentially optimal conditions for the emergence of a communication code. To the extent that the adult is capable of providing a consistent labelling of the infant vocalisations, even a quasi-random vocalisation pattern will be suitable for the establishment of a vocalisation-based communicative code [1].

The issue of the adult's interpretation of infant vocalisations is not new. A number of phonetic studies have addressed this question, providing detailed phonetic descriptions of the early stages of infant speech development [2-4]. From our point of view, however, accuracy in phonetic transcription is typically associated with a loss in the adults' response spontaneity. Phonetic transcriptions of babbling are difficult to achieve and usually demand repeated listening of recorded utterances along with strict inter-transcriber reliability criteria. Thus, to investigate the nature of the spontaneous feedback that the infant receives from the adult, we assessed the adult speakers' notion of the infant's articulatory vocalisation gesture. In this paper we report the main aspects of that study.[1]

METHOD

To assess adult-infant interaction under controlled conditions, we extracted infant vocalisations from natural mother-infant interaction situations and presented them for spontaneous adult judgements. We assume that adult judgements produced in a speeded response paradigm are close to the immediate interpretation of infant vocalisations that the adult would provide normal interactive setting.

The study was designed to investigate possible cross-linguistic differences in the adults' perception of babbling. We do not expect infants to produce language-specific vocalisations during the early stages of babbling. At later stages the infant's vocalisations will eventually converge towards the sounds that are used in the ambient language. Thus, our prediction is that the adults' judgements of vocalisations produced during the first stages of babbling will tend to reflect the adults' native language. For vocalisations during later babbling stages we expect more consistent adult classifications.

[1] A paper with a complete procedural description and extensive discussion of the resutls is currently being prepared for publication.

Subjects

A group of 12 Swedish and 11 Japanese subjects participated in the experiments. The subjects were students of phonetics who had attended an elementary course in phonetics. Although these subjects are regarded as "non-expert", some degree of phonetic awareness was necessary to be able to estimate the tongue position on the arbitrary "frontness x heightness" space.

Stimuli

The infant vocalisations were extracted digitally from recordings of babbling made at the homes of two Japanese[2] and one Swedish infant. The recordings cover an age range from 17 to 78 weeks.

The stimuli for the perception test were randomly sampled among all the vocalisations that were free from noise or the mother's speech.

Procedure

Two sets of stimuli were created: a set of 107 vowel-like utterances that had been produced by the Japanese infant and a set of 150 utterances produced by the Swedish infant.

The stimuli were presented via AKG K135 headphones. To familiarise the subjects with the task, a training session was run before the test proper. The stimuli were presented in random order, in blocks of 5 stimuli and an ISI of 3.5 s.

The subjects estimated the infant's tongue position for each vocalisation. To reduce biasing with phonetic or orthographic symbols, the responses were marks of tongue position on an idealised 5x5 grid of heightness and frontness.

RESULTS

Japanese infants

Swedish listeners

The judgements of vowel height and frontness produced by Swedish adults when listening to Japanese infants are displayed as contours in figures 1a and 1b. The contours define regions of the hypothetical vowel space within which the subjects agreed in their judgements. The contours enclose areas of the vowel space where the agreement among the adult subjects increases in 5% steps.

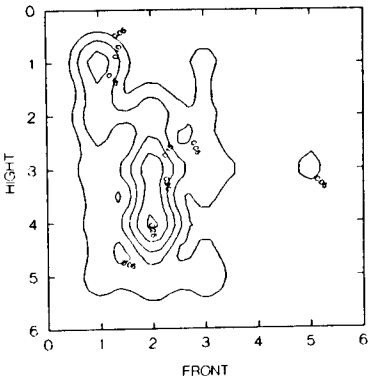


Figure 1a. Swedish listeners' judgements of tongue position. Japanese infant 4-8 months old.

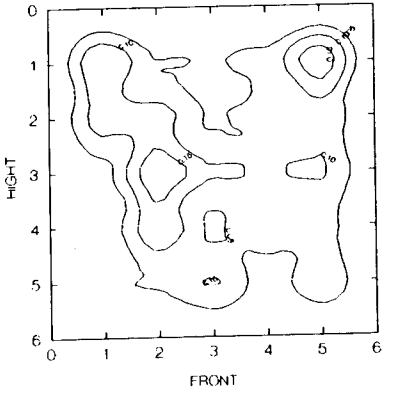


Figure 1 b. Same as fig.1a but for babbling produced at 19 months of age.

Japanese listeners

The judgements obtained from the Japanese listeners, when listening to the same babbling material are shown in figures 2a and 2b. The age groups are the same that were used in figures 1a and 1b.

[2] In the following we will not distinguish between the data from these two infants.

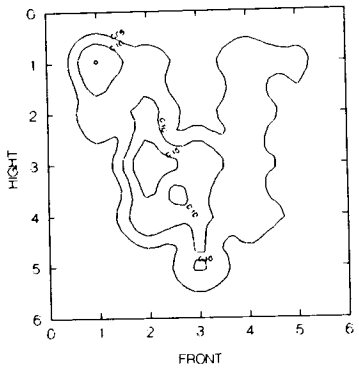


Figure 2a. Japanese listeners' judgements of tongue position. Japanese infant 4-8 months old.

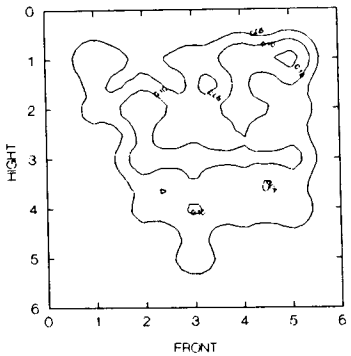


Figure 2b. Same as fig. 2a but for 19-months old Japanese infant.

Swedish infants

Swedish listeners

The distribution of non-expert Swedish listeners' judgements of the babbling of a Swedish infant are displayed in figures 3a and 3b.

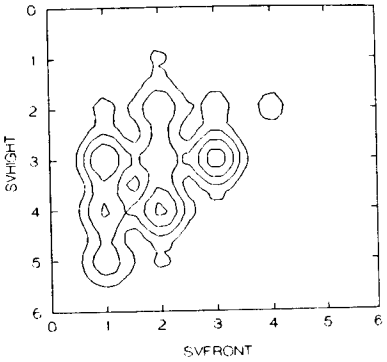


Figure 3a. Swedish listeners' judgements of tongue position. Swedish infant 4-8 months old.

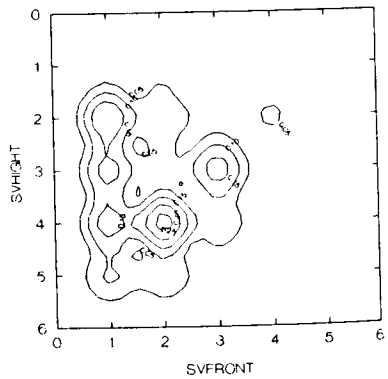


Figure 3b. Same as in figure 3a but for 9-18 months of age.

Japanese listeners

The distributions of the judgements of tongue height and frontness that were produced by the Japanese adults when listening to the babbling of the Swedish infant are shown in figures 4a and 4b. Also in this case the Japanese listeners judged the same stimuli that had been presented for the Swedish listeners.

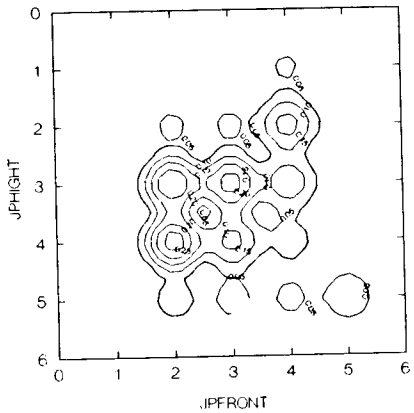


Figure 4a. Japanese listeners' judgements of tongue position. Swedish infant 4-8 months old.

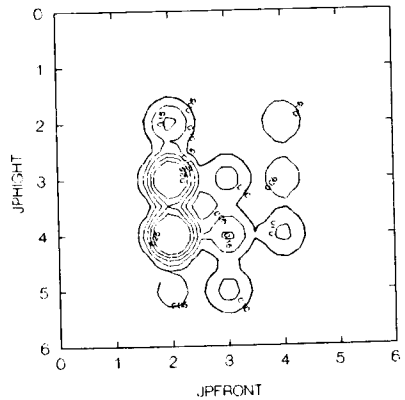


Figure 4b. Same as in figure 4a but for 9-18 months of age.

DISCUSSION

The results above are simply counts of the relative response frequency for each position on the frontness × height chart. This statistic does not reveal response stability for specific infant utterances. Instead it provides a general indication of how consistent adults are. Vocalisations equally rated by the adult listeners increase response frequency at that location in the grid. Following this line of reasoning, the data suggests that adults tend to be more consistent in tongue height judgements of early babbling and that consistency in frontness tends to appear only for later babbling, produced between 9 and 19 months.

The response pattern of the Japanese listeners tends to be more consistent than that of the Swedes in the use of frontness. This may be a reflection of the typological differences between the Japanese and the Swedish adult vowel systems.

Finally, there seems to be an indication of continuity in the judgements produced for babbling from the early and the later developmental stages. The islands of consistency observed up to 8 months tend to be still present in the later, in spite of its increased diversity.

CONCLUSIONS

This study suggests that adults may provide consistent feedback on degree of vowel opening for early babbling vocalisations. Consistence in degree of frontness appears only for later babbling.

ACKNOWLEDGEMENT

We would like to thank Lise-Lotte Roug-Hellichius and Ingrid Landberg for giving us access to the Swedish babbling data and to Christin Andersson for preparing the perception tests. We also would like to thank the students who participated in the perception tests.

Research supported by grant 90-0150 from The Bank of Sweden Tercentenary Foundation and by Fukutake Publisher.

REFERENCES

[1] Lacerda, F. (1995), "The perceptual-magnet effect: An emergent consequence of exemplar-based phonetic memory", *ICPhS 95*.

[2] Oller, D. K., Wieman, L., Doyle, W. and Ross, C. (1976), "Infant babbling and speech", *J. Child Lang. 3*, 1-11.

[3] Roug, L., Landberg, I. and Lundberg, L-J. (1989), "Phonetic development in early infancy: A study of four Swedish children during the first eighteen months of life", *J. Child Lang. 16*, 19-40.

[4] Vihman, M. (1992), "Early syllables and the construction of phonology", in C. Ferguson, L. Menn and C. Stoel-Gammon (Eds.), *Phonological Development: Models, Research, Implications*, Parkton, Maryland: York Press, 393-422.

VOCAL LEARNING IN INFANTS: DEVELOPMENT OF PERCEPTUAL-MOTOR LINKS FOR SPEECH

Patricia K. Kuhl and Andrew N. Meltzoff
University of Washington, Seattle, WA

ABSTRACT

Infants' development of speech begins with a language-universal pattern of production that eventually becomes language-specific. Our work focuses on the processes underlying this developmental change in infants' production of speech. One important contribution to this change is vocal learning — the process by which infants listen to the patterns of ambient language and attempt to reproduce them. We here discuss studies on 12- to 20-week-old infants' vocalizations in response to speech. The results show that (a) very young infants imitate sound patterns they hear, and (b) developmental change occurs in the microstructure of infants' vowel categories. We conclude that connections between auditory and articulatory representations of speech exist very early in life.

INTRODUCTION

Speech production during the first two years of life has been described by universal stages [1]. These changes encompass: *Reflexive Phonation* (0 to 2 months), *Cooing* (1 to 4 months), *Expansion* (3 to 8 months), *Canonical Babbling* (5 to 10 months), and *Meaningful Speech* (10 to 18 months). Although there is considerable consensus on the description of successive stages, there is less known about the *processes* by which change in infants' vocalizations are induced. Two factors are critical in producing the transition in infants, anatomical change and vocal learning. The young infant's vocal tract is very different from that of the adult, and anatomical growth must contribute, at least in part, to the stage-like shift seen in infants' vocalizations.

The factor emphasized here, however, is *vocal learning*. Human infants learn speech by listening to ambient language and attempting to produce sound patterns that match what they hear [2]. Cross-cultural studies show that infants from different linguistic environments exhibit differences in their vocalizations by 1 year of age, suggesting that by this age, ambient language has had an effect on spontaneous speech production [3].

The question examined here is whether there is evidence of vocal learning earlier in life. One way to investigate this point is to study vocal imitation. It offers a behavioral assay of vocal learning. Vocal imitation requires that infants recognize the relationship between articulatory movements and sound. Adults have an internalized auditory-articulatory "map" that specifies the relations between mouth movements and sound. When do infants acquire the auditory-articulatory map?

Experimental Evidence

Infants' vocalizations in response to speech were recorded at three ages, 12-, 16-, and 20-weeks of age [4]. Infants (N = 72) watched and listened to a female producing one of three vowels, /a/, /i/, or /u/. The vowels were produced once every 5 sec by a female talker and presented via a video display (life-size and in color). The stimuli were presented to infants as they reclined in an infant seat at a distance of 18" from the monitor.

All of the infants' vocalizations were recorded. Those meeting pre-established criteria for "vowel-like" utterances (greater than 100 msec and fully resonant, N = 224) were analyzed perceptually by having them phonetically transcribed, and analyzed instrumentally using computerized spectrographic techniques. The phonetic transcriber used a narrow transcription to initially code the vowels: /i, I, ɛ, ae, a, ʌ, U, u/. Once transcribed, the vowels were classified into one of three groups: /a/-like vowels (/ae/, /a/, /ʌ/), /i/-like vowels /i/, /I/, /ɛ/), and /u/-like vowels (/U/, /u/). A second individual rescored 30% of the utterances, chosen randomly. Transcriptional reliability for the three-category coding was 92%.

Instrumental analysis was conducted by a person who was blind to the transcriptional classification of each utterance. Infants' vocalizations were analyzed using Kay Elemetrics microcomputer-based equipment (CSL, v. 4.0). The first two formants of each utterance were sampled at five locations: onset, at the 1/4, 1/2, and 3/4 points, and offset.

Two outcomes of the study are noteworthy for theory. First, there was developmental change in infants' vowel productions. Figure 1 displays the /a/-like, /i/-like, and /u/-like vowels of 12-, 16-, and 20-week-old infants in an F1/F2 coordinate space. In each graph, infants' vowel utterances are classified according to the phonetic transcription data. The closed circles enclose 90% of the utterances in each category. As shown, utterances in each of the three categories formed clusters in acoustic space. These clusters are, from an acoustic standpoint, relationally consistent with productions by adult speakers.

More intriguing from a developmental standpoint, the areas of vowel space occupied by infants' /a/, /i/, and /u/ vowels become progressively more *separated* between 12- and 20-weeks of age. Infant vowel categories were more tightly clustered at 20 weeks than at 12 weeks. What causes the increased separation of vowel categories over this relatively short (8 week) period? We suggest that infants listening to their ambient language have begun to form *stored representations* of vowels in memory; these representations serve as "targets" that infants try to match. Our view is that the stored representations resulting from infants' analysis of ambient language influences not only their perception of speech [see 11 for discussion] but their subsequent productions as well.

A second result of the study suggested that infants' vowel productions can be influenced by short-term exposure to sound. Infants' vowel productions were altered depending on what they heard. Infants produced more /a/-like utterances when exposed to /a/ than when exposed to /i/ or /u/; similarly, they produced more /i/-like utterances when exposed to /i/ than when exposed to /a/ or /u/; finally, they produced more /u/-like utterances when exposed to /u/ than when exposed to /a/ or /i/. In sum, the particular vowel stimulus infants heard influenced the type of vowel infants produced.

The total amount of exposure that infants received to their target vowel was only 15 minutes (5-min exposure to a specific vowel for each of three days). If 15 minutes of laboratory exposure influences infants' vocalizations, then listening to the ambient language for 20 weeks would appear to provide sufficient exposure to induce change.

How do 12 week-old-infants know how to move their articulators in a way that achieves a specific auditory target? Research on visually-specified movements shows that newborns imitate silent movements of the articulators, such as mouth opening and tongue protrusion [5]. We do not know if this kind of mapping from auditory to articulatory events exists at birth, but it is not out of the question, given Meltzoff and Moore's findings of visual-motor mappings of human mouth movements.

Even if primitive connections exist initially, they must be rapidly expanded to create the repertoire that infants possess just a short time later. This rapid expansion is gained, we believe, through

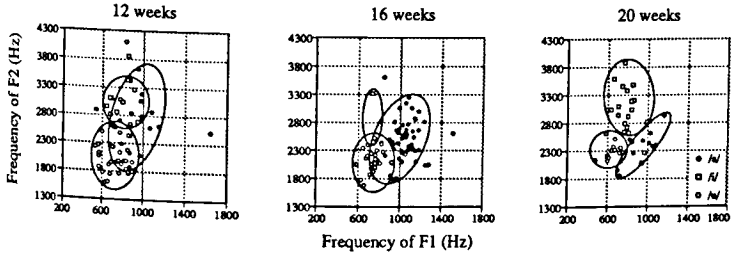


Figure 1. The location of /a/, /i/, and /u/ vowels produced by 12-, 16-, and 20-week-old infants. The curves were drawn by visual inspection to enclose 90% or more of the infants' utterances. Across age, infants' vowel productions show tighter clustering in vowel space. From Kuhl & Meltzoff (In press).

experience as infants engage in cooing and sound play. Infant cooing, which begins at about 4 weeks of age, allows extensive exploration of the nascent auditory-articulatory "map" during which (self-produced) auditory events are related to the motor movements that caused them. Presumably, infants' accuracy in producing vowels improves as infants relate the acoustic consequences of their own articulatory acts to the acoustic targets they heard. This implies that infants not only have to hear the sounds produced by others, but hear the results of their own attempts to speak. Hearing the sound patterns of ambient language (auditory exteroception) as well as hearing one's own attempts at speech (auditory proprioception) are critical to determining the course of vocal development.

Polymodal Speech Representations

Research on adults shows that speech perception is a polymodal phenomenon in which vision plays a role. This is indicated by auditory-visual "illusions" that result when discrepant information is sent to the two separate modalities [6,7,8].

Even very young infants appear to represent speech polymodally. We demonstrated that 18-20-week-old infants recognize auditory-visual connections, akin to what we as adults do when we lipread [9]. Four-month-old infants viewed two filmed faces, side by side, of a woman pronouncing two vowels silently and in synchrony, the vowel /a/ and the vowel /i/. Infants heard one of the two vowels (either /a/ or /i/), played in synchrony with the faces from a loudspeaker located midway between the two faces. The results of the test showed that infants who heard the vowel /a/ looked longer at the face pronouncing /a/ while the infants who heard /i/ looked longer at the vowel /i/. The experiment shows that by four months infants can recognize that an /a/ sound is mapped to a face articulating /a/ while an /i/ sound is mapped to an /i/ articulation. Infants' cross-modal speech perception abilities provide direct evidence that infants connect sound and articulatory movement when they observe another person speak.

Developmental Speech Theory

We would like to offer a more general developmental model about the connection between audition and articulation early in infancy. We emphasize two terms: representation and the influence of ambient language. Ambient language comes into play because it provides input to the child that shapes his or her perceptual space even before words are understood or produced [11,12]. Representation comes into play because we also believe that these speech patterns are stored in memory — that they are represented by the child and serve as "targets" which infants try to match when they vocalize.

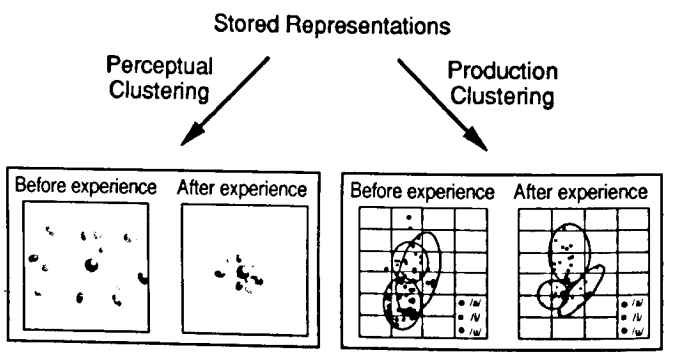


Figure 2. Infants' stored representations affect both speech perception, producing the perceptual clustering evidenced by the magnet effect, as well as speech production, producing the increased clustering seen in infants' vocalizations.

This framework is useful in interpreting the production study reported here. The study uncovered both long-term and short-term changes in infants' vocal repertoires. Long-term changes occurred over the eight-week period (from 12- to 20-weeks) during which infants' vocalizations were measured); short-term changes occurred in infants' vocalizations with relatively short periods of laboratory exposure that were manifest in vocal imitation. These findings suggest that infants' acquisition of speech is strongly influenced by the auditory information that surrounds them.

Two lines of research converge in speech development. On the one hand, there are the findings that linguistic exposure alters infants' *perception* of speech [11,12], and on the other hand, the result shown here that linguistic exposure alters infants' *production* of speech. These can be unified by the suggestion that stored representations of speech alter current processing, and that this occurs from the very earliest phases of infancy. On this view, the tighter clustering observed in infant vowel production and the tighter clustering among vowels in infant perception are both attributable to a common underlying mechanism — the *formation of representations that derive initially from perception of the ambient input and then act as targets for motor output.* The speech representational system is thus deeply and thoroughly polymodal. Early listening affects both sensory perception and motor learning.

ACKNOWLEDGMENT

Research reported here was supported by grants from NIH (HD-22514, HD-18286, and DC 00520).

REFERENCES

[1] Stoel-Gammon, C. (1992), "Prelinguistic vocal development: Measurement and predictions", in *Phonological development: Models, research, implications*, edited by Ferguson, C.A., Menn, L., Stoel-Gammon, C., pp. 439-456. Timonium, MD: York.

[2]. Kuhl, P. K., Meltzoff, A. N. (In press), "Evolution, nativism, and learning in the development of language and speech", in *The biological basis of language*, edited by Gopnik, M., New York: Oxford University Press.

[3] de Boysson-Bardies, B. (1993), "Ontogeny of language-specific syllabic productions", in *Developmental neurocognition: Speech and face processing in the first year of life*, edited by de Boysson-Bardies, B., de Schonen, S., Jusczyk, P., McNeilage, P., Morton, J., pp. 353-363. Dordrecht, Netherlands: Kluwer.

[4] Kuhl, P. K., Meltzoff, A. N. (In press), "Infant vocalizations in response to speech: Vocal imitation and developmental change", *Journal of the Acoustical Society of America.*

[5] Meltzoff, A. N., Moore, M. K. (1994), "Imitation, memory, and the representation of persons", *Infant Behavior and Development*, vol. 17, pp. 83-99.

[6] Green, K. P., Kuhl, P. K., Meltzoff, A. N., Stevens, E. B. (1991), "Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect", *Perception & Psychophysics*, vol. 50, pp. 524-536.

[7] Massaro, D. W., Cohen, M. M. (1994), "Auditory/visual speech in multimodal human interfaces", in *Proceedings of the International Conference on Spoken Language Processing*, pp. 531-534. Tokyo: Acoustical Society of Japan.

[8] Kuhl, P. K., Tsuzaki, M., Tohkura, Y., Meltzoff, A. N. (1994), "Human processing of auditory-visual information in speech perception: Potential for multimodal human-machine interfaces", in *Proceedings of the International Conference on Spoken Language Processing*, pp. 539-542. Tokyo: Acoustical Society of Japan.

[9] Kuhl, P. K., Meltzoff, A. N. (1982), "The bimodal perception of speech in infancy", *Science*, vol. 218, pp. 1138-1141.

[10] Kuhl, P. K. (1994), "Learning and representation in speech and language", *Current Opinion in Neurobiology*, vol. 4, pp. 812-822.

[11] Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., Lindblom, B. (1992), "Linguistic experience alters phonetic perception in infants by 6 months of age", *Science*, vol. 255, pp. 606-608.

[12] Kuhl, P. K. (this volume), "Mechanisms of developmental change in speech and language".

ARTICULATORY PREFERENCES IN FIRST WORDS: THE FRAME CONTENT HYPOTHESIS

Barbara L. Davis (University of Texas at Austin, Department of Speech Communication), Peter F. MacNeilage (University of Texas at Austin, Department of Linguistics)

ABSTRACT

Phonetically transcribed utterances for 1.) babbling, 2.) early speech and 3.) concurrent babbling in three normally developing infants were analyzed to asses the ratio of labials to alveolars in the three contexts. Prediction based on the "Frames then Content" hypothesis was that early words would show a predominence of labials indicating a return to more simple production patterns based on the requirements of simultaneous lexical and phonetic coding. Results showed predominence of labials in first words contrasted with a predominence of alveolars in babbling for all three infants.

INTRODUCTION

In contrast to Roman Jakobson's early theory [1] proposing discontinuity in sound use between babbling and speech, recent work has shown continuity between the two [2, 3]. With few exceptions, output patterns in babbling seem to correspond to output patterns in first words. Phones most frequently used to describe canonical babbling and early speech output in phonetic transcription studies are stops [b], [d]; nasals [m], [n],glides [j], [w], and [h] [4] and vowels [ɛ], [æ], [a], [ʌ], [ɔ] [5, 6] in CV and CVCV forms. Acoustic studies of early vowel development [7, 8, 9, 10, 11] are consistent with transcription based studies in showing early preferences for vowels located in the lower left quadrant. Evidence of continuity from prelinguistic behaviors to early words in both sound preferences and temporal organization increases the importance of understanding babbling as a crucial first phase of development toward first word production patterns.

Some evidence suggests that an aspect of use of babbling patterns in first words consists of even higher frequency of usage of certain aspects of babbling patterns in words than in babbling. For example, the favored number of syllables in babbling is one, the favored consonant is a stop consonant and the favored mode of consonant repetition, reduplication. There is evidence that these preferences actually increase in first words and concurrent babbling. Alveolars are the most frequent stops in babbling [12]. It has been noted that there is a strong trend towards favoring labials in first words and concurrent babbling, both in English and in other languages [13]. It is tempting to propose that with the onset of the demand to interface the lexicon with the motor system, the infant enters a conservative motor phase in which he/she focuses mainly on the simplest of available motor capacities. However, a simplicity based hypothesis has a circular quality as simplicity is taken as synonymous with frequency rather than being defined independently. This circularity has been a persistent problem with Markedness which is fundamental to phonological theory, and in principle is a problem for any approach that emphasizes relative frequencies. In this case the claim that this incidence of labials represents regression to simpler forms is not necessarily circular. An additional finding from babbling-to speech studies offers a potential means of avoiding the circularity problem for a simplicity hypothesis.

MacNeilage & Davis [14, 15, 16] have proposed a "Frames then Content" metaphor to describe spatio-temporal and biomechanical characteristics of babbling and continuity in transition to early speech output. Frame applies to the rhythmic regularity of mandibular oscillation cycles resulting in listener perception of syllable-like and therefore speech-like output. It is claimed that close and open phases of the cycle may often have no associated neuromuscular activity other than movement of the mandible and consequently no sub-syllabic organization of Content elements. The syllabic Frame thus constitutes the earliest temporal envelope within which segment specific Content elements develop as the child gains increasing independence of control over articulators in speech movement sequences. This perspective allows testable hypotheses regarding organization of babbling and speech The CV Co-occurence Hypothesis predicts strong associations between labial closure and central vowel open phase, alveolar closure and front vowel open phase, and velar closure with back vowel open phase as emerging from mandibular oscillation in the temporal domain. The Variegation Hypothesis predicts manner changes across successive closure (consonant) phases and height changes over successive open (vowel) phases, consistent with the principle of mandibular oscillation during output sequences. Both hypotheses are generated from an ease of articulation perspective suggesting the driving force behind sound qualities observed in pre-speech output is production based rather than perceptual. In a study of six normal infants [17], tests of the CV co-occurrence hypothesis showed either significant associations or expected trends for the three consonant places of articulation. A potential conclusion from these results is that earliest CV relationships represent a lack of independence in place of articulation between close and open phases of canonical sequences. Tests of the variegation hypothesis revealed highly significant trends for height over front back changes for vowels and manner over place changes for consonants for all subjects. This result is suggested as being due to change in tongue height based on degree of openness of the jaw over successive cycles.

The default mode for producing speech like output is considered to be found in the labial central vowel association as no tongue presetting is required to realize the sound qualities produced. A predominence of labials in first words would thus allow an additional piece of evidence to confirm the ease of articulation hypothesis suggested by "Frames then Content".

METHOD

Data analyzed for this study were collected as a part of a larger longitudinal project tracking early normal speech development from the onset of canonical babbling through age 3 1/2. Data for three infants were analyzed for this study. Normal development was established through parent case history report. In addition, each infant was administered the Battelle Developmental Screening Inventory [18] and hearing screening using sound field techniques.

Three observers collected and analyzed data for the infants. Each observer tracked the same infant over the course of the study. One hour weekly sessions were audio taped in the subject's home. An ATW-20 digital audio recorder was used for both data collection and subsequent transcription. Each infant wore an Audiotechnika ATW-1030 remote microphone in a cloth vest. The microphone was clipped at the shoulder to keep a consistent mouth to microphone distance and to keep the infant from handling the microphone.

Data selected for analysis included all speech-like canonical babbling and word forms occurring during the sessions. Vocalizations analyzed were produced with an eggressive airstream, including minimally a consonant like closure phase (articulatory obstruents, sonorants, and glides), and vowel like open phase within a single utterance string. This criterion resulted in either CV or VC monosyllables as minimal units for analysis; polysyllables included CV or VC alternations. The non-oral closant /h/ was included when it was present in rhythmically alternating sequences. All utterance strings analyzed were comfort state vocalizations produced without simultaneous background noise or speech. Tokens selected as single utterance strings were bounded by 1 second of silence, noise, or adult vocalization.

All utterances which met these criteria were phonetically transcribed

by the primary transcriber using broad phonetic transcription supplemented by diacritics available for infant speech [19]. All transcribed data was entered using a phonetic keyboard and software developed for analysis of infant data [20]. Data analyzed for this investigation included 8158 alveolar and labial consonants.

RESULTS

Results for all three subjects are displayed as ratios of alveolar consonants to labial consonants. Included in Table 1 are ratios for prespeech babbling, concurrent babbling and speech, and first words.

Table 1. Ratios for prespeech babbling, concurent babbling and first words

	Prespeech Babbling	Concurrent Babbling	First Words
S1	1.35	.50	.28
S2	4.33	3.34	.35
S3	1.65	1.76	.40

All three subjects favored alveolar consonants in prespeech babbling. Ratios were 1.35, 4.33, and 1.65. In first words, all three favored labials strongly. Ratios were .28, .35, and .40. One subject also favored labials in babbling produced concurrently with first words (S2; ratio 3.34).

DISCUSSION

The dominant perspective in the area of early speech production has been one of linguistic theory, which includes the use of linguistic formalisms to capture regularities in early child productions. The object of formal linguistic inquiry is an abstract system shared by members of a community rather than the actual phonetics of speech output. The consequence for study of acquisition is emphasis on phonological contrast rather than on details of output; the message level rather than the level of signal generation. One statement of this position on speech development is that of Macken [21] "Phonology is a cognitive/linguistic system which exists independently of the phonetic system on which it is based" (pg.436). At the same time, the limitations children have in producing speech correctly have frequently been acknowledged as being largely motor control limitations, without a corresponding attempt to develop a coherent theory of speech motor control.

A second consequence of the formalist approach to speech development has been a relative emphasis on the onset of meaningful speech as the legitimate starting point for description, although recent work has shown continuity between late babbling and early speech and dramatic similarities across languages in babbling and early speech inventories. A coherent theory of speech motor control in early development thus might more properly be seen as beginning at the onset of canonical babbling, the initial manifestation of speech-like motor behavior.

Findings of this study, viewed in the context of "Frames then Content" as well as in the context of these infants pre-speech babbling preferences for alveolars support the strength of a production-based explanation for early speech patterns as an alternative to development of linguistic categories based on perceptual distance; extensions of the child's mechanical production constraints rather than as rule-driven cognitive operations. Predominance of labials in first words related to preference for alveolars in babbling is viewed as use of basic production patterns to realize early lexical items. At the very least, more knowledge of early motor constraints will allow more careful evaluation of claims related to cognitive or phonological factors which have been suggested as being independent of motor constraints. This independence must develop at some point. The question then arises of how the child achieves this independence if it is not a given as a starting point.

ACKNOWLEDGMENTS

This work was supported in part by NICHD-R01-HD27733-03.

REFERENCES

[1] Jakobson, R. (1968). *Child language, aphasia, and phonological universals*. The Hague: Mouton.

[2] Vihman, M., Macken, M., Miller, R., Simmons, H., & Miller, J. (1985). From babbling to speech: A reassessment of the continuity issue. *Language*, 60, 397 -

[3] Vihman, M., Ferguson, C. A. & Elbert, M. (1986). Phonological development from babbling to speech: Common tendancies and individual differences. *Applied Psycholinguistics* 7:3-40

[4] Stoel-Gammon, C. (1985). Phonetic inventories 15-24 months: A longitudinal study. *Journal of Speech and Hearing Research*, 23, 506-512.

[5] Davis, B.L. & MacNeilage, P.F. (1990). The acquisition of vowels: a case study. *Journal of Speech and Hearing Research* , V 33, pp.16-27.

[6] Stoel-Gammon, C. and Harrington, P. (1990). Vowel systems of normally developing and phonologically disordered children. *Clinical Linguistics and Phonetics*, 4, 145-160.

[7] Buhr, R.D. (1980). The emergence of vowels in an infant. *Journal of Speech and Hearing Research*, 23, 73-94.

[8] Bickley, C. (1983). Acoustic evidence for phonological development of vowels in young children. Paper presented at the *Tenth International Congress of Phonetic Sciences*, Utrecht, Holland.

[9] Kent, R.D. and Murray, A.D. (1982). Acoustic features of infant vocalic utterances at 3,6, and 9 months. *Journal of the Acoustical Society of America*, 72, 353-365.

[10] Kent, R.D., and Bauer, H.R. (1985). Vocalizations of one year olds. *Journal of Child Language*, 12, 491 - 526.

[11] Lieberman, P. (1980). On the development of vowel productions in young children. In G. Yeni-Komshian, J.F. Kavanagh, & C.A. Ferguson (Eds.), *Child Phonology: Vol.1: Production* . (pp. 23-42). New York: Academic Press.113-142.

[12] Locke, J. (1983). *Phonological Acquisition and Change*. New York: Academic Press.

[13] Boysson Bardies, B. (1993). Ontogeny of language-specific syllabic production. In B. Boysson Bardies, S. de Schoen, P. Jusczyk, P. MacNeilage, and J. Morton, (Eds.), *Developmental Neurocognition: Speech and Face Processing in the First Year of Life*. Dordrecht: Klewer Academic Publishers, 353-363.

[14] MacNeilage, P.F. & Davis, B.L. (1990a). Acquisition of speech production: Frames, then content. In Jeannerod, M. (Ed.) *Attention and PerformanceXIII: Motor representation and control*, Hillsdale, N.J.: Lawrence Erlbaum. 453-475.

[15] MacNeilage, P.F. and Davis, B.L. (1990b). Acquisition of speech production: The achievement of segmental independence. In Hardcastle, W.J. & Marchal, A. (Eds.) *Speech Production and Speech Modeling*, Kluwer: Dordrecht. 55-68.

[16] MacNeilage, P.F. and Davis, B. L. (1993). Motor explanations of babbling and early speech patterns. In B. Boysson Bardies, S. de Schoen, P. Jusczyk, P. MacNeilage, and J. Morton, (Eds.), *Developmental Neurocognition: Speech and Face Processing in the First Year of Life*. Dordrecht: Klewer Academic Publishers.

[17] Davis, B.L. & MacNeilage, P.F. (1995, In press). The articulatory basis of babbling, *Journal of Speech and Hearing Research*.

[18] Guidubaldi, J., Newborg, J., Stock, J.R., .Svinicki, J., Wneck, L. (1984). *Battelle Developmental Inventory*. Allen, Texas, DLM Teaching Resources.

[19] Bush, C.N., Edwards, M.L., Edwards, J.M., Luckau, C.M., Macken, M.A. and Peterson, J.D. (1973). On specifying a system for transcribing consonants in child language. *Stanford Child Language Project.*, Department of Linguistics, Stanford University, Stanford California.

[20] Oller, D.K. (1990) *Logical International Phonetic Programs*, Intelligent Hearing Systems, Miami, Florida

[21] Macken, M. (1993). Harmony and melody templates in early words. In B. Boysson Bardies, S. de Schoen, P. Jusczyk, P. MacNeilage, and J. Morton, (Eds.), *Developmental Neurocognition: Speech and Face Processing in the First Year of Life*. Dordrecht: Klewer Academic Publishers, 435-446.

FINAL LENGTHENING AT PROSODIC BOUNDARIES IN DUTCH

E. Hofhuis, C. Gussenhoven and A. Rietveld
Department of Language and Speech
Nijmegen University
Nijmegen, The Netherlands

ABSTRACT

In this paper we describe an experiment that was set up to measure segmental lengthening before five types of prosodic boundaries, ranging from the Prosodic Word boundary to the Utterance boundary.

INTRODUCTION

It has been shown by several researchers [1], [2] that segments are longer at syntactic boundaries, and that the amount of lengthening increases with the boundary's place in the syntactic hierarchy. However, we assume that it is prosodic structure that regulates the rhythm of language, and that final lengthening therefore occurs at prosodic boundaries. In earlier experiments we have found this to be true for boundaries below the word level [3]. In the experiment described below we investigated final lengthening at boundaries ranging from the Prosodic Word boundary to the Utterance boundary.

METHOD

Our experiment was set up to test the influence prosodic boundaries have on the durations of the segments that precede them. We based our definitions of the relevant prosodic boundaries in Dutch on [4].

We devised five carrier sentences in which target words could be placed before one of five prosodic boundaries. The lowest boundary we tested was a Prosodic Word boundary within a compound. The next boundary was a Prosodic Word boundary at the end of a morphological word, for which the target word was an adjective within an NP. In prosodic theory there is no difference between these two boundaries, although morpho-syntactically there is. The next higher boundary to be tested was the Phonological Phrase (PPh) boundary, which occurred at the end of an NP in our material. The highest boundary was the Utterance boundary.

To rule out any possible effect of sentence length we made sure that all carrier sentences had the same number of words before and after the target word position. Since it is not clear whether the shortening effect of the number of words following a target word can pass the Utterance boundary, or alternatively, whether the Utterance-final lengthening effect is distinct from the lengthening before the end of a discourse, we added a small sentence after the Utterance boundary, consisting of the same two words that followed the PPh-boundary. In order to be able to answer this question we also included the Utterance boundary without this following sentence in our materials.

Material

It has been pointed out by [5], among others, that segment classes may differ in the amount of lengthening they show at boundaries. Therefore, we chose target words ending in segments from four consonant classes, each of which followed a long as well as a short vowel. This resulted in the following target words:

Table 1. Target words.

	liq	nas	fric	stop
V	bɑr	kɑn	pɑs	mɑt
VV	bar	kan	pas	mat

We also included bisyllabic target words in our material, but we will not go into that part of the experiment in this paper.

The target words were placed in carrier sentences in the five boundary positions described above. This led to meaningful sentences in nearly all cases. Our speakers received instructions about the non-meaningful cases to enable them to treat these sentences as normal, meaningful sentences. The sentences were read by two male native speakers of Dutch. Each item was repeated ten times by each speaker. The sentences were recorded in a sound-proof studio. Durations were measured using a wave-form segmenting program.

RESULTS

We performed ANOVA's on each of the four subsets (liquids, fricatives, nasals and stops). For every subset the variable 'boundary' had a significant effect on the vowel and the consonant directly preceding the boundary. For liquids and their preceding vowels this was $F(4,172)=111.4$, $p<.001$ and $F(4,172)=27.1$, $p<.001$ respectively. For nasals the values are $F(4,173)=143.4$ $p<.001$ for the vowel and $F(4,173)=161.76$, $p<.001$ for the nasal. For fricatives they are $F(4,176)=179.3$, $p<.001$ for the vowel and $F(4,176)=471.38$, $p<.001$ for the fricative and finally for stops: $F(4,169)=71.45$, $p<.001$ for the vowel and $F(4,169)=47.35$, $p<.001$ for the stop. This means that the durations of every type of consonant and all vowels preceding them are significantly influenced by the type of boundary they precede. Onsets were never significantly influenced by their boundary position.

To find out which of the boundaries contributed to this effect we performed a post-hoc analysis (Tukey's HSD). To avoid large within-group variance we did separate post-hoc tests for long and short vowels. In the four tables below we can see the means for vowels and following consonants, for each segment class. The digits correspond to prosodic boundaries, 1 is the Prosodic Word boundary within composite words, 2 the final Prosodic Word boundary, 3 the Phonological Phrase boundary, 4 the Utterance boundary followed by a second sentence and 5 the Utterance boundary without this sentence. Values that are significantly different from the preceding values are underlined.

Table 2. Means in ms. for target words ending in liquids.

	1	2	3	4	5
ɑ	119	120	122	<u>149</u>	145
r	40	42	44	<u>88</u>	79
a	161	170	163	<u>180</u>	171
r	43	41	45	<u>80</u>	71

Table 3. Means in ms. for target words ending in nasals.

	1	2	3	4	5
ɑ	80	84	<u>96</u>	<u>111</u>	113
n	46	48	51	<u>81</u>	90
a	139	135	<u>153</u>	<u>164</u>	170
n	42	40	47	<u>79</u>	82

Table 4. Means in ms. for target words ending in fricatives.

	1	2	3	4	5
α	97	103	104	133	143
s	67	71	76	145	182
a	160	159	162	185	208
s	66	69	72	117	193

Table 5. Means in ms. for target words ending in stops.

	1	2	3	4	5
α	86	86	85	101	100
t	36	37	41	55	58
a	144	140	133	155	169
t	37	33	41	49	71

Boundaries

Looking at the tables above we find some interesting results. To begin with, we never found a significant difference between boundaries 1 and 2. Since there is no phonological difference between 1 and 2 in the theory of prosodic constituency we adopted, this means that segment durations, in these cases, reflect prosodic structure rather than morpho-syntactic structure. In part, this also holds for boundaries 4 and 5. In most cases there was no difference between segment durations before these two boundaries. The fricatives (and their preceding vowels), however, showed more lengthening in 5 than in 4 and so did the long vowel before the stop. Prosodically, 4 and 5 are identical: they are both Utterance boundaries. But phonetically they differ: in 5 the boundary is 'discourse'-final, whereas in 4 another utterance followed in the same discourse. It is well known that the number of words following a target word has an effect on its duration. It is not clear, however, whether this effect can cross the Utterance boundary. The results described above suggest that it can, especially in cases where extreme lengthening is possible, as is the case with fricatives.

The effect of the Phonological Phrase boundary can only be observed in table 3, showing the words ending in nasals. Vowels that preceded nasals were significantly longer before PPh-boundaries than before word boundaries.

Segments

As has been pointed out in [5], there has been some discussion on the question which part of the syllable is lengthened, and which segments can be lengthened before boundaries. For example, in [6] it is said that most of the syllable lengthening before utterance boundaries is due to lengthening of the vowel. It is also assumed in [6] that only sonorant and continuant segments can be lengthened. In [7], however, it appears that final lengthening largely affects the later part of the syllable. In [5] it was found that stops may show considerable lengthening, even more than the preceding vowel.

When we look at tables 2-5, we see that in our material the largest share of preboundary lengthening is not borne by the vowel but by the following consonant, as was found in [5] and [7]. This was true for all segment classes, including stops. Fricatives were lengthened most (up to 272%), but even stops were lengthened by 192% after long vowels. On the whole, the values for the different classes are not as far apart as might be expected.

CONCLUSION

The experiment we described above shows that higher prosodic boundaries trigger more final lengthening than lower prosodic boundaries. An interesting finding was that compound-internal Prosodic Word boundaries have the same effect as word final ones. This means that segment duration reflects the Prosodic Word boundary instead of the morpho-syntactic word boundary. The effect of the Phonological Phrase boundary could only be observed in words ending in nasals. Thus, at least in some cases, this boundary affects the duration of the segments before it.

We found a difference between a discourse-final Utterance boundary and an Utterance boundary that occurs before another sentence. This suggests that the Utterance boundary may not be the highest boundary that needs to be recognised, or alternatively, that the shortening effect that following words have on the target word may cross Utterance boundaries.

Our experiment confirmes the findings in [5] and [6] that final lengthening affects the vowel as well as the consonant following it, but that it is the latter which is lengthened most, even when this is a stop.

REFERENCES

[1] Crystal, T.H. & House, A.S. (1988a), Segmental durations in connected-speech signals: current results, *JASA*, vol 83, pp. 1553-1573.

[2] Klatt, D.H. (1975), Vowel lengthening is syntactically determined in a connected discourse, *Journal of Phonetics*, vol. 3, pp. 129-140.

[3] Hofhuis, E. (1993), Establishing prosodic structure by measuring segment duration, *Working Papers*, vol. 42, Lund University, Dept. of Linguistics, pp. 136-139.

[4] Nespor, M. & Vogel, I. (1986), *Prosodic Phonology*, Dordrecht: Foris.

[5] Berkovits, R. (1993b), Utterance-final lengthening and the duration of final-stop closures, *Journal of Phonetics*, vol. 21, pp. 479-489.

[6] Cooper, W.E. & Paccia-Cooper, J. (1980), *Syntax and Speech*, Cambridge: Harvard University Press.

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

NOTES ON SYLLABLE DURATION IN FRENCH AND SWEDISH

Anita Kruckenberg and Gunnar Fant

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

ABSTRACT

We have studied the durational contrast of stressed and unstressed syllables as a function of the number of phonemes per syllable. Observed differences in regression lines for the French and the Swedish data are discussed with respect to the concepts of syllable timing and stress timing. We have also studied the effect of tempo in French. A noteworthy observation is the constancy of pre-pause syllable durations with tempo variations.

INTRODUCTION

In our earlier study [1] comparing the reading of a one minute long passage from a Swedish novel which was translated into English and French, we measured average syllable durations within a stressed/unstressed, i.e. an accented/unaccented, labelling and as a function of the number of phonemes per syllable. The smaller durational contrasts between stressed and unstressed syllables in French than in Swedish and English gave a support for the established notion of French as a syllable timed language and Swedish and English as stress timed languages. The concept of "stress timing" is attributed not to a physical isochrony but to an overall relative greater auditory prominence of the alternation between stressed and unstressed syllables.

The present study provides data on the same text for one more French subject and variations of reading speed and it provides a more detailed comparison of Swedish and French data. For details about the text see [1]. A perceptual binary labelling of syllables as stressed and unstressed performed by three trained listeners revealed that in all three languages about 90 % of the stresses were found in content words including adverbs. Comparing syllables of the same number of phonemes we found about 50 ms stressed/unstressed contrast in French and about 120 ms in Swedish. The average number of phonemes per unstressed syllable was 2.1 in French and 2.25 in Swedish, while stressed syllables showed a marked difference, 3.0 in Swedish and 2.3 in French, which contributes to the overall greater durational contrast in Swedish. In the present study we have investigated to what extent the particular dominance of phonemes of inherently long duration should be taken into account when comparing stressed and unstressed syllables. We have also made a more detailed analysis of pre-pause terminal stress in French versus the "minor stresses" within a phrase and in relation to speech tempo.

SWEDISH REFERENCE DATA

Figure 1 illustrates average syllable duration as a function of the number of

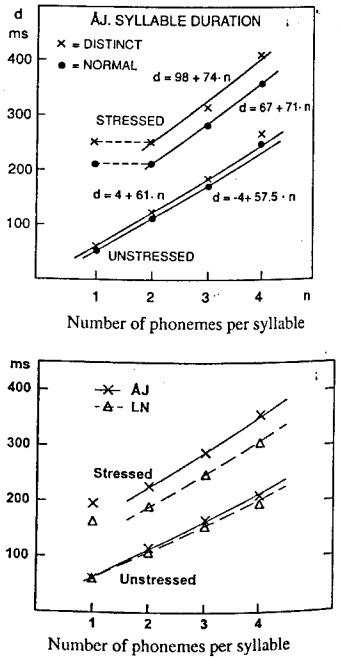


Figure 1. Syllable duration from Swedish prose reading.

phonemes in stressed and unstressed syllables of the original Swedish prose text. The two speakers in the lower graph differ significantly in the duration of stressed syllables but marginally only in terms of unstressed syllables.

A quite similar relation is to be seen in the upper graph which is a comparison of our reference subject ÅJ reading the text twice, in a distinct mode and in a normal mode. There is an apparent stability of the duration of unstressed syllables whereas the stressed/unstressed contrast is a speaker and speaking specific feature. In 2-phoneme syllables the stressed/unstressed difference was 98 ms and 112 ms in 3-phoneme syllables. In the distinct reading mode the contrast was 25 ms greater than in the normal speaking mode. Unstressed syllables averaged 125 ms and stressed syllables 290 ms.

A COMPARISON OF FRENCH AND SWEDISH

A typical French prosodic phrase is illustrated in Figure 2, "Le long de trois des murs...". The intonation contour here marks three so called prosodic words, i.e. three "stress groups", each containing a content word with an F_0 rise of the order of 6 semitones in the vowel, in non-final groups followed by a corresponding F_0 fall. The recurrent pattern of F_0 rises and falls in successive prosodic words within a phrase and the associated lengthening of stress-group final syllables produces a rhythmical regularity. The large duration of the phrase-final syllable [my:r] is apparent.

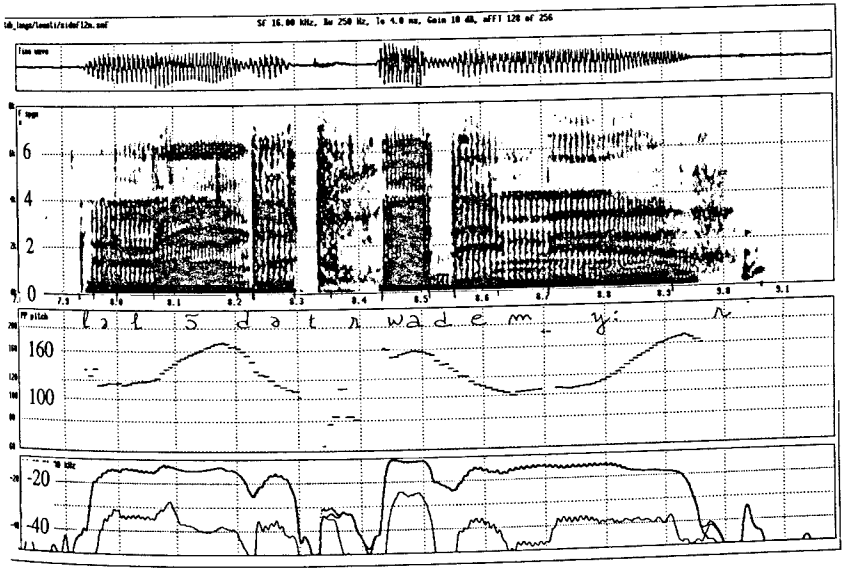


Figure 2. French prosodic phrase, "Le long de trois des murs". Oscillogram, spectrogram, log F_0 and intensity, LP 1000 and HP 1000 Hz (below).

Figure 3 shows average syllable duration as a function of the number of phonemes per syllable comparing two French speakers and our Swedish reference speaker reading the same base text. Phrase-final syllables are excluded. The stressed/unstressed contrast is apparently lower in the French than in the Swedish data. Regression lines for stressed and unstressed data converge for the French data but diverge for the Swedish data. This compression versus expansion of the duration of complex syllables adds to the language specific contrasts and can be quantified with reference to the constants of the regression equations.

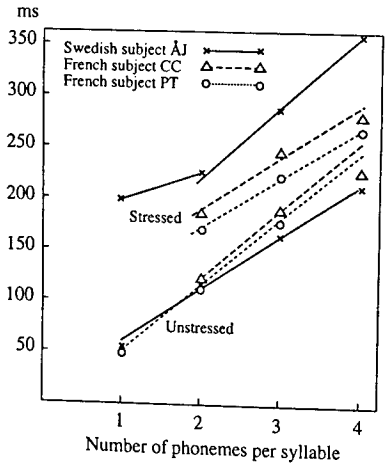


Figure 3. Syllable duration. One Swedish and two French subjects.

Table 1. Regression equations for syllable duration d=a+bn as a function of the number of phonemes per syllable.

	Unstressed	Stressed
Swedish, ÅJ	d=9+51n	d=84+67n
French, CC	d=-11+66n	d=89+50n
French, PT	d= -10+63n	d=76+48n

In Swedish the slope factor b is greater for stressed than unstressed syllables and the reverse is true for French.

Figure 4 contrasts high and low speaking rate for the French subject PT. The higher tempo reduces durations of both stressed and unstressed syllables to the effect that unstressed syllables of the lower speaking rate attain approximately the same durations as stressed syllables of the higher speaking rate. In Swedish the durational contrast is also reduced but to a less extent [2].

A wellknown feature is the extra long duration of phrase-final syllables in French [3]. A remarkable finding is that they stay approximately constant with reading rate. A study of pausing showed that in addition to 8 sentence final pauses there occurred 12 pauses within sentences at normal and high speech rate and 16 at a low speech rate. The more frequent pausing in slow speech is a well known phenomenon [4]. The effective reading time, pauses excluded, was 16% higher in the low than in the normal rate

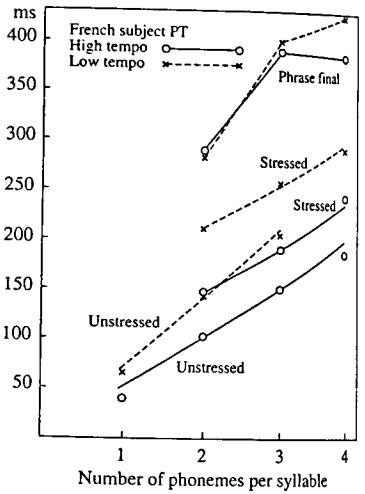


Figure 4. Syllable duration at high and low speech rate. French subject.

and 13% lower in the high than in the normal speaking rate. The ratio of pause time to effective reading time was close to 38 % at both normal and low speaking rate and 22% at the high speaking rate.

PHONEME DURATIONS AND INHERENT LENGTH

These corpora are too small to allow a representative listing of the durations of individual vowels and consonants but there is a basis for considering certain group data, i.e. to what extent syllable duration is influenced by the relative occurrence of phonemes of inherently long durations and also how the stress induced lengthening affects the vowel and associated consonants. As already stated in [1], stress in French induces almost no lengthening of consonants following a vowel in a non-terminal syllable, the main effect to be observed is in the vowel and preceding consonants. In Swedish, consonants following a short stressed vowel carry a large part of the syllable lengthening. In English preceding and following consonants have about equal importance.

An example of segment duration profiles in French unstressed, stressed non-terminal, and stressed final syllables is shown in Figure 5. It pertains to 3-phoneme syllables of type CCV and CVC with segment notations C-2, C-1, V and C-1, V, C+1.

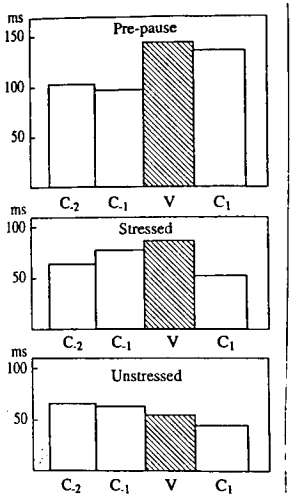


Figure 5. French segment durations in pre-pause, stressed and unstressed syllables.

It verifies the importance of the vowel as a primary object of stress induced lengthening and of consonants in non-terminal locations. In terminal prepause locations the final consonant adds significantly to the syllable duration.

An analysis of stressed versus unstressed CV sequences in non-terminal positions provided the following results. For speaker PT the difference in vowel duration was 28 ms and in consonant duration 19 ms, i.e. a total stressed/unstressed contrast of 47 ms. However, the more frequent occurrence of consonants of a relative large duration in stressed than in unstressed syllables, fricatives and unvoiced stops and also [b], accounted for 11 ms of the 19 ms difference in consonant duration. The stressed induced consonant lengthening was thus merely 19-11=8 ms. Similarly, the more frequent occurrence of nasal vowels and of the diphtong [wa] in stressed syllables contributed to 4 ms of the stressed/unstressed contrast. The true stressed induced vowel lengthening was thus 28-4=24 ms. Of the total C+V difference of 47 ms a net of 32 ms represented a true lengthening associated with stress. Similar figures were attained for speaker CC, i.e. 4 ms for inherent vowel duration and 8 ms for inherent consonant durations and a total of C+V=34+28=62 ms stressed/unstressed difference of which 50 ms represents a true segment lengthening. Differences in inherent durations may also explain a part of the stressed/unstressed convergence in the French data, Figure 3. In Swedish the role of phoneme inherent durations was found to be insignificant since the distributions were quite similar in stressed and unstressed syllables.

CONCLUSION

In French a part of the observable average durational difference between stressed and unstressed syllables of the same number of phonemes is due to a larger proportion of phonemes of a relatively long inherent duration in stressed syllables. Regression lines for stressed and unstressed syllables converge in French but diverge in Swedish. These findings add to the two basic components of durational contrast, that of stress induced lengthening and the dominance of 2-phoneme syllables in French. The smaller total durational contrast between stressed and unstressed syllables in French supports the view of French as a syllable timed and Swedish as a stress timed language. The rhythm of French is, however, a more complicated question. The sequence of prosodic words also carry a rhythmical pattern.

ACKNOWLEDGEMENT

This work has been supported by a grant from the Bank of Sweden Tercentenary Foundation, RJ.

REFERENCES

[1] Fant, G., Kruckenberg, A. & Nord, L. (1991), "Durational correlates of stress in Swedish, French and English", *Journal of Phonetics* 19, 1991, pp. 351-365.

[2] Fant, G., Kruckenberg, A. & Nord, L. (1991), "Some observations on tempo and speaking style in Swedish text reading", ESCA Workshop on the phonetics and phonology of speaking styles, Barcelona 30 Sept. - 2 Oct. 1991.

[3] Crompton, A. (1980), "Timing Patterns in French", *Phonetica* 37, pp. 205-234.

[4] Vaissière, J. (1983), "Language-independent prosodic features," in *Prosody: Models and Measurements*, Ed. A. Cutler and D.R. Ladd, Springer Verlag.

UTTERANCE-FINAL LENGTHENING: THE EFFECT OF SPEAKING RATE

F. Bell-Berti, C. E. Gelfer, and M. Boyle
St. John's University, Jamaica, NY, Haskins Laboratories, New Haven, CT, William Paterson College, Wayne, NJ, & Burke Rehabilitation Hospital, White Plains, NY

ABSTRACT

The phenomenon of utterance-final lengthening is a fairly ubiquitous one. However, there appear to be limits to the extent that the speech system can slow itself and still operate under the normal control regime. This study found that while final lengthening occurs at fast and comfortable speaking rates, it did not occur when speech was slowed. Rather, the effect of slowing speech was to increase the relative durations of vowels in utterance initial syllables.

INTRODUCTION

Although utterance-final lengthening has been widely studied, the origin of this effect has not been established. One possibility is that final lengthening is determined by linguistic characteristics of an utterance (that is, the inherent segment durations and the phonetic, semantic, and syntactic context in which the segment occurs). So, for example, it has seemed reasonable to suggest [1,2] that final lengthening is a planned effect provided to listeners to help them identify syntactic boundaries. Indeed, it has been shown [3] that listeners expect longer durations for words in phrase- and sentence-final positions. Another possibility, however, is that final lengthening arises from neurological, muscular, and mechanical characteristics of the speech system.

We have previously reported studies of ataxic dysarthric speakers [4,5] in whose speech there was an absence of utterance final lengthening. The question that arose from that result was whether final lengthening failed to occur because utterance-final events were affected differentially by the underlying neurological impairment, or simply because these speakers were already speaking so slowly that they could not slow their speech any further. We hypothesized that one would find the same absence of final lengthening in unimpaired speakers if they simply spoke slowly.

METHODS

We used four utterance types (Table 1), two five- and two seven-syllable sentences. The target word 'pat' occurred either in medial or final position at both sentence lengths. Five women between 28 and 36 years of age, with no history of speech or hearing difficulties, served as subjects. At the time of recording, all were recent graduates of a master's degree pro-gram in speech-language pathology. They were unaware of the specific purpose of the experiment until after the recordings were completed. They produced a set of sentences at three self-selected speaking rates: natural, fast, and slow.

Table 1. List of sentences

Short Sentences:
Medial: Seek a pat music.
Final: Seek a music pat.
Long Sentences:
Medial: Seek a pat grand musical.
Final: Seek a grand musical pat.

The subjects listened to a recorded, natural rate exemplar of each sentence and saw a written version of the sentence each time it was to be repeated at each rate. The only special instructions were to produce the sentences without pauses. They were given practice trials until they were comfortable with the task. No subject wanted more than two trials of each sentence at the slow rate, and none at the natural and fast rates.

The sentences were presented in six random orders; in each, the subject was asked to repeat each of the sentences twice at her natural, then twice at her slow, and finally, twice at her fast rate. In all, each subject produced 12 repetitions of each sentence at each rate (144 sentences/subject, 720 sentences in all).

Audio recordings were made using a Marantz PMD 221 cassette recorder and Sony unidirectional, low impedance, dynamic microphone. The signals were digitized at 11kHz on a Macintosh Quadra 800 microcomputer, using SoundDesigner II software and a 12-bit Digidesign audio-media board. They were analyzed using Signalyze 3.0.1 software implemented on the same computer. Segment duration measurements were made from synchronous waveform and spectrographic displays of each sentence, expanded to provide a minimum resolution of 1.5 ms.

Because it was not always possible to identify the onset of the initial [s] frication, the onset of [i]-formant structure in 'seek' was used as the beginning of each token. The acoustical point used as the end of the token depended on the final word: the end of [æ] in 'pat' for the two sentences in which the target word was in final position, and the end of [ɪ] in 'music' or the end of [l] in 'musical' for the sentences in which the target word was in medial position. We also measured the durations of [i] in 'seek,' and of [æ] of 'pat.'

RESULTS

As expected, the 'slow' utterances have the longest, and the fast ones have the shortest, durations (Figure 1). In addition, there are greater differences among subjects in the 'slow' than in either the 'natural' or the 'fast' condition, and the difference between 'slow' and 'natural' speech is greater than that between 'natural' and 'fast' speech. (This is not surprising, since there are no necessary limits other than respiratory capacity on how long a sentence might last. However, the degree of shortening must be limited--segments have some shortest duration below which they cannot be articulated--and, possibly, perceived.)

Also as expected, a 3-factor ANOVA revealed that, across subjects, both speaking rate and sentence length had significant effects on sentence duration, but the position of the target word did not. There was also a significant interaction between rate and length.

Figure 2 presents the pooled mean durations of the vowel [æ], in the target word 'pat.' A 3-factor analysis of variance revealed that for each subject, *Rate* had a significant effect on the duration of the target vowel, that *Position* of the target word was significant for four of the subjects, and that the *Rate x Position* interaction between was significant for all five subjects (Table 2).

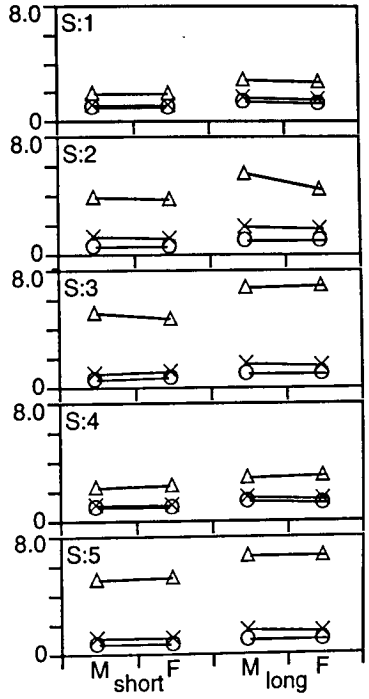


Figure 1. Average sentence duration in sec, for each subject. Circles represent fast-rate utterances; X's, natural-rate utterances; triangles, slow-rate utterances. 'Short' sentences are at the left, 'long' ones at the right. Medial target word sentences at the left member of each pair.

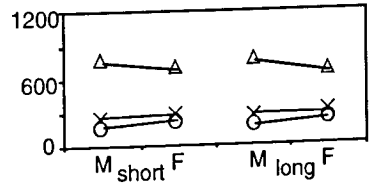


Figure 2. Five-subjects mean durations (in ms) of [æ]. Circles represent fast rate; X's, natural rate; triangles, slow rate. 'Short' sentences at the left; long sentences at the right. Within each pair, the medial-position target word is to the left, the final-position target word is to the right.

To examine further the relations among [æ] duration, speaking rate, and target word position, we calculated the ratio of [æ] duration to that of the sentence in which it occurred. In Figure 3, we see that the medial-position ratios (depicted by circles) are more similar across speaking rates than final-position ratios. Of particular interest here is that although the ratio is generally greater for final- than medial-position targets at the fast rate, the ratio decreases as speaking rate becomes slower. Thus, as sentence duration increases as a function of speaking rate, final syllables occupy a smaller proportion of the total utterance duration, compared with the same syllables produced at natural and fast speaking rates.

Table 2. Results of 3-way repeated measures ANOVA on the effects of Rate and Position and the Rate x Postion interaction on the duration of the target-word vowel.

Variable: Rate

Subject	*F*	(df)	*P*
1	284.16	(2,143)	.0001
2	115.62	(2,143)	.0001
3	188.29	(2,143)	.0001
4	115.67	(2,143)	.0001
5	1118.99	(2,143)	.0001

Variable: Position

Subject	*F*	(df)	*P*
1	242.38	(1,143)	.0001
2	4.48	(1,143)	.0400
3	34.03	(1,143)	.0001
4	11.13	(1,143)	.0103
5	1.50	(1,143)	.2456

Variable: Rate x Position

Subject	*F*	(df)	*P*
1	18.66	(2,143)	.0001
2	4.14	(2,143)	.0192
3	73.89	(2,143)	.0001
4	26.28	(2,143)	.0001
5	4.87	(2,143)	.0177

The data shown in Figure 4 are analogous to those of Figure 3, but are ratios of the duration of [i], the first vowel in the sentence, to overall sentence duration. We had no reason to expect that the position of the target word would affect the duration of [i], and so should not be surprised that the ratio patterns are the same for the two sentence types. That is, that the medial-position and final-position functions are superimposed at all three speech rates for all five subjects, showing that utterance-initial syllables occupy a relatively larger proportion of the entire utterance duration as speaking rate becomes slower. What is especially interesting, we think, is how clearly these [i]-ratio data, taken together with those for [æ], show that as sentence duration increases with speech rate (from left-to-right in each group), the early parts of sentences show the greater growth in duration; that is, they occupy a larger proportion of the sentence.

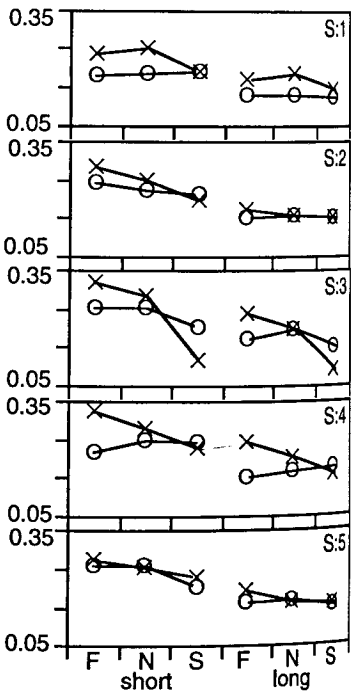


Figure 3. Ratios of [æ]-to-sentence duration for all five speakers. Circles represent medial-position target-word ratios, X's represent final-position target-word ratios. F=fast rate, N=natural rate; S=slow rate.

In summary, then, our data suggest, first, that slowing down speech results in longer relative durations of vowels in utterance-initial syllables (e.g., the [i] in seek). Second, vowels in sentence-medial position tend to occupy the same proportion of a sentence across changes

in speaking rate. Third, vowels in sentence-final syllables are not lengthened further at slow rates, and thus, have shorter relative durations at slow speaking rates.

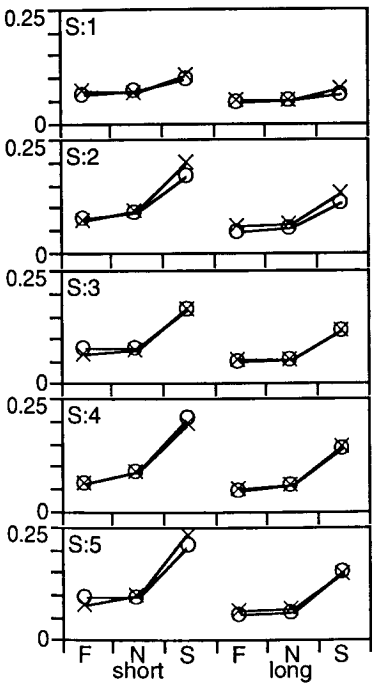


Figure 4. Ratios of [i]-to-sentence duration averaged for all five speakers. Circles represent medial-position target-word sentences, X's ratios In final-position target-word sentences. F=fast rate, N=natural rate; S=slow rate.

The simplest explanation for this last result, that our speakers ran out of air, can be rejected because if that were true, [æ] should have been shorter in long than in short utterances, and this was not so.

Another possibility is that, under normal circumstances, the speech system's natural rhythm reflects the sort of 'winding down,' or general declination, across the components of the breath group [6,7]. However, there may be limits to how slowly one may speak and still be operating under the normal regime.

ACKNOWLEDGMENTS

This research was supported by NIH grant DC-00121 to the Haskins Laboratories, and by St. John's.University, William Paterson College, and Burke Rehabilitation Hospital.

REFERENCES

[1] Lindblom, B., Lyberg, B., & Holmgren, K. (1981). *Durational patterns of Swedish phonology: Do they reflect short-term memory processes?* Bloomington, IN: IULC.

[2] Lindblom, B., & Rapp, K. (1973). Some temporal regularities of spoken Swedish. *Papers from the Institute of Linguistics (University of Stockholm), 21,* 1-59.

[3] 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.

[4] Bell-Berti, F., & Chevrie-Muller, C. (1991). Motor levels of speech timing: Evidence from studies of ataxia. In H. F. M. Peters & C. W. Starkweather (Eds.), *Speech Motor Control and Stuttering, pp. 293-301.* Amsterdam: Elsevier.

[5] Bell-Berti, F., Gelfer, C. E., Boyle, M., & Chevrie-Muller, C. (1991). Speech timing in ataxic dysarthria. *Proceedings of the XIIth International Congress on Phonetic Sciences, 5,* 262-265.

[6] Fowler, C. A. (1988). Periodic dwindling of acoustic and articulatory variables in speech production. *Perceiving-Acting Workshop, 3,* 10-13.

[7] Krakow, R. A., Bell-Berti, F., & Wang, Q. E. (1995). Supralaryngeal declination: Evidence from the velum. In F. Bell-Berti & L. J. Raphael (Eds.), *Producing speech: Contemporary issues for Katherine Safford Harris (pp. 333-354). Woodbury, NY: American Institute of Physics.*

THE MAINTAINING OF DURATIONAL RATIOS IN QUANTITY DISTINCTIONS IN CONVERSATIONAL SPEECH

Zita McRobbie-Utasi
Department of Linguistics, Simon Fraser University, Burnaby, Canada.

ABSTRACT

It is a well known fact that segments may occur in a great variety of different phonetic realizations in conversational speech. The question addressed here is to what extent the durational component of linguistic quantity in conversational speech will differ from that associated with citation forms. Durational ratios play a significant role in the realization of different quantity degrees in Skolt Sámi (a Finno-Ugric language). This language makes lexical as well as grammatical use of the quantity feature. The objective of this study is to examine the maintaining of linguistically significant durational ratios in conversational speech. The results of three experiments designed to determine the role of duration in quantity realization point to (i) recognizing the significance of certain durational ratios associated with different morphological classes in Skolt Sámi, and (ii) there being a definite tendency to maintain these ratios in speaking modes radically different from those occurrences of speech where more articulatory precision is needed.

INTRODUCTION

Previous analyses of Sámi quantity have established that disyllabic units (stress-groups) are to be regarded as the domain of quantity [1]. It has generally been accepted that within the disyllabic unit there exists a definite durational interdependency between the first syllabic vowel, the consonant(s) following it, and the second syllabic vowel [3]. The results of the acoustic analysis reported on here derive from three experiments all directed toward examining durational interdependencies from the perspective of the maintaining of the characteristic durational ratios that exist between the first syllabic vowel and the consonant(s) following it. These ratios were discussed in [4, 5]. The maintaining of these ratios were examined (i) in disyllabics were compensatory lengthening occurred (first experiment); (ii) in larger grammatical units (second experiment); and (iii) in disyllabics occurring in spontaneous conversations in which the same speakers participated as in the first two experiments (third experiment). On the basis of the result of the first experiment it was concluded that durational ratios rather than absolute durational values are relevant in the signalling of different quantity degrees in Skolt Sámi. The second experiment showed that the tendency to maintain the durational ratios overrides the apparent tendency to keep the duration of larger grammatical units (such as the paragraph) constant. The third experiment implies that there is a definite tendency to keep the durational ratios unchanged -- thus phonetic variations that occur in conversational speech will not affect durational ratios relevant in the realization of distinctive quantity degrees.

EXPERIMENTS

In all three experiments described below the recording was made with a Scully Full-Track Broadcast Machine tape recorder in a sound-proof room. The recording speed was 7.5" per second. The software for making durational measurements was the Signalyze (Version 3.12) with a Macintosh computer.

Durational ratios in compensatory lengthening

In the first experiment recordings of 1,200 disyllabics belonging to the five types of disyllabics [3,4] were made by two speakers. They were asked to place the test-words in the sentence frames *cie'lk e'pet* and *saar ... epet* 'say ... again' respectively.

Skolt Sámi has a phonological rule that either reduces or drops word-final vowels (the latter being more common in

connected speech). There are five structural types of disyllabics where this rule may apply [3]: Type 1 (containing a long geminate), Type 2 (containing a long consonant cluster), Type 3 (containing a single consonant), Type 4 (containing a short geminate), and Type 5 (containing a short consonant cluster). Type 1 has two sub-groups: 1a (containing liquids, nasals or non-sibilant fricatives), and 1b (containing plosives, affricates or sibilant fricatives). Similarly, Type 4 has two sub-groups: 4a (containing voiced fricatives); 4b (containing a plosive, affricate or voiceless fricative). Durational mesurements were made of the first syllabic vowel, the consonant(s) following and the second syllabic vowel (when present). *Table 1* summarizes the results of these measurements when there is a full vowel in the second syllable; *Table 2* summarizes the results of these measurements when there is a reduced vowel or no vowel at all word-finally. Mean durations ($\bar{x}$) and standard deviations (SD) are given for each segment in each structural type (durational values are given in milliseconds). These tables also show the V/C ratios.

Table 1. Durational measurements of disyllabics with a full vowel in the second syllable

Type	V		C		V/C
	$\bar{x}$	SD	$\bar{x}$	SD	
1a	209	13	171	20	1.22
1b	188	16	222	16	0.84
2	147	14	322	16	0.45
3	273	20	85	8	3.21
4a	206	15	146	10	1.41
4b	206	15	207	17	0.99
5	229	21	163	17	1.30

The results of the durational measurements, as presented in *Table 1* and 2, show the realization of the compensatory lengthening process. It can clearly be seen that all these structural types behave similarly in terms of durational increase as a result of compensatory lengthening: i.e. the duration of the vowel that has become reduced or deleted in the second syllable is added to the duration of both of the preceding segments. The durational increases in the relevant segments average 32 msec for vowels and 38 msec for the first syllabic consonant(s).

Table 2 Durational measurements of disyllabics with a reduced vowel or no vowel word-finally

Type	V		C		V/C
	$\bar{x}$	SD	$\bar{x}$	SD	
1a	240	22	207	18	1.15
1b	220	19	265	21	0.83
2	170	13	373	21	0.45
3	349	20	89	7	3.80
4a	233	15	179	16	1.30
4b	233	15	253	20	0.92
5	269	18	198	16	1.35

The pattern of durational increase, observable in the five structural types where compensatory lengthening is present, suggests two important trends: (i) the durational ratios remain constant, and (ii) the durational increase occurs in both the consonant(s) and the vowel segment preceding the second syllabic vowel. The different behavior of disyllabics belonging to the third type have been discussed in detail in [3,4] and its implications are not relevant in this context.

The above measurements thus indicate the maintaining of durational ratios even though absolute durational values change due to compensatory lengthening.

The maintaining of durational ratios in larger grammatical units

This second experiment consisted of the recording of six paragraphs by the same two speakers, each paragraph being recorded twice; thus the total number of recorded paragraphs was 24. Each of the six paragraphs under investigation contains three sentences. The sentences in these paragraphs were the same, except for their ordering. More details of this experiment were discussed in [5]. Suffice it here to say that the clearly recognizable timing strategies by the speakers that aim at a certain durational target are associated with shorter word duration in sentences in the third (i.e. last position). Consequently, segment durations in these words are also significantly decreased. Durational changes manifested in these segments were examined in relation to the constant

durational ratios between the first syllabic vowel and the consonant(s) following it as discussed above in connection with the compensatory lengthening phenomenon. Thus, it was to be expected that shorter word duration associated with paragraph-final sentence position will correspond with shorter segment duration. The question I tried to answer was whether or not absolute durational change -- in this case, decrease in duration -- affects the ratios of first syllabic segment duration. In particular, the issue addressed in this experiment was whether there is a tendency to keep durational ratios between the relevant segments constant.

The result of this present experiment indicates that changes in absolute duration indeed do not affect durational ratio values. *Table 3* shows mean durations of the first syllabic segments of disyllabics together with their durational ratios; *Figure 1* summarizes the changes in absolute duration in relation to constant durational ratios. The paragraph-final words that contain the segments analyzed here belong to the first structural type, Type 1a and 1b (see above).

Table 3. Mean duration and durational ratios of segments in disyllabics Types 1a and 1b of paragraph-final words

Type	V		C		V/C
	x̄	SD	x̄	SD	
1a	174	21	139	20	1.25
1b	155	19	164	18	0.94

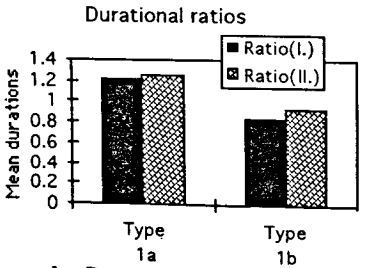


Figure 1. Durational ratios of shorter segment duration in relation to ratios of average segment duration

A comparison of the durational values in the above table with those in *Table 1* confirms the importance of durational ratios. That these segmental durational changes occur in paragraph-final positions points to the relevance of a temporal organization strategy in keeping the targeted duration of the paragraph constant. These absolute durational changes do not affect durational ratio values, which are in accordance with the same tendency observed in connection with the compensatory lengthening phenomenon referred to above. *Figure 1* illustrates the durational patterns summarized in *Table 3*.

The maintaining of durational ratios in spontaneous conversation

In the third experiment audio recordings were made of spontaneous conversation in which the same two speakers participated. 110 utterances were analyzed, 55 for each speaker. The analysis here focussed on the durational properties of disyllabic stress-groups, and, similarly to the above two controlled experiments, the first syllabic vowel and the consonant(s) following it were analyzed. According to the varying speech tempo, the relevant segment durations displayed a rather wide variety. It proved to be practical to divide them into several groups within each structural type on the basis of of the apparent durational properties associated with the segments under investigation in relation to the speech tempo variations. In this place only those disyllabics are analyzed which showed significant durational difference from those of the citation form presented in *Tables 1* and 2. While disyllabics in citation forms averaged word durations between 550 and 720 msec (depending on structural types), in this fastest speech tempo that occurred in the conversation recorded word durations averaged between 177 and 302 msec. *Tables 4* and 5 summarize the durational measurements and the V/C ratios of disyllabics occurring in spontaneous speech with or without a second syllabic vowel. The durations of the second syllabic vowel have a mean average of 56 msec.

Tables 4 and 5 indicate a definite tendency to keep durational ratios constant. This can be stated despite (i) the evident large standard deviation

values associated with varying speech tempo, and (ii) the noticeable different ratio values when compared with those presented in *Table 1* and 2 representing citation forms in controlled experiments. It has to be noticed, however, that (i) these differences are largest in connection with the third type -- but even with this difference it clearly separates this type from the others, and (ii) ratio values in all the other types can clearly be related to those ratio values characteristic of the citation forms. *Figure 2* summarizes durational ratio values associated with spontaneous speech and those associated with the citation forms.

Table 4. Segment durations and V/C ratios in spontaneous speech with a full vowel in the second syllable

Type	V		C		V/C
	x̄	SD	x̄	SD	
1a	124	28	103	23	1.20
1b	96	28	127	32	0.75
2	92	25	160	31	0.57
3	139	35	61	18	2.27
4a	155	23	110	27	1.40
4b	139	24	157	34	0.78
5	149	34	102	27	1.46

Table 5. Segment durations and V/C ratios in spontaneous speech with no vowel in the second syllable

Type	V		C		V/C
	x̄	SD	x̄	SD	
1a	135	31	120	26	1.12
1b	110	29	138	34	0.79
2	104	25	172	31	0.60
3	151	35	83	21	1.81
4a	172	30	132	29	1.30
4b	153	27	173	35	0.88
5	157	31	121	34	1.29

Conclusion

Articulatory simplicity associated with spontaneous speech, though changing absolute durational values, will not affect the realization of significant durational ratios in Skolt Sámi, a language with contrastive quantity degrees. It can thus be concluded that this study supports the claim that languages with distinctive duration tend to maintain characteristic durational patterns more consistently in conversational speech, while in those languages where duration is not contrastive, characteristic durational values tend to be less stable in a more casual speaking mode [6].

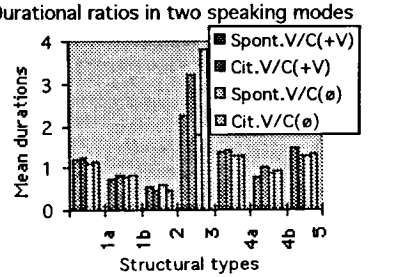


Figure 2. Durational ratios in spontaneous speech in relation to ratios of segment duration in citation forms

REFERENCES

[1] Itkonen, E. (1946), *Struktur und Entwicklung der ostlappischen Quantitätssysteme*, Mémoires de la Société Finno-Ougrienne 88. Helsinki: Suomalais-ugrilainen Seura.

[2] Magga, T. (1984), *Duration of quantity of disyllabics in the Guovdagaeidnu dialect of North Lappish*, Acta Universitatis Ouluensis, series B. Oulu: Oulu University Publications.

[3] McRobbie-Utasi, Z. (1991), *An acoustic analysis of duration in Skolt Sámi disyllabics*, Ph.D. Dissertation. University of Manitoba.

[4] McRobbie-Utasi, Z. (1991), "Durational ratios of disyllabics in relation to the syllable boundary in Skolt Sámi", in *Proceedings of the XIIth International Congress of Phonetic Sciences, Aix-en-Provence, France.* pp. 306-309.

[5] McRobbie-Utasi, Z. (1994), "Timing strategies within the paragraph", in *Proceedings of the 1994 International Conference on Spoken Language Processing, Yokohama, Japan.* pp. 382-386.

[6] Engstrand, O. & Krull, D. (1994), "Durational correlates of quantity in Swedish, Finnish and Estonian: Cross-language evidence for a theory of adaptive dispersion", *Phonetica*, vol. 51, pp. 80-91.

PROSODIC BOUNDARY STRENGTH IN SWEDISH: FINAL LENGTHENING AND SILENT INTERVAL DURATION

Merle Horne*, Eva Strangert** and Mattias Heldner**
*Dept. of Linguistics & Phonetics, Lund Univ. **Dept. of Phonetics, Umeå Univ.

ABSTRACT

Production data are presented that support the assumption of Final Lengthening and Silent Interval duration as parameters of prosodic boundary strength in Swedish.

INTRODUCTION

Within the hierarchical model of prosodic constituents assumed for Swedish [1], four boundary strengths (0-3) have been assumed: the 0-boundary is thus associated with the end of a word *within* a Prosodic Word (PW), boundary strength 1 with the end of a PW, boundary strength 2 with the end of a Prosodic Phrase (PPh), and boundary strength 3 with the end of a Prosodic Utterance (PU). Although we have a good idea as to how tonal parameters are associated with the various prosodic categories, it is not clear how other parameters, in particular Final Lengthening (FL) and Silent Interval (SI) duration are associated with the various prosodic constituents.

Prosodic constituents have often been assumed to be associated with specific degrees of FL. The domain of lengthening is generally believed to be the rhyme of the final syllable [2-3]. Recent studies have also shown that the lengthening is progressive, i.e. the final consonant is lengthened to a greater extent than the preceding vowel [4]. It is also known that FL is influenced by the presence of prominent accents/boundary tones [3]. Researchers on Swedish [5-6] have provided evidence from production experiments that shows that segment lengthening in final position is indeed influenced by the presence of a focal accent on the last word in an utterance. Lyberg & Ekholm, basing themselves on measurements made on the stressed vowel rather than the final rhyme consonant in their test words [7], do not find any consistent evidence for the appearance of FL as an independent marker of the end of a phrase. Fant and colleagues [8-9], in studies on read prose, show that there is a negative correlation between SI duration and FL in their analysis of stress foot structure in Swedish. These results prompted us to make an investigation in order to tease out the relation between accenting, FL and SI duration in order to relate the findings mentioned above to boundaries assumed in the prosodic constituent hierarchy.

DATABASE STUDY

We made a preliminary data base study to determine 1) whether boundaries between the four types of prosodic constituents we have assumed for Swedish are associated with different degrees of perceived boundary strength and if so, to determine 2) if and how these boundary strengths correlate with segment lengthening and silent intervals for the radio commentator style we are modelling.

17 broadcasts from Radio Sweden on Stock-Market rates were studied. The boundary strength after the Accent 1 word *procent* 'percent' [pru'sent] was chosen for analysis since this word was uttered on the average of 5 times during the broadcasts comprising the database. Also the 5 different occurrences of *procent* always had the same respective syntactic position in each text. The material was presented to 2 native listeners, who scored the strength of the perceived boundary after *procent* on a 4 point scale, where 0 corresponds to no boundary (0-boundary), and 3 corresponds to the strongest boundary (PU-boundary). An example of one of the texts follows (subscripts after the word *procent*: refer to the predicted boundary strength):

Vid 13-tiden noterades Stockholms fondbörs generalindex till 1026,1. Det är en uppgång med 0,1 procent(1) jämfört med gårdagens slutindex. 16-i-topp-index hade då gått upp med 0,4 procent(3).

Marknadsräntorna vid middagstid: den 4-åriga standardobligationen låg då stilla på gårdagens slutränta på 10,12 procent(2), 12-månaders statsskuldväxlar hade gått tillbaka 1 räntepunkt till 10,58 procent(2), medan sexmånadersväxlar gått upp 5 punkter till 10,50 procent(3).

Preliminary results

The boundaries were given scores of 1, 2, or 3 in accordance with the predictions. (The text contained no test word followed by a predicted 0-boundary and there there were no 0-scores either.) A MANOVA analysis demonstrated significant differences in the rhyme of the stressed syllable for the scored boundaries. The greatest differences were found in the [t] segment which differed significantly between all three perceived boundary strengths and was furthermore positively correlated with the strength of the boundary. The SI durations were also positively correlated with the strength of the boundary, and significant differences were found between all three boundary strengths.

Although the results indicated a clear correlation between type of boundary and the phonetic correlates, there were certain gaps in the database: the test word was non-focal in 4 cases and focal in one case. Furthermore, as there were no examples of the test word followed by a 0-boundary, we decided to conduct a more structured lab study in order to include all boundary types.

LAB STUDY

The study was thus undertaken in order to include all boundaries after both [+focus] and [-focus] words to determine to what extent 1) the boundary-type and 2) the focal/non-focal status of a word interacts with FL and SI duration.

One text from the database study was modified so that in new versions *procent* was followed by all 4 types of boundary, 0-boundary, PW, PPh and PU-boundaries, respectively. Moreover, 2 subcategories of boundary within the PPh and PU were distinguished: clause-final/sentence-final position for PPh and paragraph-final/textfinal position for PU. Thus there were 6 boundary categories: 0, PW, PPh/C, PPh/S, PU/P and PU/T. All 6 categories occurred after both [+focus] and [-focus] *percent*. In this way 6 new texts were created which were read 10 times by the same speaker as in the database study. Altogether the material contained 120 occurrences of the testword (10 readings x 6 boundaries x +/- focus).

Results

Figure 1 shows the SI duration for the 5 types of boundary, and Figure 2, the duration of the entire word *procent* before each boundary. The data are given separately for +/-focus (abbreviated +/-fa), thus giving an overview of the general trends. Concerning SI duration, there is a gradual increase as the rank of the boundary becomes higher. The SI varies between 0 (0-boundary) to more than 900 msec (PU/T). However, only three levels of boundary differ significantly from each other on the basis of the measured intervals: 0 and PW versus PPh/C and PPh/S versus PU/P ($p<.05$). The focus distinction, moreover, has no significant effect on the SI duration. The duration of *procent*, on the other hand, differs considerably between the +/-focus condition ($p<.001$). The [+focus] increase in duration varies over the different boundaries between 40 and 100 msec. Figure 2 also demonstrates complex effects of boundary type on word duration, an increase in the higher-rank end of the curve (PPh/C - PU/T and a decrease in the lower-rank end (0 - PPh/C). Data disentangling the complex information in Figure 2 are presented in Figure 3 showing in separate parts the duration of the segments [s], [ɛ], [n], [t].

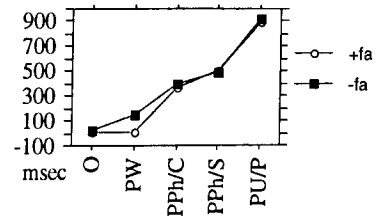


Figure 1. Silent interval duration following the test word with +/- focus accent.

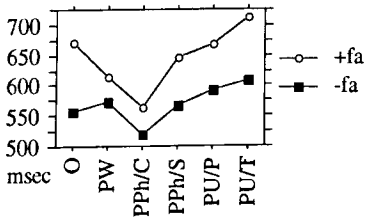


Figure 2. Duration of the test word with +/- focus accent before each boundary.

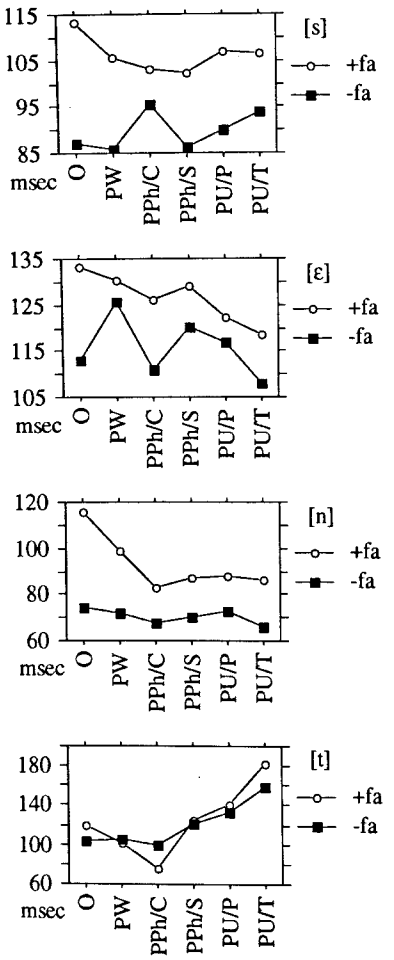


Figure 3. Stressed final syllable segment durations in +/- focus accented test word before each boundary.

First, whether the test word is focussed or not has significant effects on all segments except the final [t] ($p<.001$). (The first syllable of the test word is affected similarly.) Secondly, the final [t] as well as the preceding [ε] and [n], are significantly ($p<.001$) affected by boundary type (as is also the initial [pr]-segment). However, the three segments are affected in very different ways. For [ε] there seems to be a negative correlation between segment duration and the rank of the boundary. For [n] there is also a negative correlation, but only for the boundaries ranked lowest, 0, PW and PPh/C. The higher ranked boundaries appear to be unaffected. For both [ε] and [n] these adjustments primarily affect [+focus] words. For [t] on the other hand, there is a positive correlation between SI duration and the boundaries with higher ranking, PPh/C, PPh/S, PU/P and PU/T. The interaction between boundary type and +/-focus makes it difficult to state the effects of boundary type for 0, PW and PPh/C. Thus, the segment duration data demonstrate where the effects found in Figure 2 come from. The increase of duration in the higher rank end of the curve stems primarily from the [t], while the decrease in the lower rank end is a combined effect of adjustments made in [ε], [n] and, to some extent, [t].

CONCLUSIONS

We may first conclude that the database study indicates that the boundaries associated with the four environments we have analysed can be perceptually distinguished. The subcategorizations of the PPh and PU categories in the present study have not been tested perceptually yet. Thus, we cannot be certain as to how many distinguishable categories there are.

A major adjustment affecting the rhyme segment durations is associated with the focus accent, with [+focus] associated with significantly longer durations than [-focus). This is to be expected, as focus accent has temporal correlates in additon to the primary F0-correlates [6]. However, [t] as well as the SI following the test word do not conform to the general trend, both being unaffected by the +/-focus distinction.

Concerning the temporal adjustments associated with the boundary types investigated we found a more or less gradual increase in the duration of the SI upon an increase in the rank of the boundary observed (see also [8,10] for Swedish). Concerning segment durations, the increase in [t] duration associated with the higher rank end of the boundary scale and the decrease in [ε], [n] and [t] duration in the lower end, together, as we have seen, sum up to a v-shaped curve with PPh/C forming the lowest point.

The increase in [t] duration between the higher ranks in the boundary hierarchy (starting with the PPh/C) is evidence for FL existing independent of focus accent in Swedish (cf. also results by Edwards & Beckman [11] who claim that FL does not exist below the PPh). In contrast, Lyberg and Ekholm [7] measuring only the stressed rhyme vowel, could not find any evidence of FL as an independent marker of the end of a phrase. (Neither could we find any independent lengthening when measuring the stressed [ε] in the present study (see also [4]). However, the observations on Swedish made by Lyberg and Ekholm may be given an alternative interpretation in the light of the present study. The claim they made that FL is a consequence of focus position on the last content word - they, like us, observed lengthening of the [+focus] stressed *vowel* - may therefore be a consequence of the segment they chose for analysis.

We also have reported a decrease of duration in the rhyme segments at the lower ranks in the boundary hierarchy. In our test word with a stressed syllable containing a short vowel [ε] followed by two consonants [n] and [t], it is particularly the consonants which are affected, though only in the [+focal] condition. Duration is greatest at the 0 boundary, less at the end of a PW and least at the end of the PPh/C word. Combined with the silent interval data we present, these results corroborate previous observations of a trading relation between FL and SI duration [8-9]. In contrast to the findings of the other researchers, however, negative correlations are not obtained for our data above the PPh/C level.

The trading effect may be looked upon as a means to optimize boundary signalling, maximizing segment duration cues when SI duration is at its minimum. However, why is this pattern more or less restricted to the [+focus] conditions? And how does one explain the fact that the significant +/- focus differences we have reported are much more pronounced, especially in the [n], at the lower ranked boundaries; the difference is greatest at the lowest ranked boundary, 0, and least at the PPh/C boundary. These questions need to be answered in future research.

In summary, what one can conclude from this study is that the phenomenon of Final Lengthening does exist in Swedish. Its domain would appear to be PPh and PU. It affects the final segment of the rhyme. Silent Intervals, moreover, are intimately tied to the higher-ranked boundaries, PPh and PU. Further, there appears to be a trading relation such that at the lower-ranked boundaries, segment and Silent Interval duration are negatively correlated.

ACKNOWLEDGEMENTS

We would like to thank Suzanne Haage-Palm for reading the material, Antonio de Serpa-Leitao for assistance with the recordings and Björn Granström for generously placing the sound studio at KTH at our disposal. This research was supported by grants from the Swedish HSFR/NUTEK Language Technology Programme.

REFERENCES

[1] Horne, M. & Filipsson, M. (1995), "Computational modelling and generation of prosodic structure in Swedish", *Proceedings, XIII ICPhS 95*, Stockholm.

[2] Crystal, T.H. & House, A.S. (1990), "Articulation rate and the duration of syllables and stress groups in connected speech", *J. Acoust. Soc. Am.*, vol. 88, pp. 101-112.

[3] C.W. Wightman, S. Shattuck-Hufnagel, M. Ostendorf & P. Price (1992), "Segmental durations in the vicinity of prosodic phrase boundaries", *J. Acoust. Soc.Am.*, vol. 91, pp. 1707-1717.

[4] Berkovits, R. (1994), "Durational effects in final lengthening, gapping, and contrastive stress",. *Language and Speech* , vol 37, pp. 237-250.

[5] Lyberg, B. (1981), *Temporal properties of spoken Swedish*. MILUS, vol. 6, Stockholm U.

[6] Bruce, G. (1981), "Tonal and temporal interplay", *Working Papers* (Dept. of Linguistics, U. of Lund), vol. 21, pp. 49-60.

[7] Lyberg, B. & Ekholm, B. (1994), "The final lengthening phenomenon in Swedish – a consequence of default sentence accent?" *Proceedings of ICSLP* 94, 135-138.

[8] Fant, G. & Kruckenberg, A. (1989),. Preliminaries to the study of Swedish prose reading and reading style. *STL-QPSR* 2.

[9] Fant, G., Kruckenberg, A., & Nord, L. (1991),. "Prosodic and segmental speaker variations", *Speech Communication* , vol. 10, pp. 521-531.

[10] Strangert, E., (1990), "Perceived pauses, silent intervals and syntactic boundaries", PHONUM (Dept Phon., Umeå Univ.), vol. 1, pp. 35-38.

[11] Edwards, J. and Beckman, M. (1988), Articulatory timing and prosodic interpretation of syllabic duration. *Phonetica* 45, pp. 156-174.

The Intonational Disambiguation of Potentionally Ambiguous Utterances in English, Italian, and Spanish

Cinzia Avesani[*], *Julia Hirschberg, and Pilar Prieto*
AT&T Bell Laboratories

Abstract

We investigated the role that intonation plays in disambiguating potentially ambiguous utterances in English, Italian, and Spanish, to see a) whether speakers employ intonational means to disambiguate these utterances, and b) whether speakers of the three languages employed consistently different intonational strategies in this disambiguation. In a preliminary production study, speakers of the three languages did differentiate among some types of syntactic and scopal ambiguity intonationally. Their strategies differed among languages, with Spanish and Italian patterning together more often than either patterned with English.

INTRODUCTION

It is often been claimed that phenomena such as the scope of negation and quantifiers and the attachment of prepositional phrases and relative clauses can be disambiguated intonationally (Ladd, 1980; Bolinger, 1989). In this preliminary study, we investigated the strategies native speakers of English, Italian, and Spanish might use to disambiguate structurally identical utterances.

METHOD

We conducted a production study to identify intonational variations associated with different readings of potentially ambiguous utterances embedded in disambiguating contexts. We focused on the following types of ambiguity: 1) scope of negation; 2) quantifiers; 3) PP attachment. An Italian example of (1) is: *Non sono scappato da casa perché mia madre mi faceva paura*; an English example of (2) is: *None of the students would embarrass them*; a Spanish example of (3) is: *Ganó a la mujer con los dados*. Each sentence has two possible interpretations, a wide and a narrow scope reading for the negation, wide vs. narrow scope for the quantifier, and VP vs. NP attachment for the prepositional phrase. We constructed potentially ambiguous utterances in Italian, embedding each in two disambiguating contexts, and then translated the resulting paragraphs into English and Spanish.[1] We intended that subjects be able to infer each of the two interpretations of the sentences from the surrounding context. For example, a wide scope interpretation of negation for a sentence like *William does not drink because is unhappy* was conveyed by embedding it in the following paragraph:[2]

> I know William very well. Since his girlfriend left him, he's done nothing but drink. Now, such a long time since his separation, he's used to living alone. Now, *William doesn't drink because he's unhappy.* He drinks because he's an alcoholic.

A narrow scope was induced by embedding it in the following context:

> There's something about William that puzzles me. When he's happy, he has a good time with his friends, and certainly he does't dislike drinking. I think I understand what's wrong. *William doesn't drink because he's unhappy.*

We recorded four native speakers of each language (3 males and one female per language) reading these paragraphs. Two Italian speakers (GR, CA) are speakers of northern Italian, one (RP) of Tuscan, and one (RS) of a southern variety. Among them, only one (RS) can be said to have a strong regional (southern) prosodic characterization. Of the Spanish speakers, one is from the Ecuadorian Andes (JG) and the three others are Catalan, speaking Castilian for this experiment; of these, one is from Murcia (JP), and two from Barcelona. The English speakers are all American, from New Jersey (AB), Missouri (JH), and California (MK, GW).[3] Recording was done in a sound-proof room, results were analysed using Entropic Research Laboratory's Waves+ speech analysis software, and speech was transcribed using the ToBI. annotation conventions (Pitrelli, Beckman, and Hirschberg, 1994).

[*] and the University of Ferrara

[1] Our corpus is unbalanced: we have three pairs of utterances for scope of negation, two for quantifiers, and one for PP attachment.

[2] See the appendix for examples of additional sentence types, embedded in disambiguating contexts.

ANALYSIS AND RESULTS

English

For our English speakers, wide vs. narrow scope of negation in sentences like '*William doesn't drink because he's unhappy*' was distinguished in two ways, with speakers following one or both strategies in all cases. In the majority of cases, speakers placed an intermediate or intonational phrase boundary between the material within the narrow scoped negative (i.e., after *drink*, meaning "William doesn't drink") and uttering the wide scope version of the sentence (meaning "William does drink, but not because he's unhappy") as a single intermediate phrase. Also, in about 90% of cases, speakers employed falling intonation (a LL% ending) for narrow scope utterances but a continuation rise (LH%) for readings where the interpretation was wide scope.[4]

Quantifier scope shows no such pattern: While two speakers (AB, JH) distinguished wide from narrow scope for the negative quantifier *none* in sentences like '*The presence of none of the students would embarrass them*' by accenting the focus associated with the quantifier (i.e., *student*) in the wide scope case and deaccenting it in the narrow, the other speakers produced different patterns. And for ambiguous association of focus with *only*, no common intonational variations among any of the speakers distinguished between readings.

Ambiguous prepositional phrase attachment in sentences like '*He won the woman with the die*' was distinguished by three speakers (JH, MK, GW) by the presence of an intonational phrase boundary setting off the PP from the direct object to indicate VP attachment, compared to the presence of an intonational boundary between the verb and direct object or the absence of any internal prosodic boundary for the NP-attached reading. That is, a boundary was placed after *woman* to indicate VP attachment; for NP attachment readings, either a boundary was placed after *won* or the sentence was uttered as a single phrase. The fourth speaker (AB) produced no prosodic differences between the two readings.

So, our production studies suggest that speakers of American English may disambiguate scope of negation by varying prosodic phrasing and/or utterance-final tones (final fall vs. continuation rise). PP attachment ambiguities are also distinguished by three of our speakers by differences in prosodic phrasing. However, productions of quantifier scope ambiguous sentences (containing *none* and *only*) exhibit no such clear generalities, although two speakers did use accent placement to distinguish ambiguities involving the scope of *only*.

Italian

All our Italian speakers were quite consistent in the way they disambiguated ambiguous scope of negation. All instances of wide scope utterances were uttered as single intonational phrases; all the narrow scope utterances were uttered as two intermediate phrases, with a phrase boundary delimiting the scope of negation. So, for example, in *Guglielmo non beve perché é infelice*, speakers placed an intonational phrase boundary after *beve* in tokens with narrow scope readings, and no internal boundaries for those uttered in wide scope contexts. In uttering the wide scope utterances all speakers associated a prominent nuclear pitch accent with the negative verb, deaccenting the remainder of the utterance. In the narrow scope utterances, uttered as two phrases, one nuclear pitch accent was associated with the verb and one nuclear pitch accent was associated with *infelice*. As a combined effect of phrasing and accent placement, the lexical material in the subordinate clause was deaccented in the wide scope utterances, accented in the narrow ones.

Speakers were also consistent in the way they intonationally disambiguated the scope of quantifiers. In sentences like '*La presenza di nessuno studente potrebbe metterle in imbarazzo*', the strategy for disambiguating the scope of the negative quantifier *nessuno* for all speakers was: for narrow scope ("there will be no student who can embarrass them"), all speakers produced an utterance with one intonational phrase, placing the nuclear pitch accent on the quantifier itself and deaccenting the subsequent lexical material. For wide scope ("if no students come, they will be embarrassed") two speakers (GR, CA) produced

[3] The three authors participated as speakers.

[4] In one pair of paragraphs an orthographic difference may induced this distinction; however, even excluding tokens from this pair, only one pair of productions fails to exhibit this distinction.

utterances with a single intonational phrase, placing nuclear stress on the last content word of the utterance ("imbarazzo"); two others (RP, RS) produced utterances with two intermediate phrases, separated by a high intermediate phrase accent. Note that all speakers appeared to use same phrasing and same intonational contour for disambiguating the narrow scope of the negative quantifier and the wide scope of negation in type (1) sentences.

A different strategy was used for disambiguating the quantifier *solo*, in sentences like '*E' necessario che venga solo Maria*'. Accent placement and relative prominence appear to be the relevant means employed to disambiguate here, but speakers were inconsistent in their productions. One (RP) used pitch accent placement as a main prosodic cue, accenting the quantifier and deaccenting the noun (*Maria*) in the narrow scope utterances, while deaccenting the quantifier and accenting the noun in the wide scope ones. CA and GR accented both quantifier and noun in both cases, but assigned greater prominence to the quantifier than to the noun in the narrow scope contexts.

Intonational phrasing seemed to be the most important cue in disambiguating VP from NP attachment for prepositional phrases in sentences like '*Vinse la donna con i dadi*'. All speakers distinguished VP attachment by producing two intermediate phrases, with the phrase boundary occurring after the direct object (*la donna*). NP attachment differed among subjects: For three speakers (RP, CA, GR), the sentence was uttered as one intonational phrase (RP, CA, GR); for the fourth (RS), the sentence was uttered as two intermediate phrases, but the boundary occurred after the verb *vinse*; so, this speaker delimited the domain of attachment using phrasing in each case.

Summarizing, it appears that intonational phrasing was the only means used consistently by our Italian speakers to disambiguate the scope of the negative quantifier and to disambiguate ambiguous PP attachment. In type (1) utterances, intonational phrasing and nuclear accent placement were used by all speakers to disambiguate. Accent placement and prominence were the means through which our speakers disambiguated the scope of the quantifier *solo*. When speakers differ in their production of one member of the pairs, speakers of the northern Italian generally pattern together, as do speakers of Tuscan and southern Italian. In only one case (NP attachment) did northern and Tuscan speakers exhibit similar behavior among themselves, differing from the southern Italian speaker.

Spanish

Spanish-speaking subjects used phrasing to disambiguate ambiguous scope of negation in utterances like '*Guillermo no bebe porque está triste*. All four speakers produced wide scope utterances as single intermediate phrases and narrow as two intermediate phrases, with a high phrase accent at the end of the first phrase. For wide scope utterances, speakers deaccented *triste*, while accenting it in narrow scope utterances.

Quantifier scope disambiguation in sentences like '*La presencia de ningún estudiante podría ponerlas nerviosas*' was disambiguated through phrasing variation. Our Spanish speakers produced wide scope utterances as two intermediate phrases, and narrow scope utterances as a single intermediate phrase. However, the scope of the quantifier *sólo* was disambiguated by three speakers (PP, JG, JP) though pitch accent assignment. Wide scope utterances were produced with a deaccented *sólo* or a low accent (**L***), and the narrow scope reading was uttered with a peak (**H*** accent) on the quantifier.

Spanish subjects were inconsistent in the disambiguation of PP attachment. While speakers JG and PP did not distinguish between the two readings, JS and JP disambiguated the sentences through variation in phrasing. NP attachment was indicated by producing utterances as single intonational phrases, and VP attachment by producing two intermediate or intonational phrases.

So, our Spanish speakers consistently disambiguated scope of negation by varying prosodic phrasing and by varying accent placement. They disambiguated negative quantifer scope by varying phrasing alone, and the scope of *sólo* by varying accent placement and type. PP attachment was less consistently treated by these speakers.

DISCUSSION

We found that most of our speakers used intonational means to disambiguate the potentially ambiguous sentence types under investigation in this study. English, Spanish, and Italian speakers were most similar in their disambiguation of the scope of negation, employing variation in prosodic phrasing to distinguish wide from narrow scope productions, with wide scope utterances produced as a single phrase and narrow produced as two phrases. Italian and Spanish speakers also differentiated wide from narrow scope by similar variation in phrasing; however, they also placed nuclear stress on the verb to indicate wide scope negation, while English speakers located nuclear stress later in the utterance. Also, English speakers further distinguished wide from narrow scope by utterance-final tonal variation, with continuation rise employed for wide scope readings and falling intonation for narrow. While our Italian speakers consistently used phrasing variation to indicate differences in PP attachment (between NP and VP attachment), English and Spanish subjects were inconsistent in this regard. For quantifier disambiguation, the picture is more complex: For Italian and Spanish speakers, renditions of sentences containing scope-ambiguous negative quantifiers were disambiguated by variation in nuclear stress placement and in prosodic phrasing; for two English speakers, accent placement served to disambiguate these utterances. However, *only/solo/sólo* was treated less consistently by speakers of all three languages.

Inconsistencies among speakers of all three languages could be due to regional differences in the use of prosodic variation. Our limited evidence for different patterning of the Italian speakers according to language variety suggests that this may be an area worth exploring further. A partial analysis of the present data for differences in pitch accent prominence and duration also suggest that prosodic cues other than those discussed might also contribute to the disambiguation of ambiguous utterances. Collection of a larger corpus with more speakers for each language and more paragraphs for each ambiguity type should shed light on both these areas.

SAMPLE PARAGRAPHS

VP attached PP: I remember that scene in the officers club. There were four of them, and they were playing dice. One of them, the youngest, was in love with the commandant's wife. The commandant was older than she was, and had a wild passion for gambling. That night he lost all he had. The youngest player proposed the woman as a stake. The commandant accepted. They rolled the die. The young player won. *He won the woman with the die.*

NP attached PP: Paradiso worked in the carnival. In the next stand, there was a target-shooting game, where the prizes were old paintings. Paradiso's favorite one showed a woman throwing a pair of dice. Paradiso tried and tried to win this painting, but try as he would always failed. Finally, one night he decided that he no longer wanted the painting. And what do you suppose happened then? *He won the woman with the die.*

wide scope negative quantifier: Usually our university organizes at least one seminar per year. Every student and every researcher is supposed to attend that seminar. Next week, Maria will give a talk with Marina on quantifiers. *The presence of none of the students would embarrass them.*

narrow scope negative quantifier: Maria and Marina are close to getting their degrees. Tomorrow they will rehearse their thesis defenses. I've heard them already. They're really good. *The presence of none of the students would embarrass them.*

wide scope quantifier: Mary is organizing a party for next weekend in her parent's place. I think that she wants to invite a bunch of people I don't really care about. It's really not important to me whether they come or not. There's only one person I'm interested in. All of you know who it is. For me, *it is important that only Mary comes.*

narrow scope quantifier: I have a problem. Mario likes Mary but he is a little timid about asking her out. He's asked me if I could organize something so that the two of them can be alone. It needs to be something casual, and, naturally, with nobody else around. I've thought of organizing a party at home and inviting the two of them, as well as some other people. At the last minute I will explain to everyone but Mary and Mario that the party has to be postponed. I don't know what else I could do. *It is important that only Mary comes.*

REFERENCES

Bolinger, Dwight. 1989. *Intonation and Its Uses: Melody in Grammar and Discourse.* Edward Arnold, London.

Ladd, D. Robert. 1980. *The Structure of Intonational Meaning.* Indiana University Press, Bloomington, Ind.

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

PITCH VARIATIONS AND EMOTIONS IN SPEECH

Sylvie Mozziconacci
Institute for Perception Research / IPO, Eindhoven, The Netherlands

ABSTRACT

Speech of a Dutch male professional speaker enacting seven emotions was analysed with respect to pitch. Because observed pitch variations could not easily be captured in a two-component intonation model, the perceptual relevance of the differences observed between natural contours analyzed and contours as described by the model was tested. Results showed that modelling the departures from the model did not result in improved recognition performance.

INTRODUCTION

Utterances expressing different emotions show systematic differences with respect to pitch, temporal properties, and voice quality [1, 2, 3]. Recent work on speech recognition and speech synthesis has made clear the need for insight into the prosodic characteristics associated with different emotions. The work to be reported here is part of a study of these characteristics.

It is not known exactly how the relevant prosodic features should be controlled to get optimal recognition of different emotions in synthetic speech. Hence, acoustic analyses and perceptual evaluations of tentative rules are needed. This involves an adequate representation of the acoustic data.

The research reported here focuses exclusively on the role of pitch. In a previous study [1], tentative rules for synthesizing utterances conveying particular emotions were expressed in terms of a two-component model [4]. In this model, one component represents *pitch register* variations, concerning how high or low the utterance is produced in the speaker's overall range. Register is operationalized by means of a baseline which is anchored in the utterance-final low pitch and which has a certain slope. The other component represents *pitch range* variation, operationalized in terms of the distance between local F0 minima and maxima (i.e., the size of pitch changes). For sake of simplicity, the pitch range has been held constant throughout the utterance in formulating tentative rules, although this is not a necessary assumption within the model.

Comparing the output of the tentative rules to the contours in utterances conveying different emotions produced by a human speaker (these last contours will further be called natural contours), we observed that the rule-based contours differed in several respects from the natural contours, for most emotions. This observation gave rise to the questions to be addressed in the current study:

1. What are the detailed characteristics of pitch contours of utterances expressing different emotions?
2. To what extent does the modelling of these characteristics lead to improved recognition of the emotions?

I. ACOUSTIC ANALYSIS

Materials and method

A male Dutch speaker enacted seven emotions (neutrality, joy, boredom, sadness, anger, fear, and indignation), producing three tokens of each of five sentences for each emotion. The five sentences had previously been found to have neutral semantic content (e.g. *Het is bijna negen uur* "It is almost nine o'clock").

Intonation contours were labelled according to the description by 't Hart, Collier and Cohen [4].

Because natural contours differed substantially from synthetic contours produced by means of rules based on the outcome of preliminary research [1,5], an accurate description of the natural contours was needed. Therefore, pitch was measured at six "anchor points" in each utterance: onset, two peaks (all utterances contained two accented words), two values in the intermediate "valley" (after the first peak and before the second peak), and offset.

Results and discussion

The labelling of intonation contours revealed that the so-called 1&A pattern (pitch rise-and-fall on a single accented syllable) was used for all emotions. Indeed, for most emotions it was the most frequently used pattern. This suggests that the 1&A pattern is appropriate for the expression of all emotions under study. Therefore, in order to exclude variation due to phonological structure, further acoustic analyses will be restricted to utterances realised with the 1&A pattern.

For each emotion a "mean natural contour" was calculated by taking the means for the successive "anchor points". They are shown in figure 1. This figure shows that pitch register and range vary systematically as a function of emotion. Further inspection of figure 1 indicates that there are two major sources of deviation between the natural and rule-based contours:

1. Whereas in the rule-based contours all minima fall on a single declining baseline, the utterance-final low pitch in the natural contours does not simply fall in line with the minima in the earlier part of the utterance. Instead, the offset of pitch can be considerably lower than expected on the basis of the earlier part of the contour.
2. Whereas the size of pitch changes is the same throughout the utterance in rule-based contours, this is not the case in the natural contours. Instead, for contours higher in register, the second pitch accent becomes increasingly larger than the first one.

In fact the only emotions for which the natural contours compare well to the rule-based versions are neutrality and boredom. This is not surprising, since the model was developed on the basis of neutral, uninvolved utterances.

The question arises whether detailed modelling of these departures from the rule-based versions may improve identification accuracy for different emotions.

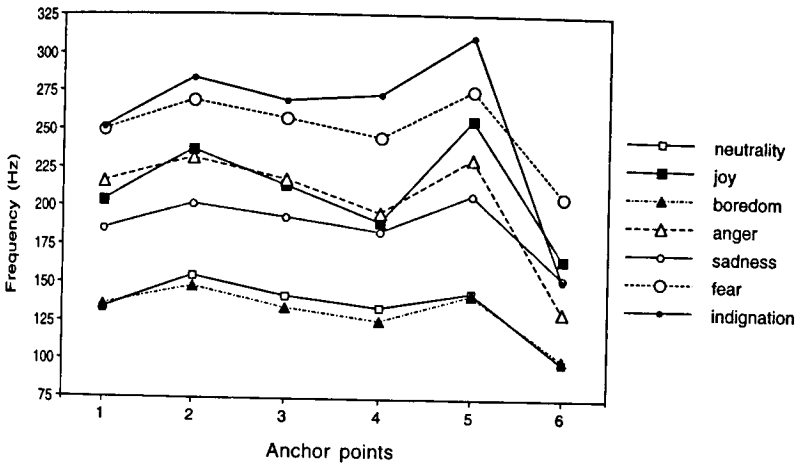


Figure 1. Results of F0 measurements averaged over successive points (so called anchor points) in 1&A realisations of five sentences. Each anchor point is based on at most 15 tokens for each of the seven emotions N: neutrality, J: joy, B: boredom, A: anger, S: sadness, F: fear, and I: indignation. Symbols representing successive anchors for each emotion are connected with lines.

II. PERCEPTUAL EVALUATION

The perceptual test investigates to what extent modelling the detailed aspects of contours for different emotions helps to improve their recognition.

Materials and method

Synthetic pitch contours were generated for a single carrier sentence (*Zijn vrienDIN kwam met het VLIEGtuig* "His girlfriend came by plane"), with accents on /din/ and /vlieg/. Both accents were realized with 1&A type pitch accents.

For each emotion, five different

synthetic pitch contours were produced, representing five conditions. Condition 1 was included for comparison and conditions 2 to 5 resulted in approximations of the pitch contours in figure 1. All synthetic contours had a fixed baseline declination of 3.5 semitones / s.
Sentence duration is 1.77 s.

1. *Two-component optimal perceptual values.*
Contours were generated using values for pitch register and range that had been obtained in a previous adjustment experiment aiming to determine optimal perception-based parameter values [1, 5]. The end frequency and the size of the pitch movements vary as reported in table 1. The slope of the baseline is fixed.

Table 1. Parameter values per emotion for synthetic contours of condition 1. Freq: Endfrequency in Hertz; Exc: Size of the pitch movements in semitones; N: neutrality, J: joy, B: boredom , A: anger, S: sadness, F: fear, and I: indignation.

	N	J	B	A	S	F	I
Freq	65	155	65	110	103	200	170
Exc	5	10	4	10	7	8	10

2. *Two-component best matches to natural contours.*
Values for scaling of the declination line and for the excursion size of the pitch movements were determined to get a close fit to the natural contours (see table 2). The declination line was anchored at utterance onset rather than offset. Table 2 shows end frequencies instead of onset frequencies, to allow comparison with table 1. The distance between the second peak and the baseline (in semitones) was used to determine the excursion size for both pitch accents (this choice was inspired by the fact that the excursion size of the second peak varied much more in relation to emotion than that of the first peak). This means that in this condition the size of pitch movements was equal for both peaks.

Table 2. Parameter values per emotion for synthetic contours of condition 2. Freq: Endfrequency in Hertz; Exc: Size of the pitch movements in semitones; N: neutrality, J: joy, B: boredom , A: anger, S: sadness, F: fear, and I: indignation.

	N	J	B	A	S	F	I
Freq	95	125	100	145	125	160	180
Exc	6	12	5	8	8	9	9

3. *Peak modelling.*
The first peak was manipulated independently of the size of the second one, so as to make the relation between peaks as shown in Figure 1.

4. *Offset modelling.*
Starting from the utterances in condition 2, utterance-final low pitch for each emotion was modelled after the contours in Figure 1.

5. *Peak & offset modelling.*
Condition 5 combines the effects of conditions 3 and 4.

Since for neutrality and joy condition 2 provided accurate offset modelling, contours for conditions 2 and 3 were the same as conditions 4 and 5 respectively. Hence, there were only 31 test utterances instead of 35. A series of 55 stimuli was presented to the subjects; the first 19 gave an idea of the kind and amount of pitch variations allowed in the stimuli, the next 31 were the test-stimuli, and the last 5 were end-of-list fillers.

Sixteen subjects participated in this experiment, which involves a seven-alternative forced choice paradigm with the seven emotion labels. The subjects performed individual interactive listening tests. They listened only once to each stimulus and decided which emotion had been expressed. The 31 test stimuli were presented to different subjects in different random orders.

Results and discussion

Table 3 gives the number of subjects correctly identifying each emotion in the different conditions. Notice that the number of correct responses for neutrality and joy in conditions 2 and 3 are in parentheses. As explained, the contours for neutrality and joy in conditions 2 and 3 had the same utterance-final low pitch as the natural contours so that the contours generated in conditions 2 and 3 for these two emotions actually instantiated conditions 4 and 5. To compare an equal number of judgments for each condition, the results for these stimuli were also included in conditions 2 and 3.

Table 3. Number of correct responses per condition (C1-C5) and per emotion (N: neutrality, J: joy, B: boredom , A: anger, S: sadness, F: fear, and I: indignation.)

	N	J	B	A	S	F	I
C 1	3	3	11	2	2	6	3
C 2	(10)	(7)	6	0	0	5	3
C 3	(9)	(6)	6	1	4	7	3
C 4	10	7	2	0	4	7	9
C 5	9	6	5	0	3	5	5

The total number of correct responses is about the same for all conditions: 30 for condition 1 (Nmax=112), 31 for condition 2, 36 for condition 3, 39 for condition 4, and 33 for condition 5 (chance level = 16). Thus, we find that emotions are recognized better than chance on the basis of pitch alone, but that modelling of contour details does not lead to substantial improvement. Recognition performance in condition 5, which produces the closest match to the natural contours, is similar to the performance in conditions 1 and 2.

The rather low identification performance is probably due to the fact that no characteristics other than pitch have been manipulated. Especially anger and sadness gave poor performance. Previous investigations [1,5] suggested that voice source was an important component for sadness for example. Further looking at the detailed outcome, we observe that there is considerable trade-off between neutrality and boredom. Whereas the condition 1 values for neutrality (judged to be perceptually optimal in a previous study [5]) give a bias towards boredom, the values used for the conditions 2 to 5 give a bias towards neutrality. Further discussion of detailed aspects of the data and of confusions between emotions is beyond the scope of this paper.

CONCLUSION

In sum, we find that, even though there is considerable discrepancy between contours based on a simple two-component model and natural contours, it does not appear necessary to extend the two-component model in order to capture these differences. No clear improvement in recognition is obtained beyond what is achieved in terms of the two-component model.

Furthermore, it is clear that high recognition performance of emotions cannot be obtained through pitch manipulation only, and that other aspects such as duration and voice quality must also be taken into consideration.

ACKNOWLEDGEMENTS

This research was supported by the Co-operation Centre Tilburg and Eindhoven Universities (SOBU).

REFERENCES

[1] Vroomen, J., Collier, R., & Mozziconacci, S. (1993). Duration and intonation in emotional speech. *Eurospeech 93 ESCA Proceedings*, Berlin.

[2] Carlson, R., Granström, B., & Nord L. (1992). Experiments with emotive speech acted utterances and synthetic replicas. *ICSLP 92 Proceedings*, Alberta, 1, 671-674.

[3] Bezooijen, R. A. M. G. van (1984). *Characteristics and recognizability of vocal expressions of emotion*. Dordrecht, The Netherlands: Foris.

[4] Hart, J. 't, Collier, R., & Cohen, A. (1990). *A perceptual study of intonation; an experimental-phonetic approach to speech melody*. Cambridge: Cambridge University Press.

[5] Mozziconacci, S.J.L. (1994). Pitch and duration variations conveying emotions in speech. *IPO Report No. 961*. IPO, Eindhoven.

OBJECTIVE VOICE PARAMETERS TO CHARACTERIZE THE EMOTIONAL CONTENT IN SPEECH

Gudrun Klasmeyer, Walter F. Sendlmeier
TU Berlin, Institute of Communication Science, Germany
klasmeyer@kgw.tu-berlin.de

ABSTRACT

In the present study, the hypothesis is tested that voice parameters derived from clinical measurement of pathological voices (Jitter, Shimmer, Harmonics-To-Noise-Ratio) and the glottal pulse shape could serve as useful features to characterize specific emotional contents in spoken utterances.

DATABASE

The emotionally loaded speech material was produced by three students of acting (2 male, 1 female). It was DAT-recorded in separate sessions in an anechoic room using a B&K measuring microphone. The utterances are 10 short sentences frequently used in everyday communication which could appear in all emotional contexts without semantic contradictions. Each utterance was spoken several times in a neutral voice and several times with each of the following emotions: happiness, sadness, anger, fear, boredom and disgust. The most appropriate realisation was selected by the authors and the respective actor for a listening test.

LISTENING TEST

For each actor, the selected 70 sentences were randomized. Every sentence was repeated three times in short intervals followed by a 30 second pause. A 1 kHz tone announced the next triplet of sentences. The series of stimuli were acoustically presented via headphones to 20 naive listeners in separate sessions to evaluate the emotional content within 8 categories: neutral, happiness, sadness, anger, fear, boredom, disgust and not recognizable emotional content. Only those sentences recognized by at least 80% of all listeners were used for the analysis.

PARAMETERS

The sound of emotional speech differs from that of neutral speech, which is partly due to differing articulator movements and partly due to differeing glottis behaviour. In clinical measurement of pathological voices, sustained signals of open vowels are used for analysis. In fluent speech however, the role of articulator movements and intonation has to be considered. The glottal pulse signal can be derived from the acoustic speech signal by inverse filtering. The pulse shape contains information about glottis movement. In this context also voicing irregularities (Jitter [1] and Shimmer [1]) are discussed. Another parameter investigated in this study is the Harmonics-To-Noise-Ratio [2]. A phoneme based set of Energy Distribution Parameters is developed to differentiate between specific emotional contents in the frequency domain.

I. GLOTTAL PULSE SHAPE

Theoretically the process of glottal closure resembles an impulse. In the following closed-glottis-interval, the acoustic speech signal can be interpreted as impulse response of the vocal tract, because the subglottal volume is decoupled from the upper tract during this time. The filter coefficients for inverse filtering are calculated during the closed-glottis-interval. The actual point of glottal closure can be determined by inverse filtering with filter coefficients derived from a time interval (3-4 times) longer than one period duration of the acoustic signal [3]. In the present study, 18th order covariance LPC and rectangular data windows are used. The filter coefficients for inverse filtering are calculated during the closed-glottis-interval of the middle period of each realisation of the German phoneme /a/ in the emotional speech database. Due to different period durations, the length of the data window had to be adapted with regard to reasonable spectral shaping of the inverse filter. 100 ms of the speech signal are filtered. Only the middle period within which the LPC coefficients were calculated is examined further. In emotional speech the glottal cycles vary considerably from normal speech. So the inverse filtered signal should be interpreted as glottal pulse signal with great care, because some of the premises for this theory may be violated. For example, pulses filtered from sad speech hardly show any obvious closed-glottis-interval and (or because) the closure is not abrupt. But still the shape of the inverse filtered signal does represent important characteristics of the glottal cycle. In general it is more difficult to determine the exact point where the opening begins. The relative duration of the closing phase can be measured more reliably. To parameterize its shape the filtered signal is amplitude-normalized. The duration of the closed-glottis-interval (T1), the opening-phase (T2) and the closing-phase (T3) are measured as fractions of the full period. (See figure 1. for explanation.)

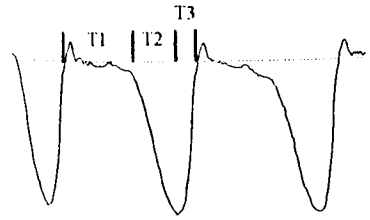


Figure 1. Parametrization of the inverse filtered signal

Within the speech database a relation between specific glottal pulse shapes and specific emotions could be proved. The general correlations are not speaker-dependent. For example in angry speech, pulses show very abrupt closure and remarkably long closed-glottis-intervalls. In contrast sad utterances show hardly any closed-glottis-intervals, and in anxious speech the relative duration of glottal closing is nearly similar to that of glottal opening. Table 1. gives the precise data.

	T1	T2	T1+T2	T3
Neutral	0.15	0.65	0.80	0.20
Happiness	0.33	0.43	0.76	0.24
Sadness	0.0	0.72	0.72	0.28
Anger	0.47	0.40	0.87	0.13
Fear	0.26	0.40	0.66	0.34
Boredom	0.25	0.48	0.73	0.28

Table 1. Average durations of closed-glottis-interval (T1), opening-phase (T2) and closing-phase (T3) as fraction of the full period

It is clear that narrow pulses with short opening and closing phases show less high frequency damping of their harmonics. Discussing the perceptible influence on the speech signal the absolute amplitude and period duration has to be taken into consideration, too. The pulse shapes measured in this study do not correlate with period duration. It can rather be assumed that pulse shapes correlate with loudness. This could not be tested, however, because the microphone signals for different specific emotions were recorded at the same level.

II. VOICING IRREGULARITIES

In clinical voice measurement, the patient produces isolated vowels with flat F0 contours, whereas in fluent speech there is a permanent rise and fall due to the intonation pattern. This implies that the parameter Jitter has to be measured taking into account such meaningful variation of F0. This is done by calculating a polynom approximation of the F0 contour and subtracting this approximation from the measured values. The difference values are used to calculate the absolute jitter, which has to be interpreted taking into consideration the absolute period duration [4].

In the emotional speech database remarkably high Jitter values were found in all anxious utterances produced by the female actress and in most anxious utterances produced by one male actor. This male actor also

laryngealized most vowels in sad utterances. The other male actor showed no voicing irregularities.
Shimmer was not found in the speech database used for this study.

III. HARMONICS-TO-NOISE-RATIO

In neutral speech the energy of harmonics is damped in higher frequencies, so there is very little energy above 4 kHz in vowels, whereas in fricatives there is very little energy below this frequency. In emotional speech, high frequency noise can be present in vowels which has its origin in abnormal articulation. For example with fear, the face can be stiff, and teeth are pressed together. This produces fricative noise in vowels, which does not derive from the voice source, but can be measured with the same parameter used in clinical measurement of pathological voices to detect noise produced by imperfect closure of the glottis. Especially in bored speech the articulation is imprecise, and voiceless fricatives in VCV clusters tend to be voiced. This effect can also be detected with the Harmonics-To-Noise-Ratio.

IV. SPECTRAL ENERGY DISTRIBUTION

Discussing the Harmonics-To-Noise-Ratio some phenomena were explained, by which energy is shifted to different frequency regions. Also specific glottal pulse shapes correlate with specific spectral damping of harmonic energy. From these considerations a set of Energy-Distribution-Parameters is developed on a phoneme basis to differentiate between specific emotional contents in spoken utterances. A spot check revealed that the introduction of 4 frequency bands leads to meaningful parameters for the characterization of different emotional contents in spoken utterances on a phoneme basis. The acoustic speech signal is lowpass filtered with an adaptable filter cutting all energy above F0. This is the very-low-frequency-band (VL). A second lowpass filter with constant 1.5 kHz cut off frequency is used to produce a low-frequency-band (L) signal. The middle-frequency (M) region between 1.5 and 4 kHz is filtered from the acoustic signal with a bandpass. The high-frequency (H) region between 4 and 8 kHz is filtered from the speech signal with a highpass. The energy distribution is compared within different emotionally loaded realisations of the same phoneme by measuring the energy within single bands as fraction of the total energy of that phoneme. As an example, the results for the vowel /a/ for one male speaker are presented in table 2. Significant frequency shifts for specific emotions are also apparent in fricatives. It is clear that the energy distribution within the frequency bands is speaker dependent; for example, female speakers have higher formants than male speakers. But the general tendency of energy shifts correlating with specific emotions is not speaker dependent. Ratios of different frequency bands can be calculated to discriminate specific emotions even stronger, but for a general survey the fractions of total energy are more illustrative.

	VL	L	M	H
Neutral	0.036	0.964	0.024	0.013
Happiness	0.069	0.879	0.045	0.017
Sadness	0.364	0.966	0.007	0.031
Anger	0.016	0.850	0.081	0.024
Fear	0.224	0.843	0.084	0.060
Boredom	0.170	0.988	0.006	0.007
Disgust	0.150	0.683	0.178	0.084

Table 2. Average distribution of energy within frequency bands as fraction of total energy in the vowel /a/

In the vowel /a/ spoken in a neutral voice most energy is below 1.5 kHz. In most emotionally loaded utterances some energy is shifted to the middle and even to the high frequency region. Only in sad and bored speech there is less energy in middle and high frequencies. A remarkable difference between fear and anger is that in angry voice there is little energy in the very-low-frequency band, though F0 is often high in angry utterances, whereas in utterances spoken with fear the appearance of energy in high frequencies is combined with high values in the very-low-frequency band. In bored and sad utterances there is also much energy in the very-low-frequency-band, even though F0 is usually very low in these utterances. The results of the energy-distribution measurement confirm very well what could be expected from the analysis of the glottal pulse shapes.
Examining the energy distribution on a phoneme basis leads to many interesting results. There is also emotion specific information in the shift of energy in vowel-fricative transitions and in the ratio of total vowel-to-fricative energy. The emotion specific differences in the distribution of energy are visualized in figure 2. The narrowband spectrogram of the preemphasized signal is calculated and amplitude-normalized. All values below a fixed threshold are presented in white, all values above this threshold are printed in black. In these 'binary spectrograms' different energy distributions in emotionally different realisations of the same utterance are obvious.

SUMMARY

It could be shown that parameters derived from clinical measurement of pathological voices and the glottal pulse shape are useful to characterize specific emotions in spoken utterances. Discrimination of these specific emotional contents is possible from the examination of spectral energy distributions on a phoneme basis. The results correlate very well with predictions from the examination of glottal pulse shapes.

REFERENCES

[1] Orlikoff R.F., Baken R.J., *Clinical Speech and Voice Measurement*, Sing.Publ.Group, 1993
[2] Deliyski D.D., *Acoustic Model and Evaluation of Pathological Voice Production*, Kay Elemetrics Corp., Dept. of Development and Research, 1993

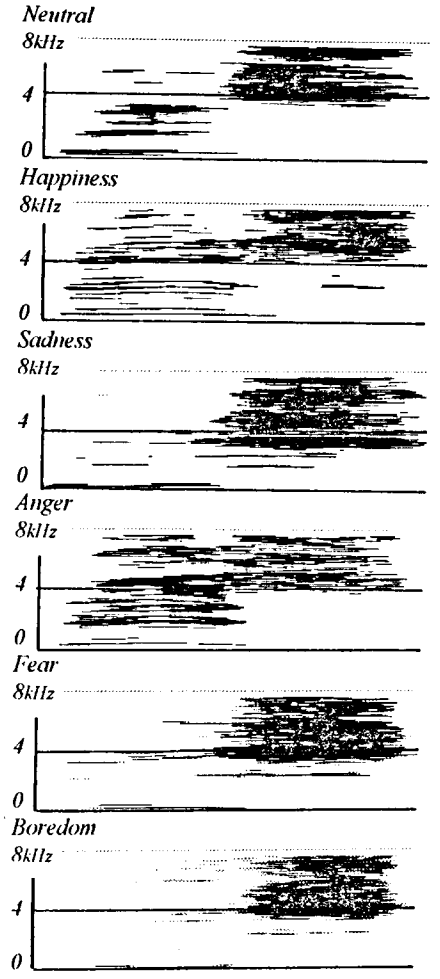


Figure 2. 'Binary Spectrograms' of diffrent emotionally loaded realisations of the phonemes /i/, /s/

[3] Wong D., Markel J.D., Gray A.H., *Least Squares Glottal Inverse Filtering from the Acoustic Speech Waveform*, IEEE Transactions ASSP Vol. 27(4), 1979
[4] Rosken W., Klasmeyer G., *Erfassung von F0-Irregularitäten in gesprochener Sprache als messbarer Parameter zur Beschreibung von Stimmqualitäten*, Fortschritte der Akustik, DAGA '95

AURAL IDENTIFICATION OF UNFAMILIAR VOICES BY COMPARISON

M. Falcone(*), F. Ferrero(+), A. Paoloni(*), P. Cosi(+)
(*) Fondazione Ugo Bordoni, Roma, Italy
(+) Centro di Studio per le Ricerche di Fonetica (CNR), Padova, Italy

ABSTRACT

The performance of human listeners is the final objective of the automated speech input systems. In speaker recognition the need of a human-reference to assess automated method is a common procedure. We address the problem of identify unfamiliar voices, assuming that the listeners use a limited amount of speech to create the reference template. Our aim is to define and assess standard procedures to evaluate listener's capability in speaker recognition.

INTRODUCTION

The human speaker recognition capability is based on two main characteristics of the speech signal: '*acoustic-phonetic*' matching, '*prosodic*' matching. In this work we are interested in the first item, so this paper is concerned with the following experimental situation: a subject listen to a pair of *short* utterances and then he/she had to decide if the listened utterances belong to the same speaker or no. This is a common situation in the experimental evaluation of listener's ability to perform some speaker verification task [1][2]. Unfortunately in the past these tests were mainly intended to provide a basis for comparison with performance of automated systems. So the test design result substantially different time to time, and performance analyses do not allow cross comparison among the several experiments.

THE TEST DESIGN: A PROPOSAL

The test consists of listening to a pair of the same word, spoken by same/different speaker and then to give a judge on the speaker identity. The stimuli we use are: monosyllabic, trisyllabic, polysyllabic (more than five, less than eight syllables).

The response is gauged to a fixed number of choices. We fixed he following constrains to the listening test material:

- the same speech segment is never presented twice to the same listener;
- the amount of speech signal presented is the same for all speakers used in the test;
- the number of same-speaker pairs is the same of different-speaker pairs;
- the frequency distribution of the used words is uniform.

If **NS** is the number of the speakers available, **NR** is the number of repetitions of the stimulus (for each speakers), and **NC** is the numbers of pairs to be presented to the listener, the previous conditions set the following rules:

$$NR = 4 * (NS - 1)$$
$$NC = 2 * NS * (NS - 1).$$

So you have the following possible solutions:

Table 1. A list of possible values to be used in order to have a "balanced" test

NS	NR	NC
2	4	4
3	8	12
4	12	24
☞ ***5***	***16***	***40***
6	20	60

and so forth.

The duration of the listening test should be about 30 minutes long. Considering that we want to use three different words, the pointed solution of NS=5 results a good choice, as the number of pairs for listening session is 3*40=120, i.e. we consider, on the average, a total duration of 15s for each pair presentation plus user response. The definition of standard procedures for listening test in speaker recognition is a very important point.

As speaker verification and identification technology finally seem to have reached a mature degree, we expect a renewed and greater interest on these topics. The test design we propose may be a good starting point.

INSTRUMENTAL SET-UP

The listening tests have been executed in Rome and in Padua. The used hardware was the same in both laboratories, i.e.: a personal computer equipped with an audio OROS AU2X board, a CD reader and a colour VGA monitor. In addition an external amplifier and a monitor headphone AKG mod.K141 complete the required hardware to run the experiment. As hearing level is a crucial point in any listening test, special attention has been devoted to the calibration of the whole audio instrumentation chain.

Calibration

This problem may be split in two parts: the digital 'calibration' of the speech files; the analogue 'calibration' of the electrical chain from the line out of the audio board to the output of the headphone. The numerical normalisation of the speech signal is executed on line during the restitution of the speech file. The normalisation factor has been computed in order to amplify to a fixed dB value the frame (26ms) of maximum energy of the given stimulus. So the frame-peak energy is the same for all the stimuli. To calibrate the electrical equipment a reference 1kHz sinusoidal tone is used (see CCITT G711 recommendation). A MCL (Most Comfortable Level) strategy has been used. A measure of the mean speech levels after calibration, stated a value about 80dBA, that is, according to the measures reported in literature, a *reasonable* calibration level.

EXPERIMENT DESCRIPTION

The experiment is described by a test control file that contains information on the utterances to be played and the relative normalisation factors, as well the number of pairs to be presented and the filename where the results are saved. We build four tests: each test consists of 120 pairs.

Table 2. The list of the utterances used in the four tests executed in the experiment

	1sill	3sill	polisill
Test1	si	cancella	unoecinquedi
Test2	no	corretto	aeffebiotto
Test3	tre	indietro	novesetteduetre
Test4	sei	esegui	richiestadiaccesso

The speech material is part of the SIVA database [3], and it is real telephonic quality signal. We only use five speakers in this experiment; the same for all four tests. They belong to the same regional area of the South of Italy. The listening sessions have been executed in Rome (central Italy), and Padua, (north Italy) where listeners have not acquaintance with the southern speaker behaviours. Listening session have been executed in a *silent* room. Listeners do not perform any training, they only receive a page of written description of the test and relative instructions. Each laboratory contributed with 5 sessions per test, for a total of 20 tests. In summary we have responses on 4800 pairs' presentation. The subject can not listen more than once a pair. In fact after the pair presentation a menu describing the following four choices:

1.voices are certainly different
2.voices are probably different
3.voices are probably the same
4.voices are certainly the same

is displayed, and after the subject's selection the relative choice is highlighted on the monitor and the other choices are cancelled, then a confirmation is requested before the next pair will be submitted, otherwise the main menu is displayed again for a new selection. So corrections are possible, but the subject is not allowed to listen more than once the same pair.

ANALYSIS OF THE RESULTS

The obtained results have been analysed separately for the two groups, and then compared. Our goal is to measure the "human" performance in comparing speech samples in relation to the duration of the utterance (fig.1, fig.4); to analyse the listener variability in performing the identification task (fig.2, fig.5); and last to trace a '*relative operating characteristic*' (ROC) of the 'human' system in solving the given task. Direct measures of the obtained performances are the '*false acceptance error rate*', (FA) and the '*false rejection error rate*' (FR). The first is also referred as error TYPE I° and it is the probability that utterances of two different speakers

are assigned to the same person, it is the most severe error as it measures the probability that an impostor get in your system. The second is also referred as error TYPE II° and it is the probability that utterances of the same speaker are assigned to two different speakers, it is a less severe error as it measures the probability that your system do not let you get in, although you have the authorisation. These are 'crude' measures that only reflect the YES/NO decision taken. A more interesting measure is the ROC (fig3, fig.6) [4]. This is a standard XY dispersion plot where on the X axes there is the ***probability of listener deciding same when samples are, in fact, by different speaker*** (error), while in the Y axes there is the ***probability of listener deciding the same when samples are, in fact, by the same speaker*** (correct). If you have a total (positive plus negative) N rating scale the result is a set of (N-1) points on the graph. The fitting of these points, plus the origin point (0;0) and the infinite point (1;1), gives you the ROC curve. Roughly speaking, we may say that curves approximating the diagonal line from (0;0) to (1;1) describe more difficult tasks. If a real (automated) system measures a distance between two samples, it will be possible to set several thresholds and design a *real* ROC; in case of listening test this is *simulated* using a rated scale. Unquestionably a ROC curve gives a more detailed information than FA and FR values, but the standard procedure, well described in [4], considers the rating scale a linear discriminative scale.

Rome group result

The results obtained from the FUB group confirm the well-known fact that the performances in identifying speaker do not vary meaningfully after the 1.2 second duration. We see from fig.1 that errors decrease about 20% if we move from monosyllabic words to trisyllabic words, but only 5% from trisyllabic to polysyllabic words. This goes against our intuitive belief, but it is a well experimental accepted fact. From fig.3 we realise that the listener population has a great variability and that some subject also has a *strange* behaviour. For example one subject has a total error that is greater than 50% (a random generator works better!) and another has a 0% FA error and a 32% FR error (he/she is a good guardian!). Finally we had to compare the fig.1 and fig.3. They both report FUB listener group performance in relation to the used word, but in the first case we only use the binary Y/N information, while in the second we also utilise the degree of confidence the listener express utilising the two rates scale.

Padua group result

As expected the CNR results follow the same trend of the Rome listener group. We only find two light differences. First the overall error (FA+FR) is just a little bit greater, and this may be because northern listener may have less familiarity than central listener with the used database. Second the FA/FR ratio is smaller. This fact may not be explained easily. Also for CNR listener group we have some subject with strange behaviours. For example we have, again, a listener that has a 0% FA error, and another with a total error greater than 50%. The ROC curves clearly set that the speaker identification using monosyllabic word is really a hard task, while it makes no particular differences recognising people using trisyllabic words or polysyllabic words.

CONCLUSION

We execute a round-robin experiment for the evaluation of human capability in speaker recognition, when pairs of short utterances are submitted to the listener. Particular attention has been devoted to the calibration and balancing of the test itself, to avoid drift effects. The obtained results show high consistency among the two groups, and clearly set that: listeners belonging to a regional area farther (in a phonetically sense) from the one of the speaker to be recognised, have, on the average, a worse performance of few percents; performance response in relation to word length shows a threshold effect situated between monosyllabic and trisillabic words (other works report a value around 1.2s) for pairs of words or short utterances. The results are promising and although more efforts are necessary before a final solution will be reached, the possibility of using standard listening test as a reference in speaker recognition seems a good choice.

REFERENCES

[1] Stevens, K.N., et alii, "Speaker Authentication and Identification", *J.A.S.A.*,vol.44,n.6,pp.1596-1607

[2] Federico, A. et alii, (1989) "Comparison between automatic methods and human listeners in speaker recognition task", *proceedings EUROSPEECH 89*, pp.279-282

[3] Falcone, M., Contino U., (1995) "Acoustic characterisation of Speech Databases: An Example for the Speaker Verification", *these proceedings*

[4] Kecker, M., (1971), *Speaker Recognition: An Interpretive Survey of the Literature*, AHSA Monographs n.16

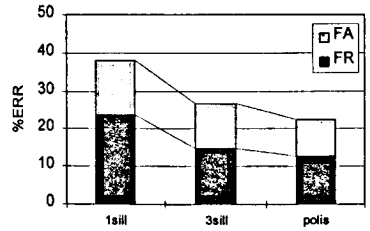


Figure 1. False Acceptance (FA) and False Rejection (FR) in relation to the utterance length. FUB listener group.

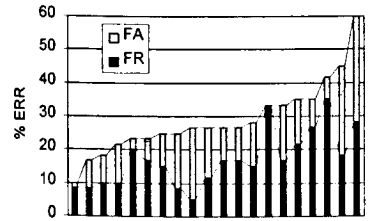


Figure 2. False Acceptance (FA) and False Rejection (FR) for each single listener. FUB listener group.

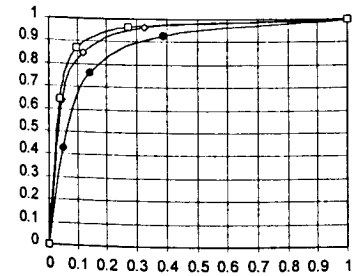


Figure 3. ROCs for monosyllable (•), trisyllable(+), and polysyllable (□) utterances. FUB listener group.

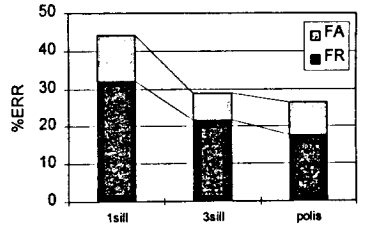


Figure 4. False Acceptance (FA) and False Rejection (FR) in relation to the utterance length. CNR listener group.

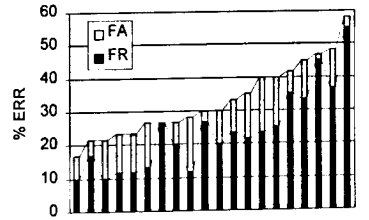


Figure 5. False Acceptance (FA) and False Rejection (FR) for each single listener. CNR listener group.

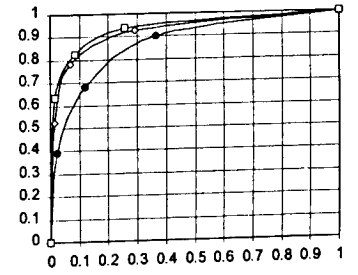


Figure 6. ROCs for monosyllable (•), trisyllable(+), and polysyllable (□) utterances. CNR listener group.

ON THE IDENTIFICATION OF FAMILIAR VOICES

Judith Rosenhouse
Dept. of General Studies

Yizhar Lavner and Isak Gath
Dept. of Biomedical Engineering

The Technion, I.I.T., Haifa, Israel

ABSTRACT

This study aimed at examining voice identification levels in and by a relatively large group of subjects. These subjects had a high degree of familiarity as they lived in the same kibbutz. We analyzed 5 modern Hebrew vowels, 5 voiced consonants, and the Hebrew phrase "good morning". The results suggest differential perception and identification processes among subjects and among speech elements. We suggest an explanation to the results in the framework of the prototype model.

PROBLEM

Voices of different people sound different, and a few seconds of speech suffice to identify a speaker without being in direct visual contact with him/her. Speaker identification is possible under various detrimental circumstances, after long time periods, in various speech contexts, when the speaker expresses different attitudes, etc. These facts seem to imply that there are some acoustic features that do not change under different conditions. Despite these observations, the topic is still not well understood as to the parameters that affect speaker identification or how much they are related to phonetics.

Human speaker recognition has been defined as any process of decision about the identity of a speaking person by certain features of the speech signal. A presupposition for such a decision is previous acquaintance with the speaker.

Acoustic characteristics of voices have thus been described, inter alia, in long-term- and short term-, inherent- and learned-, glottal source- and vocal tract-dependent features.

Listeners' role in voice identification has been studied so far from numerous angles, e.g., voice recognition learnability, time effect on recognition (memory), number of voices recognized, test type, utterance type/length effect on recognition, language dependence, masking effects, the effect of various acoustic parameters on recognition, etc. Experiments often used small numbers of subjects and/or voices.

This paper has the following goals: 1. testing voice identification of a large group of speakers by a large group of listeners well acquainted with the speakers; 2. Phoneme-dependency of voice identification. Reported here are results of our psycho-acoustic tests, performed as part of a systematic study of acoustic cues important for voice identification as described.

METHOD AND SUBJECTS

The method includes a few stages for recording the test material and testing the subjects.

The subjects were from a kibbutz in the north of Israel, all native speakers of Hebrew without speech or hearing impairments or foreign features.

In Stage I, 20 men of this kibbutz (age range: 26-59) were recorded saying the same test materials (see below). They were recorded (mono-channel) on a 486 PC computer using a voice card at 22 kHz sampling rate.

In Stage II the listeners were men and women from the same kibbutz (age range: 25-55). All know the speakers well or very well. Each listener was asked to fill in a form grading in a 5 grade scale his/her acquaintance with the speaker. There were 28 people's voices, 20 of which were later used in the tests. Subjects had to grade in a 5 grade scale the "uniqueness" of each speaker's voice, and describe this feature in words.

Speaker identification tests included recognition by: 1. /a, e, i, o, u/, the 5 vowels of Hebrew uttered in isolation; 2. /aCa/ syllable sequences, C being the nasals /m,n/, based on the literature which described them as good predictors of individual features and /l, r, z/ which tend to have numerous allophones; and 3. the 2-word Hebrew utterance /'boker 'tov/, i.e., 'good morning'.

Each session lasted up to an hour and a half, in which each listener heard a 100 vowels of 20 different speakers in random order. The listeners were asked to identify the speakers from a list of 28 people's names, i.e., more speakers than actually used in the test. After hearing the vowels, they had to fill in another questionnaire in which they wrote down the speaker's name (as they identified him), their confidence level in it, their evaluation of this voice's uniqueness and (optionally) a verbal description of the voice. They were allowed up to 8 times listening to each stimulus. On the same session they were also asked to identify speakers by the utterance "good morning". Speaker identification by /aCa/ syllables was tested in a separate session.

The tests were aimed to give answers to the following questions: 1. What is the average speaker identification level by individual listeners? 2. Are there inter-listener differences in speaker identification? 3. Are there inter-listener differences in speaker identification by different phonemes? 4. Does successful idenification of a voice by a certain phoneme imply successful speaker identification by the same phoneme by other listeners?

FINDINGS

1. The identification test of the utterance /'boker 'tov/ yielded an average correct identification rate of 60% (216 correct identifications for 360 stimuli). The test was an open test, in the sense that the listeners did not know which speaker out of 28 possible people they were about to hear. This proportion of successful identification is much higher than reported in previous studies (see e.g., Ladefoged & Ladefoged, 1980, van Lancker et a., 1985). The listeners identified speakers, except for three cases where the voices of certain speakers were erroneously considered those of others. Thus, most errors were of the type "cannot identify". In addition, the successful identification range by individual listeners (45% - 85%) is smaller than the successful identification range of the speakers' voices (11% - 100%).

2. Speaker identification by voice recognition of vowels: The results of 20 listeners (men and women) were included in the data analysis and are summarized in Table 1 and demonstrated as an example in Table 3. The results show that there are vowel-dependent significant differences in listeners' identification abilities. The best identification was yielded for /a/ - 37.6%. Next come /i/ and /e/ without any difference between them - 29%. These three vowels are better identified than /o/ (25%) and /u/ (17%) (See Table 1). The average identification rate for the total number of vowels was 29% (range: 16%-51%), which is much lower than for the words (60%). This result may be expected owing to the little information in isolated vowels as compared to two-word-utterances.

Table 1. Percentage of speaker idetification by vowels

vowel	N	correct	percent
/a/	330	124	37.6%
/e/	295	89	30.2%
/i/	300	87	29.1%
/o/	285	72	25.3%
/u/	189	32	16.9%

3. Speaker identification by voice recognition of voiced consonants in /aCa/ syllables: Differences were also found for identification of speakers in this environment. The best identified consonant in this environment was /z/ (63%), followed by /n/ (62%) and /m/ (58%). The speakers of syllables with /l/ were correctly identified in 53% of the cases, and for /r/ - in 50% of the stimuli (see Table 2).

Table 2. Percentage of speaker idetification by consonants

consonant	N	correct	percent
/r/	198	99	50.0%
/l/	198	105	53.0%
/m/	198	114	57.6%
/n/	197	123	62.4%
/z/	198	124	62.2%

4. Confusion matrix results revealed that some of the speakers were more successfully identified for certain vowels than other speakers (See for example, Table 3).

5. Speakers whose voices were correctly identified by all the listeners in all the phonemes had special vocal features.

6.Most speakers had considerable pitch variations even for relatively short utterances (<300 ms.). But as F0 ranges were very similar for most speakers, it may be assumed that F0 is not the most important cue for speaker identification, at least in isolated vowels.

DISCUSSION

This paper presents results of a study of speaker identification by human listeners. The test lanaguage was Hebrew, which to the best of our knowledge, has so far not been studied from this respect. For the purpose of this study both speakers and listeners were native speakers of this language.We are dealing here with Modern Hebrew,a Semitic language with a phonetic system comprisng 5 vowels and 20 consonants (traditionally 22 consonantal and 10 vowel phonemes). This is apparently the first report on this issue based on such a new source of database.

Another issue that this study tackles is the number of subjects used in the tests. Most previous experiments used very small numbers (e.g., 3,5,7) of listeners and/or recordings. The present experiments were performed with a much larger group of subjects, both speakers and listeners, and thus results are probably more valid.

The tests can be considered of the open-test type, in the sense that the listeners did not know which persons of the list were going to be heard. In this sense, this test type is closer to the real-world situation of speaker identification.

At least two basic models for speaker identification by listening can be suggested:

1. All listeners use one and the same voice identification strategy using the same features.

2. Different listeners use different strategies to identify speakers' voices.

The results of our tests suggest a third model which combines the above two to some extent:

3. Listeners use the same strategy for speaker identification but different acoustic features of the speakers' voices. This model is based on the protype model (Rosch 1973, Rosch, 1976). According to this model, learning a new voice is achieved by comparing it to a prototype voice (e.g., mens' vs. women's or children's voices) and extracting from this comparison those features which deviate from the prototype pattern. Thus, voices which are less similar to the prototype will be easier to learn and memorize than voices similar to it. Thus, the more a voice deviates from the prototype, it will also be easier to identify it when presented as a stimulus for identification, and vice versa: the more similar it is to the prototype the harder it will be to identify it. The results of our experiments and of other experiments reported in the literature make this a likely hypothesis. This hypothesis is also useful, for it allows predicting results of other experiments.

Further research is required to prove whether this model is correct. Current research in speech sciences often applies acoustic analysis of speech signals and systematic resynthesis while controlling individual features and observing listeners' responses. We intend to use this method to examine the prototype model in the next stage of our study of voice identification.

REFERENCES

1. Ladefoged, P. and J. Ladefoged (1980) "The Ability of listeners to identify voices" UCLA Working Papers in Phonetics, 49, 43-51.

2. van Lancker D., J. Kreiman and K. Emmorey (1985) "Familiar voice recognition: patterns and parameters, Part 1: Recognition of backward voices" Journal of Phonetics, 13, 19-38.

3. Rosch, E. (1973) "On the internal structure of perceptual and semantic categories" in: T. E. Moore, (ed.) Cognitive Development in the Acquisition of Language, New York: Academic Press.

4. Rosch, E., C.B. Mervis, W.D. Gray, D.M. Johnson and P.Boyes-Braem,(1976) "Basic objects in natural cetegories", Cognitive Psychology, 8, 382-439.

Table 3. Stimulus-Response Confusion matrix for the vowel /a/

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	other	unrec	total
	1				7				2											1	1					2	3	16
	2	4	4										1					1			1					2	4	17
	3			12														1				2						16
S	4				12		1								2											1	1	16
	5					8								1									3			3		15
T	6				1		5							1												4	5	16
	7							4									4			3						4	1	16
I	8	3			3				5																		5	16
	9									6						1										2	7	16
M	10		1					1			7									1						1	5	16
	11								1			15															1	17
U	12				1			1					8								1					3	3	17
	13	1						2	6	1				1	1						1					1	3	17
L	14				2										4	1					1					3	6	17
	15		1		1			1								4										1	8	16
U	16				2				1								4			1						3	3	14
	17				1				1			2	1					3							7	1	1	17
S	18													1					13							1	1	16
	19				4				1															4		4	3	17
	20		1						1												8					4	3	17
	total	8	7	12	34	8	6	9	18	7	7	18	9	4	7	6	8	5	13	6	13							325
	false	8	3	0	22	0	1	5	13	1	0	3	1	3	3	2	4	2	0	6	5	2	3	4	7	40	138	

1÷24 - SUBJECTS' I.D. NUMBERS; OTHER: VOICES NOT USED AS STIMULI; UNREC: UNRECOGNIZED VOICES

TALKER- AND TASK-SPECIFIC PERCEPTUAL LEARNING IN SPEECH PERCEPTION

Lynne C. Nygaard and David B. Pisoni
Speech Research Laboratory, Indiana University, Bloomington, Indiana, U.S.A.

ABSTRACT

The present investigation was designed to assess the specificity of perceptual learning employed in the linguistic processing of spoken language. Two groups of subjects were trained to identify a set of talkers from sentence-length utterances. After training, one group of subjects was tested with isolated words produced by familiar or unfamiliar talkers and the other group was tested with sentence-length utterances. The results showed that the ability to identify a talker's voice from sentence-length utterances only modestly improved intelligibility of isolated words, but significantly improved the intelligibility of sentence-length utterances. Listeners appeared to focus their attention during perceptual learning on talker information that is specific to sentence-length utterances. The results suggest that task- as well as talker-specific perceptual learning occurs during the processing of spoken language.

INTRODUCTION

The speech signal simultaneously carries information about a talker's voice and about the linguistic content of the intended message. Traditionally, the unraveling of talker and linguistic information has been characterized as a normalization process in which talker information is discarded in the listener's quest for the abstract, idealized linguistic processing units thought to underlie speech perception [1,2]. Recent studies, however, have demonstrated that the processing of voice and the processing of linguistic content are not independent. Nygaard, Sommers, and Pisoni [3] found that learning a talker's voice facilitates subsequent phonetic analysis. In their study, listeners were trained to identify talkers' voices from isolated words and were then given a word intelligibility task. Listeners who heard familiar talkers at test were better able to extract the linguistic content of isolated words than those who heard unfamiliar talkers at test. The results suggest that perceptual learning of voice can modify the linguistic processing of isolated words.

The present investigation was designed to assess the nature and extent of this kind of perceptual learning. Subjects in two experiments were trained to recognize a set of ten talkers from sentence-length utterances.

In Experiment 1, after training was completed, intelligibility was assessed using isolated words produced by familiar and unfamiliar talkers. The aim was to determine if the information learned about a talker's voice from sentences generalizes to the perception of spoken words. The assumption was that training with sentence-length utterances would focus listeners' attention at a different level of analysis than training with isolated words. It was hypothesized that because sentences contain extensive prosodic and rhythmic information in addition to the specific acoustic-phonetic implementation strategies unique to individual talkers, perceptual learning of voices from sentences would require attentional and encoding demands specific to those test materials.

In Experiment 2, after training was completed, listeners were given an intelligibility test consisting of sentence-length utterances produced by familiar and unfamiliar talkers. Two issues were addressed here. First, does specific training on sentence-length utterances generalize to similar test materials? Second, are sentence-length utterances which have higher-level semantic and syntactic constraints susceptible to the effects of familiarity with a talker's voice?

EXPERIMENT 1

In Experiment 1, two groups of subjects learned to identify talkers' voices from sentence-length utterances over a three-day training period. The experimental group was then tested with isolated words to assess intelligibility of talkers they had been exposed to in training. The control group was tested with isolated words produced by a set of unfamiliar talkers.

METHOD

Subjects

Subjects were 33 undergraduate and graduate students at Indiana University. Sixteen subjects served in the experimental condition and seventeen subjects served in the control condition. All subjects were native speakers of American English and reported no history of a speech or hearing disorder. Subjects were paid for their participation.

Stimulus Materials

Two sets of stimuli were used in this experiment. The sentence training stimuli consisted of 100 Harvard sentences produced by 10 male and 10 female talkers. The isolated word stimuli consisted of 100 monosyllabic words produced by 10 of the same talkers (5 male and 5 female) that produced the sentence materials. All stimuli were digitized on-line at a sampling rate of 20 kHz using 16-bit resolution. The root mean squared (RMS) amplitude levels for all stimuli were digitally equated.

Procedure

Pretest Word Intelligibility. A pretest-posttest design was used in this experiment to directly evaluate the effects of talker familiarity on word intelligibility. In both pretest and posttest, 100 isolated words produced by ten talkers (5 male and 5 female) were presented at either 80, 75, 70, or 65 dB (SPL) in continuous white noise low-pass filtered at 10 kHz and presented at 70 dB (SPL), yielding four signal-to-noise ratios: +10, +5, 0, -5. An equal number of words was presented at each of the four signal-to-noise ratios. Subjects were asked to recognize the word by typing their response on a keyboard. For subjects in the experimental condition, the words were produced by the ten talkers they heard in training. For subjects in the control condition, the talkers' voices were unfamiliar.

Training. Two groups of listeners completed three days of training to familiarize themselves with the voices of ten talkers. The experimental group of 16 subjects learned the voices of the same ten talkers that were used for the pre- and posttests. The control group of 17 subjects learned the voices of ten different talkers. Both groups were required to identify each talker's voice and associate that voice with one of 10 common names.

On each day of training, both groups of listeners completed three different phases. The first was a *familiarization task* in which one sentence from each talker was presented in succession. Each time a sentence was presented, the name of the talker appeared on a CRT screen in front of the listener. Subjects were asked to listen carefully to the words presented and to attend specifically to the talker's voice.

The second phase of training consisted of a *recognition task* in which subjects were asked to identify the talker who had produced each sentence. Ten sentences from each of ten talkers were presented in random order to listeners who were asked to identify each voice by pressing the appropriate button on a keyboard. On each trial, after all subjects had responded, the correct name appeared on a CRT screen.

The third phase of training was identical to the second phase except that no feedback was given.

Posttest Word Intelligibility. The posttest was identical to the pre-test. Subjects were asked to identify isolated words produced by familiar or unfamiliar talkers at four signal-to-noise ratios.

RESULTS AND DISCUSSION

Training

All subjects showed continuous improvement over the three days of training. Both groups of subjects identified talkers consistently above chance even on the first day of training and performance rose to nearly 85% correct by the last day of training. A repeated measures analysis of variance (ANOVA) with learning and days of training as factors showed a significant main effect of day of training, $F(2,62) = 74.04$, $p<.001$, and also a significant main effect of group $F(1,31) = 20.27$, $p<.001$. The control group performed significantly better than the experimental group learning their set of talkers.

Isolated Word Intelligibility

Figure 1 shows the difference in percent correct word identification from pretest to posttest for both the experimental and control groups averaged across signal-to-noise ratio. Although there is more improvement for subjects in the experimental condition who were hearing familiar voices at posttest than for subjects in the control condition, the effects of familiarity on word intelligibility were small ($p<.08$). A repeated measures ANOVA with signal-to-noise ratio and training group as factors showed no significant main effects or interactions.

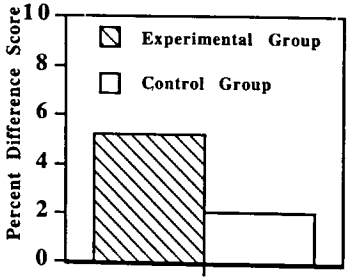


Figure 1. Percent difference is plotted for the control and experimental groups.

These results suggest that perceptual learning of talkers' voices from sentence -length utterances does not generalize to the perception of isolated words.

EXPERIMENT 2

As in Experiment 1, two groups of subjects learned to identify talkers' voices from sentence-length utterances over a three-day training period. However, the experimental and control groups in this experiment were then tested with sentences produced either by talkers they had encountered in training (experimental) or by a set of unfamiliar talkers (control).

METHOD

Subjects

Subjects were 20 undergraduate and graduate students at Indiana University. Eleven subjects served in the experimental condition and nine subjects served in the control condition. All subjects were native speakers of American English and reported no history of a speech or hearing disorder. Subjects were paid for their participation.

Stimulus Materials

Training and test stimuli consisted of 100 Harvard sentences produced by 10 male and 10 female talkers. All stimuli were digitized on-line at a sampling rate of 20 kHz using 16-bit resolution. The root mean squared (RMS) amplitude levels for all stimuli were digitally equated.

Procedure

Training. Training was identical to that used in Experiment 1 except that subjects were trained on a set of 50 sentences rather than 100 sentences. Again, two groups of listeners completed the three days of training. The experimental group of 11 subjects learned the voices of the same ten talkers that were used for the sentence intelligibility test. The control group of 9 subjects learned the voices of ten different talkers. All other aspects of training were the same as in Experiment 1.

Sentence Intelligibility Test. In the sentence intelligibility test, 48 novel sentences produced by ten talkers (5 male and 5 female) were presented at either 75, 70, or 65 dB (SPL) in continuous white noise low-pass filtered at 10 kHz and presented at 70 dB (SPL), yielding three signal-to-noise ratios: +5, 0, -5. An equal number of words was presented at each of the three signal-to-noise ratios. Subjects were asked to transcribe the sentence on a sheet of paper. For subjects in the experimental condition, the sentences were produced by the ten familiar talkers they heard in training. For subjects in the control condition, the talkers' voices were unfamiliar.

RESULTS AND DISCUSSION

Training

All subjects showed continuous improvement over the three days of training. As in Experiment 1, both groups of subjects identified talkers consistently above chance even on the first day of training and performance rose to nearly 85% correct by the last day of training. A repeated measures analysis of variance (ANOVA) with learning and days of training as factors showed a significant main effect of day of training, $F(2,36) = 78.029$, $p<.001$, and no other

significant effects.

Sentence Intelligibility

Subjects' performance on the sentence intelligibility task was assessed by determining the number of key words correct in each test sentence, adding up the total number of correct key words across sentences and averaging these totals across subjects. Each Harvard sentence contained 5 "key" words and the test set of 48 Harvard sentences contained 240 key words.

Figure 2 shows the total number of key words correct averaged across subjects for the experimental and control groups.

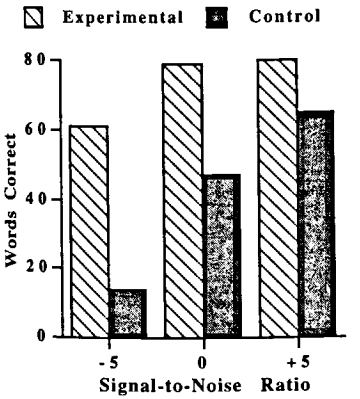


Figure 2. Percent key words correct as a function of signal-to-noise ratio for the experimental and control groups.

A repeated measures ANOVA with signal-to-noise ratio and training group as factors showed a significant main effect of training group, $F(1,18) = 220.378$, $p<.001$, indicating that subjects in the experimental condition who heard sentences produced by familiar talkers were able to transcribe more key words correctly across all signal-to-noise ratios than control subjects who heard sentences produced by unfamiliar talkers. A significant main effect of signal-to-noise ratio, $F(2,36) = 286.26$, $p<.001$, was also found indicating better performance at the higher signal-to-noise ratios. Finally, there was a significant interaction between training group and signal-to-noise ratio, $F(2,36) = 44.41$, $p<.001$, indicating that the effect of talker familiarity became larger as signal-to-noise ratio decreased.

These results suggest that perceptual learning of talkers' voices from sentence -length utterances facilitates the linguistic processing of sentence-length utterances produced by familiar talkers.

GENERAL DISCUSSION

The results of our experiments suggest that perceptual learning in speech perception is both talker- and task-specific. Perceptual learning of voice transfers to linguistic processing of spoken language in a task-specific manner such that attention must be directed to learning the specific voice attributes that will be relevant at test. Our findings also show that long-term talker-specific effects on linguistic processing occur with sentence-length materials which contain higher-level semantic and syntactic constraints suggesting that talker-specific effects operate in a variety of listening situations from isolated words to sentence-length utterances.

Familiarity with a talker's voice involves long-term modification of speech and language processing. Listeners appear to retain talker-specific information about individual articulatory idiosyncrasies both at the level of acoustic-phonetic implementation and at a more global level found in sentence-length utterances.

ACKNOWLEDGMENTS

We are grateful to Luis Hernandez for technical support, and to Lisa Burgin and Matt Peuquet for subject running. This work was supported by NIDCD Training Grant DC-00012 and by NIDCD Research Grant DC-00111 to Indiana University.

REFERENCES

[1] Halle, M. (1985), "Speculations about the representation of words in memory," In V.A. Fromkin (Ed.), *Phonetic Linguistics* (pp. 101-104). New York: Academic Press.

[2] Joos, M.A. (1948), "Acoustic Phonetics," *Language,* vol. 24, pp. 1-136.

[3] Nygaard, L.C., Sommers, M.S., & Pisoni, D.B. (1994), "Speech perception as a talker-contingent process," *Psych. Sci.,* vol. 5, pp. 42-46.

SOME SOURCES OF VARIABILITY IN SPEECH INTELLIGIBILITY

Ann R. Bradlow, Gina M. Torretta and David B. Pisoni
Speech Research Laboratory, Indiana University, Bloomington, Indiana, U.S.A.

ABSTRACT

Talker-specific correlates of intelligibility were explored using a large, multi-talker speech database. This database includes both sentence productions from multiple talkers and intelligibility data from multiple listeners. We examined global, talker-specific characteristics (e.g. gender, fundamental frequency, and overall speaking rate), as well as individual differences in phonetic implementation (such as vowel space compactness and fine-grained, segmental differences) as possible correlates of variation in overall talker intelligibility. Results indicated that individual differences in segmental articulation, rather than global characteristics, correlated well with overall intelligibility.

INTRODUCTION

The speech signal simultaneously encodes both linguistic and paralinguistic information [1]. Thus, in response to an utterance, a listener is made aware of both its content (the linguistic message) and of a host of information specific to the instance of the utterance. For example, due to both inter- and intra-talker differences, the speech signal conveys information about the talker's sex, age, geographical origin, physical and mental state, as well as the linguistic message he or she is trying to communicate. As a consequence of the simultaneous encoding of linguistic and paralinguistic information, we might expect an interaction between these two aspects of the acoustic signal. This study addressed this issue by investigating the correlation between talker-specific characteristics and speech intelligibility.

METHODS AND MATERIALS

The materials for this study came from the Indiana Multi-Talker Sentence Database. This database consists of 100 Harvard sentences [2] produced by 20 talkers (10 males and 10 females) of General American English. The sentences are all mono-clausal and contain 5 key words. Examples of the sentences are given below in Table 1.

Table 1. Two sample Harvard sentences with keywords underlined.

Rice is often served in round bowls.
Two blue fish swam in the tank.

Along with the audio recordings, this database includes intelligibility data in the form of sentence transcriptions by 10 listeners per talker. In the collection of these transcriptions, the listeners heard the full set of 100 sentences produced by a single talker. The listeners heard each sentence over headphones, and then typed what they heard at a computer keyboard. The sentences were presented in the clear (no noise was added) at a comfortable listening level. The listeners were all students at Indiana University with no speech or hearing impairments.

The sentence transcriptions were scored by a criterion that counted a sentence as correctly transcribed if, and only if, all 5 keywords were correctly transcribed. Any error on a keyword resulted in the sentence being counted as mistranscribed. With this scoring method, each sentence for each talker received an intelligibility score out of a possible 10. Each talker's overall intelligibility score was then calculated as the average intelligibility score across all 100 sentences. Across all 20 talkers, there was considerable variation in overall intelligibility. The intelligibility scores ranged from 81% to 93%, with a mean and standard deviation of 88% and 3%, respectively. Thus, the materials in this database covered a range of talker intelligibility and could be used as the basis for an investigation of the effect of talker-specific characteristics on overall intelligibility.

Our general approach to this investigation was to focus on two aspects of talker-specific variation. First, we examined the correlation of global talker characteristics, such as gender, overall speaking rate, and fundamental frequency, with overall talker intelligibility. Second, we looked at several aspects of the acoustic signal that provide information about the pronunciation characteristics of the talker. Specifically, we compared vowel space compactness across talkers, and performed an analysis of specific listener errors and their correlation with fine-grained talker variation at the acoustic-phonetic level.

GLOBAL CHARACTERISTICS

One of the most salient paralinguistic factors that is conveyed by the speech signal is the sex of the talker. Due to physical differences between the typical male and female vocal apparatus, as well as due to socio-linguistically determined differences between male and female pronunciation patterns, the sex of the talker is a very prominent paralinguistic factor. Furthermore, there is evidence in the literature that females tend to exhibit fewer instances of reduced speech than males [3]. Thus, we might expect female talkers to have higher overall intelligibility scores than male talkers.

In the Indiana Multi-Talker Sentence Database, the overall intelligibility scores indicated a significant sex-based difference in sentence intelligibility. The female talkers had a significantly higher average intelligibility score than the male talkers (89.4% versus 86.3%, with standard errors of 0.67% and 1.00%, respectively, $t(18)=2.57$, $p=.02$ by an unpaired, 2-tail t-test). Furthermore, in this database, the four talkers with the highest intelligibility scores were female and the four talkers with the lowest intelligibility scores were male. Thus, these data suggest that overall speech intelligibility is affected by the talker's sex. We now turn to an investigation of other paralinguistic factors that might help to explain the acoustic-phonetic reasons for this sex-based difference in intelligibility.

Overall rate of speech is a global characteristic of speech production that not only varies across talkers, but also has an impact on speech perception [4]. Using mean sentence duration as a measure of overall speaking rate, we investigated the correlation between overall rate and intelligibility. We hypothesized that slower speaking rates would correlate with higher overall intelligibility scores. However, across all 20 talkers, there was no correlation between overall rate and intelligibility. We also hypothesized that less variance in speaking rate would correlate with better intelligibility. When all talkers were included in this analysis, we found no correlation between rate standard deviation and mean intelligibility. However, when the three talkers with the lowest mean intelligibility score were excluded from the analysis, we found a high negative correlation ($R^2=-.82$) between rate standard deviation and mean intelligibility. Thus, for a subset of talkers, although mean speaking rate does not predict intelligibility, the less the variability in speaking rate the higher the intelligibility. With respect to rate differences for the males and females, we did not find that the females had generally slower rates than the males. This suggests that the sex-based difference in overall intelligibility does not result from a difference in overall speaking rate.

Another global talker characteristic that we investigated as a possible correlate of overall intelligibility was fundamental frequency. Here we hypothesized that both the mean and range of a talker's fundamental frequency might affect his or her overall intelligibility. For the male talkers, we found no correlation between mean fundamental frequency and mean intelligibility score across all 100 sentences. However, for the females, we found a moderate, negative correlation ($R^2=-.32$). Thus, these data provide some suggestion that females with lower mean fundamental frequencies might be more intelligible. With regard to fundamental frequency range, we found a moderate positive correlation ($R^2=0.38$) between F0 range and overall intelligibility for all 20 talkers, indicating that a wider range of pitch variation can enhance sentence intelligibility.

From these investigations of global talker characteristics and overall talker intelligibility, we concluded that the correlations are generally weak to moderate for the twenty, normal talkers in our database. Even though we did find a significantly higher mean intelligibility score for the females than for the males in our database, we were

unable to reliably trace this difference to global talker characteristics, such as overall speaking rate and fundamental frequency characteristics. In light of this result, we turned our attention to some of the indicators of talker variability in pronunciation. Our expectation here was that inter-talker differences at the fine-grained acoustic-phonetic level would correlate with variance in overall intelligibility.

PHONETIC IMPLEMENTATION

We began with an investigation of vowel-space characteristics. Talkers differ in the extent to which they differentiate the vowel categories in the F1 by F2 vowel space [5]. Thus, the compactness of a talker's vowel space is an indicator of the talker's pronunciation characteristics. Since a relatively expanded vowel space indicates less reduced vowels, we hypothesized that a more expanded vowel space would correlate with higher overall intelligibility.

In order to compare vowel spaces across talkers, we selected vowels from the sentence materials that provided an indication of the extremes of each talker's general vowel space. We selected three tokens of each of three point vowels, /i, u, a/. Each token came from a separate sentence, giving us a subset of nine sentences. First and second formant frequencies were measured from the steady-state portion of each of the target vowels for each of the 20 talkers. These measurements were then transformed according to the perceptually motivated mel scale, and plotted in the F1 by F2 mel space. Euclidian areas were then calculated for the triangles formed by the most extreme vowel tokens of each talker's measured vowel space.

Since these vowel space areas are representative of a subset of the total set of 100 sentences, we used the intelligibility scores across this subset of sentences in our analysis of the correlation between vowel space and intelligibility. A rank order correlation between talker vowel space area and overall intelligibility was moderately positive (Spearman rho =+0.36), indicating that across all talkers a more expanded vowel space can lead to higher overall intelligibility. Furthermore, in a comparison of the vowel space area of the male and female talkers, we found that the area within the female vowel spaces tended to be larger than the male vowel space areas (p=.037 by a 1-tail, unpaired t-test). Thus, the results of this analysis of vowel space expansion and overall intelligibility indicate that talkers who have more differentiated vowel articulations tend to be more intelligible. Furthermore, the vowel-space data suggest that the sex-based intelligibility difference might be related to sex-based differences in articulatory precision.

In order to further investigate the pronunciation characteristics that might correlate with talker intelligibility, we performed analyses of the acoustic-phonetic correlates of consistent listener errors. In these analyses we focused on specific portions of sentences that resulted in consistent listener errors, and attempted to find talker pronunciation differences that were responsible for the occurrence of listener errors.

One such case occurred in the phrase "the play seems," which was often mis-transcribed by listeners as "the place seems." In order to investigate the timing characteristics that determined the syllabification of the medial /s/, we measured the durations of the target /s/ as well as of the surrounding segments for each of the 20 talkers. We then examined the correlations between these measurements and the likelihood of correct transcription by the listeners across all talkers. Results of these measurements showed a fairly strong negative correlation (R^2=-0.65) between the duration of the medial /s/ as a proportion of the duration of the preceding word /plej/, and the rate of correct transcription. In other words, the shorter the /s/ relative to the preceding word, the more likely it was to be syllabified by listeners as onset of the following word, rather than as both coda of the preceding word and onset of the following word. Thus, in order to be correctly transcribed, this phrase required a high degree of inter-segment timing accuracy. Furthermore, there were fewer listener errors for the female productions of this phrase, indicating that the female talkers in our database were more accurate in this regard than the males.

Another case of a consistent listener error across all talkers was in the phrase, "the walled town" which was often transcribed as "the wall town." In order to explore the acoustic-phonetic factors that determined whether the word final /d/ was detected, we performed duration measurements on various portions of the target word sequence, "walled town." Results showed a positive rank-order correlation between the absolute vowel-to-vowel duration (i.e., the duration from the offset of the /a/ in "walled" to the onset of the /a^{U}/ in "town") with the likelihood of /d/ detection across all 20 talkers (Spearman rho =+0.702). However, we found an even higher correlation between rate of /d/ detection and the absolute duration of voicing during the /d/ closure (Spearman rho =+0.744). In addition to investigating the correlations of these durations in an absolute sense, we also investigated the correlation between rate of /d/ detection and these durations relative to the surrounding segment and word durations. However, the highest correlation was between absolute duration of voicing during closure and rate of /d/ detection. Since voiced stops in this pre-stop environment are typically not released, the only cue to the presence of a voiced stop is voicing during the closure. And, as demonstrated by the consistent listener error in this example, this normally variable cue can be crucial in this environment. This case is another example of talker-variation at a fine-grained, acoustic-phonetic level that has a direct effect on sentence intelligibility.

In addition to these examples of common listener errors that occurred across all 20 talkers, there were several cases of common listener errors for certain individual talker's productions of particular sentence portions. For these sentences, we compared the acoustic characteristics of the target sentence portion from the talker who was often misheard with those of a talker who received no listener errors on that sentence portion. One such instance occurred for the target phrase " smooth planks," which for one talker was often transcribed as "smooth banks." As compared to a talker whose utterance produced no listener errors on this word, this talker had a reduced /p/ closure duration, as well as a reduced /p/ VOT duration. Thus, for this talker, the cues to the unvoiced consonant were reduced in duration, resulting in listeners perceiving a voiced initial consonant.

In general, our investigations of the acoustic-phonetic correlates of specific listener errors show that variation in talker intelligibility can depend on fine-grained variation in articulation. Sentences spoken by talkers who are more precise in their articulations are more likely to be correctly transcribed.

CONCLUSIONS

The results of this investigation indicate that differences in fine-grained, articulatory-acoustic patterns correlate with variability in overall speech intelligibility. In contrast, global talker characteristics (such as mean fundamental frequency and speaking rate) are not well correlated with differences in talker intelligibility. Furthermore, this study indicated that female speakers, who tend to have more precise articulations, also have higher overall intelligibility scores than males. These findings indicate that talker-specific variations at the acoustic-phonetic level have an impact on both the paralinguistic information carried by the utterance and on its intelligibility.

ACKNOWLEDGMENTS

This work was supported by NIDCD Training Grant DC-00012 and by NIDCD Research Grant DC-00111 to Indiana University in Bloomington, IN.

REFERENCES

[1] Laver, J. & Trudgill, P. (1979), "Phonetic and linguistic markers in speech." In K. Scherer & H. Giles (Eds.), *Social Markers in Speech.* Cambridge, UK:Cambridge Univ. Press.

[2] IEEE (1969), "IEEE recommended practice for speech quality measurements," *IEEE Report No. 297.*

[3] Byrd, D. (1994), "Relations of sex and dialect to reduction," *Speech Communication* vol. 15, pp. 39-54.

[4] Miller, J. (1987), "Rate-dependent processing in speech perception." In A. Ellis (Ed.), *Progress in the Psychology of Language.* Hillsdale, NJ: Erlbaum.

[5] Bond, Z. and Moore, T. (1994), "A note on the acoustic-phonetic characteristics of inadvertently clear speech," *Speech Communication* vol. 14, pp. 325-337.

AN INVESTIGATION OF SINGER PITCH DEVIATION AS A FUNCTION OF PITCH AND DYNAMICS

Perry R. Cook
Center for Computer Research in Music and Acoustics, Stanford, CA, USA

ABSTRACT

Drift and jitter were measured in singer voices, and compared across loudness and pitch, in both vibrato and non-vibrato productions. Jitter showed a slight dependence on dynamic level, and drift showed no clear dependence on dynamic level. Results correlating jitter and drift to produced pitch were more consistent if absolute sung pitch, rather than position within an individual singer's range, was used. Jitter and drift showed a slight dependence on pitch.

VOCAL PITCH DEVIATION

Deviations of pitch in the voice are important perceptual features [1][2]. Some amount of pitch deviation is present in the voice at all times, no matter how much the speaker/singer endeavors to remove it. The intentional quasi-sinusoidal modulation of the fundamental pitch is called *vibrato*, and occurs at a frequency of 5-7 Hz. in trained western BelCanto singing voices. Modulation components at frequencies higher than the vibrato are called *jitter* or *flutter*. Modulation components at frequencies lower than the vibrato rate have commonly been called *wow* or *drift*. The author prefers the terms *drift* and *jitter* because of the negative connotations of *wow* and *flutter* as distortions to be removed if possible. The production of jitter is generally regarded as an involuntary process, caused by random neural firing and a low level feedback mechanism which, in the singing voice, can be trained to cause the periodic oscillation of vibrato [3]. Drift components of very low frequency are directly related to intentional corrections in fundamental pitch. Drift is generally considered to be consciously controllable by means of an auditory feedback loop [4][3][5], but it is not possible to completely remove the drift component at will.

Most synthesis models of singer (and instrument) pitch deviation involve a single sinusoid to model the vibrato, mixed with some random signal to model both the drift and jitter components, such as simple low-pass filtered noise [6]. Maher and Beauchamp [7] proposed a more elaborate model of vocal pitch control, involving one sinusoidal oscillator, three sources of lowpass filtered noise, various summing elements, and a multiplier. The pitch perturbation research covered in this paper was conducted to investigate the behavior of the jitter and drift regions of the pitch signal spectrum as a function of sung pitch and intensity, to formulate a set of rules for pitch deviation control, and to suggest a suitable set of synthesis control parameters.

A STUDY OF SINGER JITTER AND DRIFT

Many past studies of jitter and drift have typically been conducted on tones produced by singers instructed to sing with no vibrato, because the jitter and drift components are easier to isolate and study when vibrato is absent, and many pitch detection methods yield noisy pitch estimates. Signal processing on low amplitude components in the presence of a large vibrato peak is difficult, because the jitter and drift components are often below the noise floor of the pitch detection algorithm itself [8][9]. The Periodic Predictor Pitch Tracker (PPPT) [10][11] has been shown to exhibit a noise floor of less than -55 dB relative to a sinusoidal modulation signal and -30 dB additive noise, and was used to extract the fundamental frequencies in this study. Another method [12] was used to verify the results on a randomly selected 10% of analyzed vocal tones.

Four professional singers were selected for the study, one each of the voice parts soprano, alto, tenor, and bass. The singers were instructed to sing 30 long tones on the vowel /a/ (father). Five notes were performed each at Mezzo Forte (medium loud), Pianissimo (very soft), and Fortissimo (very loud), both with and without vibrato. The singers breathed between each note, and were allowed to repeat any notes which they felt were uncharacteristic of their ability. The frequencies produced were selected for each individual singer to evenly span that singer's comfortable range. The sound files were digitized directly to DAT, digitally transferred to computer disk, down-sampled (-96 dB stop-band rejection filter) to a sampling rate of 5512.5 Hz., then pitch signals were extracted by filtering and sampling at intervals of 55 samples. This 100 Hz. pitch signal sampling rate ensures that modulation information up to 50 Hz. was available for analysis. Once the pitch signals were obtained, Power Spectral Densities (PSDs) were calculated by performing multiple Fourier transforms in 256 point frames on each pitch signal, and averaging the magnitudes. Average and standard deviations were calculated across various groupings of spectra. To aid in generalizing characteristics of levels and rolloffs, a line was fit to the average spectra between 1 Hz. and 4 Hz, and another was fit to the region between 8 Hz. and 32 Hz.

OVERALL RESULTS

Consistent with the study of [6] was that the overall amplitude of jitter decreased with vocal range. That is, high sopranos exhibit less jitter than low basses. In the vibrato case, singers exhibited jitter spectra of about -65 dB (0.97 cents average) at 8 Hz, and rolled off at about 6 dB per octave. In the non-vibrato case, the jitter spectra were about -70 dB (0.55 cents average) at 8 Hz, and exhibited an average 8 dB per octave roll off. The standard deviations were consistently smaller in the drift region than the jitter region. The drift spectrum fell off slowly (roll off of about 1.5 dB / octave) from -50 dB (5.5 cents average) at 1 Hz. out to the vibrato peak at -50 dB average in the vibrato tones, and showed a decrease in the non-vibrato tones to -53 dB at 1 Hz. rolling off at about 2 dB per octave. This decrease implies that singers can hear their voices and control them better in the non-vibrato case than in the vibrato case, and is consistent with the model of drift as a random mechanism with control input from auditory feedback.

Dependance on Loudness

To investigate the dependence of jitter and drift on loudness, the PSD's of all pitch signals at a particular dynamic level were averaged in the vibrato and non-vibrato case. Figures 1 and 2 show plots of the PSD's of the pitch signals of all singers in the vibrato and non-vibrato cases, arranged by dynamic level. The broad dual peak nature of the aggregate vibrato peak shows the variability of vibrato rate between different singers. The average PSD jitter curves show an increase of 4 dB total from pianissimo to fortissimo. No significant change in

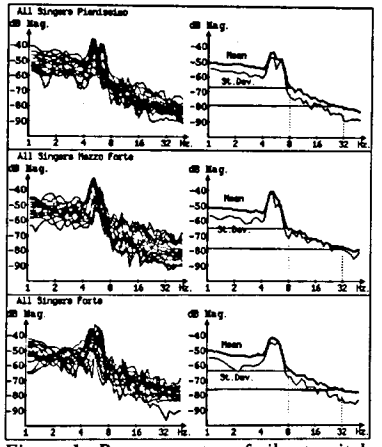


Figure 1. Power spectra of vibrato pitch signals of all singers grouped by dynamic level.

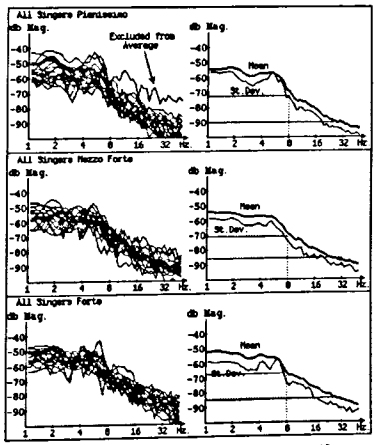


Figure 2. Power Spectra of non-vibrato pitch signals of all singers grouped by dynamic level.

spectral slope was observed, with all vibrato curves exhibiting a 6 dB/octave roll-off, and all curves without vibrato exhibiting an 8 dB/octave roll-off. The drift regions of the spectra showed no clear dependence on dynamic range, implying that the singers in this study could hear themselves and tune well at all dynamic levels.

Dependance on Pitch

To investigate how jitter and drift depend on sung pitch, two sets of spectral averages were formed. The PSD's of all singers at a particular region in their vocal range were averaged in the vibrato and non-vibrato case. Figure 3 shows the plots of the power spectral densities of the pitch signals of all singers for both vibrato and non-vibrato tones, arranged by position within the singer's range. The standard deviations of all of these plots are significantly larger than the mean spectra, indicating that grouping spectra in this way is an unreliable method of classification.

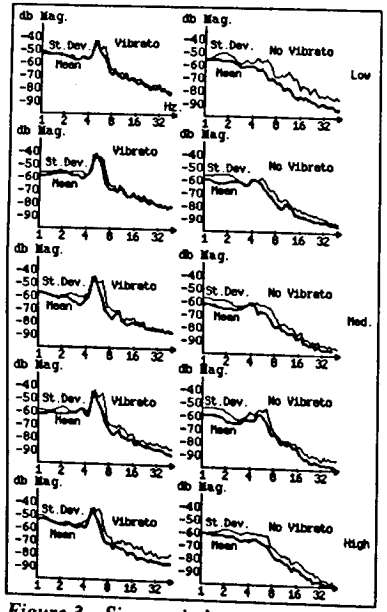


Figure 3. Singer pitch spectra averaged according to position within each singer's range.

Averages were also done within 4 one-octave frequency ranges; 90-179 Hz., 180-359 Hz., 360-719 Hz., and 720-1439 Hz. Figure 4 shows the PSD plots of the pitch signals of the singers for both vibrato and non-vibrato tones, arranged by absolute pitch. The standard deviations for these plots are quite small, indicating that the grouping of spectra by absolute pitch is a more reliable method of classification. The jitter spectra showed a slight dependence on pitch, decreasing 2 dB per octave from low pitch to high pitch. The jitter curves exhibited a consistent slope for all ranges of 8.5 dB per octave in the non-vibrato case and 6 dB per octave in the vibrato case. The drift curves showed a weak dependence on pitch, decreasing about 1 dB per octave of increasing pitch.

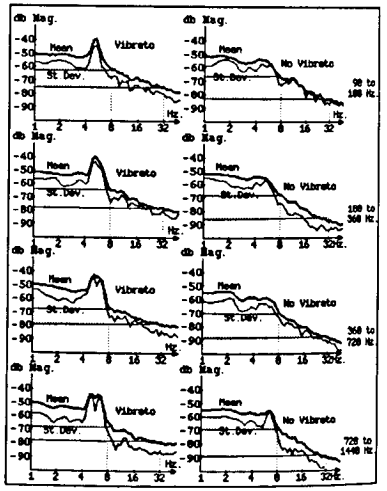


Figure 4. Singer pitch spectra averaged according to absolute pitch in octave bands.

RULES FOR SYNTHESIS

Figures 5 and 6 show the line segment approximations to the jitter and drift spectra, in the vibrato and non-vibrato cases, arranged by pitch and dynamic level. The data indicates that a suitable control space for jitter must allow control over spectral height and slope as a function of dynamic level, phonation pitch, and presence/absence of vibrato.

The minimum jitter is exhibited with no vibrato, at high pitch, and low dynamic level. This jitter is about -70 dB (.55 cents) at 8 Hz., rolling off at 8.5 dB per octave. The maximum jitter is exhibited with vibrato, at low pitch, and high dynamic level. This jitter is about -60 dB (1.7 cents), rolling off at 6 dB per octave. In both the vibrato and non vibrato case, increases in dynamic level account for about 4 dB increase in jitter across the entire dynamic range, and decreases in pitch account for about 2 dB per octave of jitter increase. From the data and the model of drift production, the drift modulation component is most strongly affected by the singer's ability to hear. An extremely simple but nearly complete model of drift is a flat spectrum at -50 dB (5.5 cents) extending to the vibrato peak. The only significant deviations from this model found in this study were in the vibrato/non-vibrato comparison, which indicated a small increase in spectral roll-off in the non-vibrato case.

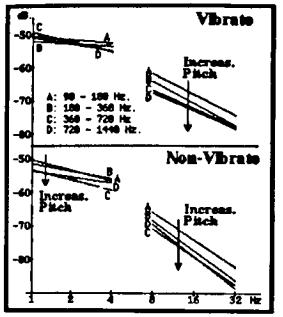


Figure 5. Line segment fits to jitter and drift spectra as function of sung pitch.

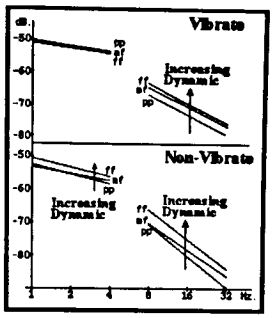


Figure 6. Line segment fits to jitter and drift as function of dynamic level.

REFERENCES

[1] Hillenbrand, J. (1970), "Perception of Aperiodicities in Synthetically Generated Voices," Journal of the Acoustical Society of America, vol. 83, no. 6, pp. 2361-2371.

[2] Kobayashi, T. & H. Sekine (1990), "Statistical Properties of Fluctuation of Pitch Intervals & its Modeling for Natural Synthetic Speech," Proc. of the IEEE Int. Conf. on Acoustics, Speech, & Signal Processing, S6a.8, pp. 321-324.

[3] Shipp, T., J. Sundberg & E. Doherty (1988), "The Effect of Delayed Auditory Feedback on Vocal Vibrato," Journal of Voice, vol. 1, no. 2, pp. 123-141.

[4] Ward, D. & E. Burns (1978), "Singing Without Auditory Feedback," Journal of Research in Singing & Applied Vocal Pedagogy, vol. 1, no. 2 pp. 24-44, 1978.

[5] Elliot, L. & A. Niemoeller (1970), "The Role of Hearing in Controlling Voice Fundamental Frequency," International Audiology, vol. 9, pp. 47-52.

[6] Ternstrom, S. & A. Friberg (1989), "Analysis and Simulation of Small Variations in the Fundamental Frequency of Sustained Vowels," Quarterly Report of the Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, no. 3, pp. 1-14.

[7] Maher, R. & J. Beauchamp (1990), "An Investigation of Vocal Vibrato for Synthesis," Applied Acoustics, vol. 30, pp. 219-245.

[8] Deem, J., W. Manning, J. Knack & J. Matesich (1989), "The Automatic Extraction of Pitch Perturbation Using Microcomputers: Some Methodological Considerations," Journal of Speech & Hearing Research, vol. 32, pp. 689-697.

[9] Hess, W. (1983), *Pitch Determination of Speech Signals.* Berlin: Springer Verlag.

[10] Cook, P. (1991), "Identification of Control Parameters in an Articulatory Vocal Tract Model, With Applications to the Synthesis of Singing," Electrical Engineering Ph.D. Dissertation., Stanford University.

[11] Cook, P., D. Morrill, & J. Smith (1993), "A MIDI Control and Performance System for Brass Instruments," Proc. of the International Computer Music Conference, pp. 130-133.

[12] Serra, X. (1989), "A System for Sound Analysis/Transformation/Synthesis Based on a Deterministic Plus Stochastic Decomposition," Ph.D. Dissertation., Stanford University.

TIMING AND ACCURACY OF FUNDAMENTAL FREQUENCY CHANGES IN SINGING

Guus de Krom and Gerrit Bloothooft
Research Institute for Language and Speech, University of Utrecht, the Netherlands

ABSTRACT

This study deals with relations between the musical score and the acoustic F_0 pattern measured in a sung passage. Four trained singers performed a song in which magnitudes and directions of pitch changes varied systematically. The songs were recorded on the vowels /i/, /a/, and /u/, at three different tempi. Timing differences in the F_0 transitions and deviations from the target values were investigated with a Dynamic Time Warping procedure.

INTRODUCTION

For a singer, the musical score indicates how and when vocal pitch should be varied. However, there exists no one-to-one relation between the prescribed (discrete) note sequence and the F_0 pattern measured in a recorded song, due to for instance F_0 vibrato and inertia of the organic structures involved in phonation. Also, singers have an expressive freedom, which allows them to deviate to some extent from the norm. It may be expected that discrepancies between the pattern of note sequences prescribed in the musical score and the actual F_0 patterns measured in a sung passage depend on the rate at which note sequences are sung (larger deviances at fast tempi), as well as the magnitude of the F_0 difference between successive notes (larger deviances on large intervals).

METHODS

Material

The material used in this study consisted of a song (composed by G. Bloothooft), in which magnitudes and directions of note transitions were systematically varied. The transitions of interest always followed on a particular sequence of "leader" notes (F3-A3 [170-220 Hz] for males, and F4-A4 [340-440 Hz] for females). We chose a fixed leader sequence in order to minimise variations in the immediately preceding context. Nine different transitions were distinguished on the basis of the interval and direction of the steps (see Table 1).

Table 1. Interval magnitude (in semitones) and direction (- downward, + upward) for nine types of transitions.

	1	2	3	4	5	6	7	8	9
interval	-7	-5	-4	-2	+1	+3	+5	+7	+8

Each singer produced nine (legato) versions of the song, using the Dutch vowels /i/, /a/, and /u/, and a slow, medium, and fast tempo (2, 4, and 6 notes per second, respectively).

Recordings and acoustic analyses

The recordings were made in a large, sound treated room. The singers (two males, two females) were standing upright, with the musical score mounted on a stand in front of them. The singers had synthesised piano accompaniment presented over headphones at a comfortable loudness level (simple chords at each measure). A condenser microphone was placed at about 50 cm from the mouth of the singer. The microphone signal was recorded on a DAT recorder.

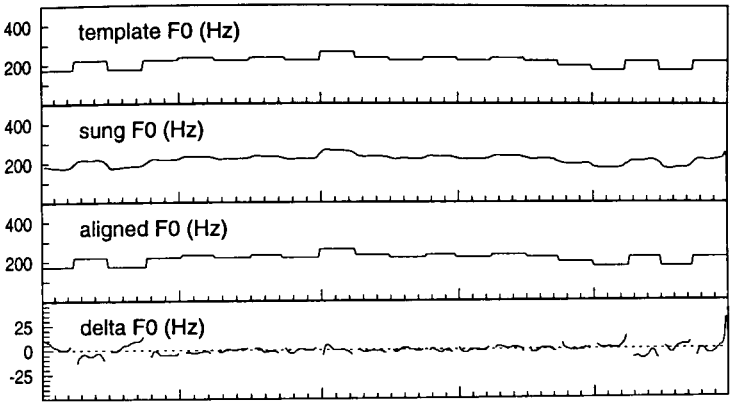


Figure 1. Example of DTW output traces (male singer, medium tempo, duration 5.0 s). From top to bottom, the traces represent the musical score (= template F_0), sung F_0, a trace with the optimal alignment of template F_0 to sung F_0 (aligned F_0), and a trace with the difference between sung F_0 and aligned F_0 (delta F_0).

The DAT recorded songs were transferred to a computer and downsampled to 20 kHz. F_0 was measured with a time-domain algorithm [1], yielding one value (in Hz) for every 10 ms. The musical score was translated into a similar time-by-F_0 format, yielding six template files (3 tempi × 2 versions [for males and females]).

Dynamic Time Warping (DTW)

Relations between the template F_0 traces prescribed in the musical score and sung F_0 traces were investigated by means of a Dynamic Time Warping (DTW) procedure [2]. DTW analyses were performed on the entire lengths of the traces, yielding template F_0, sung F_0, and a DTW-aligned F_0 trace as the optimal match of template F_0 to sung F_0 (see Figure 1).

Parameter definitions

The DTW output files were further processed to obtain measures describing the relative timing of the note transitions and the accuracy of the transitions.

Deviations in the ***timing*** of transitions were investigated by comparing transition instants in the aligned F_0 and template F_0 traces. Three parameters were examined: (1) the transition **lag,** defined as the time difference between the transition instants in the template F_0 and aligned F_0 traces. Negative **lag** values indicate that a transition was made at a later instant than prescribed in the musical score. (2) The absolute value of lag (**|lag|**) yielded another timing parameter. (3) The transition **duration** was defined as the time difference between local maximum or minimum values in the sung F_0 trace just before and after the transition moment.

Parameters indicative of the ***accuracy*** of the transitions were determined by comparing the sung F_0 and aligned F_0 traces. Because these have an identical time basis, a subtraction of sung F_0 from aligned F_0 results in a trace with the instantaneous F_0 deviation in Hz (delta F_0). The local minimum (falling transitions) or maximum (rising transitions) in the sung F_0 trace shortly following the transition moment in the aligned F_0 trace was sought. The difference between this local minimum or

maximum and the prescribed value was calculated and expressed in semitones (**F_0dev,** positive values indicating that the singers' F_0 was too high). A second accuracy parameter was defined as the absolute value of F_0dev (**|F_0dev|**).

RESULTS

Analyses-of-variance were performed with **lag, |lag|, duration, F_0dev, and |F_0dev|** as dependent variables, and interval, tempo, vowel, and the singer's sex as factors (SPSS procedure MANOVA [3]). Because the levels of the interval factor were spaced unequally (see Table 1), polynomial contrasts were applied to these levels. Main effects and two-way interactions were investigated.

Timing parameters

As could be expected, tempo had a significant effect on **lag** ($p < .001$), with values of -276, -80, and -64 ms for the slow, medium, and fast tempo, respectively. Sex ($p = .008$), tempo ($p < .001$), and interval ($p = .039$) had a significant effect on **|lag|**, as did the sex × tempo interaction ($p = .025$). Figure 2 gives the effect of size on **|lag|**.

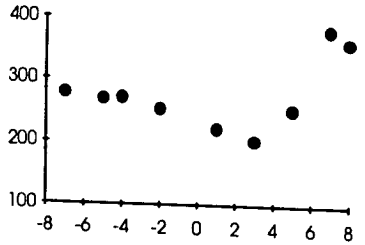


Figure 2. Mean |lag| in ms as a function of the transition interval in semitones.

As Figure 2 shows, |lag| decreased from an interval of -7 semitones to an interval of +3 semitones. |lag| was longest for intervals of +7 and +8 semitones, respectively. Thus, it seems that note steps that involve large (especially upward) F0 transitions give rise to large deviations in the timing of the note transitions.

For both males and females, mean |lag| decreased with increasing tempo. For females, the largest difference was found between the slow tempo on the one hand, and the medium and fast tempi on the other (384 ms, versus 176 and 154 ms). For males, these values were 370, 313, and 208 ms, respectively.

Sex, tempo, and interval had significant effects on the **duration** of the transitions ($p < 0.001$, all factors), with significant interactions for sex × vowel ($p = .010$), sex × tempo ($p = .015$), and interval × tempo ($p = .001$). Figure 3 gives the interval × tempo data.

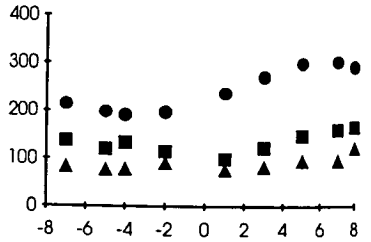


*Figure 3. Mean transition **duration** in ms as a function of the transition interval in semitones. Data are given for slow (circles), medium (squares) and fast tempi (triangles).*

Variations in **duration** related to interval size were most outspoken for the slow tempo. **Duration** was not much influenced by interval size for the medium and fast tempi.

Overall, the **duration** of transitions was shorter for females than for males. The difference was some 30 ms in the slow and medium tempi. No difference was found in the fast tempo.

For the vowels /a/ and /u/, females had a shorter mean transition **duration** than males (135 versus 168, and 130 versus 169 ms). For the vowel /i/, female mean transition duration was slightly longer (154 versus 149 ms).

Accuracy parameters

Interval had a significant effect on mean **F_0dev** ($p < 0.001$). As can be observed in Figure 4, the F0 of the note following the transition was on average too low for downward intervals. A less systematic pattern was found for upward steps, although F_0 was on average slightly too high for intervals between +1 and +5 semitones. These data suggest that the singers exaggerated the prescribed pitch transitions.

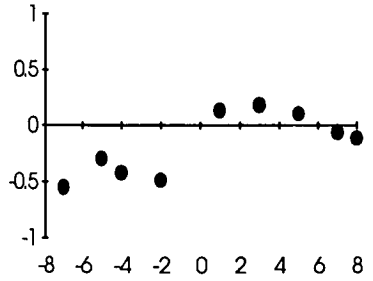


*Figure 4. Mean **F_0dev** in semitones as a function of the transition interval in semitones.*

Interval also had a significant effect on **|F_0dev|** ($p < 0.001$), with a significant (but not systematic) interval × sex interaction ($p < 0.001$). Figure 5 gives **|F_0dev|** data for different intervals.

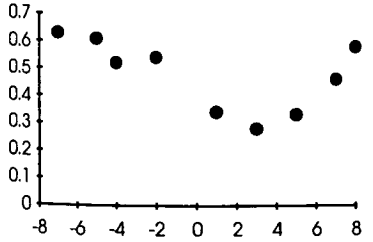


*Figure 5. Mean **|F_0dev|** in semitones as a function of the transition interval in semitones.*

Figure 5 shows that the largest (absolute) deviations in F_0 were found for the largest downward or upward intervals, indicating that these were more difficult to produce than sequences with minor pitch changes. If we compare the **F_0dev** and **|F_0dev|** data, it appears that F_0 of notes following downward transitions was on average too low (negative **F_0dev** values), while F_0 of notes following upward transitions deviated in a more random fashion from the target value (**F_0dev** approximately zero, **|F_0dev|** nonzero).

CONCLUSIONS

We found that the (absolute) interval of note transitions had an influence on the timing of note transitions, as well as the accuracy of the actual F_0 values.

Transition timing was most variable for note sequences that involved large pitch transitions. In the slow tempo, the duration of transitions was longer for upward transitions than for downward transitions.

Systematic F_0 overshoot occurred with downward intervals (more overshoot for larger intervals). Upward intervals resulted in more random F_0 deviations. Tempo and vowel type had no effect on the F_0 accuracy measures.

All four singers reported having difficulties in singing at the fast tempo. Tempo had an effect on the timing data, but not on the accuracy measures. We might therefore tentatively conclude that singers tried to compensate the difficulties encountered in the singing of fast note sequences by adjusting their timing of these transitions.

REFERENCES

[1] Reetz, H. (1989). A fast expert program for pitch extraction. *Proceedings Eurospeech '89*, 476-479.

[2] Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimalization for spoken word recognition. *IEEE ASSP, 26*, 43-49.

[3] Norušis, M.J., (1985). *SPSS-X Advanced Statistics Guide*. McGraw-Hill, New York

AN ACOUSTIC STUDY OF FRENCH VOWELS IN SPEECH AND SINGING VOICE

E. Florig
Institut de Phonétique de Strasbourg - USHS
22, rue Descartes. 67084 Strasbourg. France.
monpiou@ushs.u-strasbg.fr

ABSTRACT

The major question addressed in this study is, to find out to what extent certain French vowels undergo acoustic tranformations as a function of frequency variation from soprano voices. The vowels investigated are /i, e, ɛ, a, y/ and the main thrust of this research is to verify if the nature of spectral modification is comparable for all vowels.

INTRODUCTION

Singing voice, because of its specific demands, its complex articulatory movements, represents a particular object of analysis. In speech, the main objective is to communicate, *i.e.* to render the phonetic message comprehensible. In singing voice, however, this intelligibility does not seem to be the foremost desired goal. The search for precision, the aesthetic aspect, voice compass and quality play a fundamental role, with an aim to conserve a given homogenous quality along the entire voice range. It is obvious that it is more difficult to understand a sung text than a spoken text; the further the singer gets into the high notes, the more problematic phoneme distinction becomes.

The purpose of this study is thus to examine to what extent certain French vowels undergo acoustic transformations as a function of frequency variation.

METHOD

Corpus

As the purpose of the investigation is to look at acoustic differences between vowels in speech and in singing voice, the corpus is composed of two parts:

- the first part for singing voice that is comprised of vocal exercises: a string of the same vowel at different frequencies satisfied the conditions necessary for our study. Ascending vocal exercises were obtained for each of the vowels cited above. Vocal exercises carried out on the same vowel has the advantage of excluding intervocalic consonants, thus avoiding potential consonantal effects on vowel spectra. Each of the vocal exercises started with the cluster /ts/ thus avoiding the problem of sound attack and also optimizing respiration strategies during the vocal exercises (Figure 1).

- the second part for speech: to the vocal exercises are added sentences that contain the target vowels. The vowels appeared in French words, embedded in the carrier sentences. Vowel context was varied using one of the following consonants /b, v, z, l, ʒ, ʁ/ that vary constriction location. Unvoiced consonants were deliberately excluded to avoid possible vowel devoicing, so also were nasals to avoid vowel nasalizing. The sentences were constructed as follows:

The carrier sentence "Je vais chanter sur /i/, comme dans bise, vie, Suzie, lit, magie et Paris." means "I am going to sing on /i/, as in kiss, life, Suzie, etc.".
"Je vais chanter sur /e/, comme dans bébé, privé, rusé, blé, léger et paré."
"Je vais chanter sur /ɛ/ comme dans bête, vert, zèbre, laide, geste et raide."
"Je vais chanter sur /a/, comme dans bas, vase, visage, lame, jade et rat."
"Je vais chanter sur /y/, comme dans butte, vue, zut, lutte, juste et rustre."

Measurement zones were restricted to the mid-portion of the vowels, thus avoiding transitions due to adjacent contexts.

Recordings

Recordings were carried out using a DAT TDC-D3 recorder and a Neumann K54 microphone. The recordings were made in a sound proof anechoic room for two lyrique sopranos from the Conservatoire National of Strasbourg, digitized and analyzed by software. Analyses were carried out on spectrogrammes, using wideband and narrow band filters. Wideband spectra were used for formant measurements whilst narrow band spectra served for measurements of sound pitch.

Extracts of the corpus were submitted to a dozen subjects for identification. This was not a perception test, rather it served as a control test where subjects had to note the quality perceived and also to say if the vowel was recognizable or not. Such a task would served as an indicator that would confirm or invalidate our results.

RESULTS

Speech

Measurements were carried out on 31 /i/, 28 /e/, 17 /ɛ/, 24 /a/ and 17 /y/ for each speaker. Two values were obtained from two points on each vowel. Typical values were then calculated as the means of measured values (*cf.* Table 1.). These values served as reference values in our analyses of singing voice, since the influence of frequency variation on the acoustic composition of vowels in singing voice, will be determined in relation to the acoustic properties of the same vowels in speech.

Singing voice

10 vocal exercises on /i/, 8 on /e/, 7 on /ɛ/, 9 on /a/ and 10 on /y/ were analyzed for each speaker. This amounts to 90 notes for /i/, 72 for /e/, 63 for /ɛ/, 81 for /a/ and 90 for /y/. Two measurement points were retained, here also, on each vocalic portion.

Determining thresholds

The acoustic composition of the different vowels in singing voice was not clear along the entire frequency range of vocal exercises. Actually, as from certain high notes, it becomes very difficult, let impossible, to distinguish formants in a precise manner. Also, it was observed that in these high frequencies, subjects could no longer recognize the different vowels. Another observation was that the first hamornic and fondamental frequency were rapidly above the first reinforcement zone that had been detected for the same vowel in speech. It is thus Fo that will be reinforced to play the F1 role. However, the Fo value will also rapidly attain a level so high, such that the different harmonics will not coincide with reinforcment zones characteristic of vowels. Formant structure of vowels in the singing voice context will therefore no longer ressemble that of vowels in speech. Thresholds were established, beneath which the formant structure of the vowel is relatively close to a speech-like vowel and above which, formant detection becomes difficult as values no longer correspond to speech reference values (*cf.* Table 2). – [5] refers to this phenomenon as "intelligibility thresholds", since vowel intelligibility depends on the distinctiveness of its formants. A mean formant value was obtained for all vowels on each note of the vocal exercises. These formant values allowed us to establish thresholds, by comparing obtained values with reference values for speech and by controlling our results with observations that had been made during the audition test.

It was also noticed that in zones where acoustic patterns were relatively unclear, formant structures were close to those of speech, on certain notes. This, presumably, is due to Fo frequency that allows harmonics to coincide with reinforcement zones, charactersitic of vowels.

The presence of a supplementary reinforcement zone was also detected, the "Singing Formant", as described by [7]. This formant was located, for the two female speakers, around 360 Hz, regardless of the vowel analyzed

Quality variation

Vowel quallity was also compared across the singing and the speech conditions, using mean values (*cf.*

Figure 2). Two values were determined: the first in an average pitch, not very far from speech and the second, in a higher pitch.

Notice the difference in formant values between speech and singing voice for /i/, regardless of the note on which it was produced: F1 increases slightly — due to the combined effect of pitch and a slight aperture increase — whereas F2 decreases — due certainly to a slight lip rounding (*sic*). Actually, around 20% of listeners hesitated in recognizing /i/ and /y/ in the high frequencies.

For vowel /e/, the strategy is to increase both F1 and F2 thus attaining values for /i/ in speech. The two vowels are confused by 70% of the listeners in the high frequencies.

Vowel /ɛ/ has formant values different from those in speech as from the onset of low frequencies, because F1 is immediately superior while F2 starts descending. It seems that this vowel was slightly rounded since some listeners perceived a timbre close to the rounded vowel /œ/.

F1 for /a/ remains stable for quite a while — up to around D 4 — with an F2 lower than in speech. This vowel is seemingly rounded in singing voice and is perceived, in the high frequencies, as /ɑ/ or /ɔ/ by 86% of listeners.

For /y/ the tendency is to increase F1 — provoked certainly by pitch increase and by jaw lowering —, while F2 starts lowering. Apparently, jaw lowering, that is responsible for frequency increase, also causes the unrounding of the vowel.

CONCLUSION

It is clear that frequency variations provoked modifications in the acoustic structure of sung vowels. It is possible to determine thresholds or progressive steps beyond which formant distinction and vowel recognition becomes difficult. Moreover, it was noticed that vowel quality was somewhat different across conditions. This is a general tendency that is manifest on all vocal exercises. All vowels are characterized by an increase in F1, probably due to the conjugated effect of pitch increase and jaw lowering. Vowels /i, ɛ, a/ seem to undergo slight rounding while vowel /y/ seems to lose some of its rounding feature.

REFERENCES

[1] Cornut, G., Lafon, J.-C. (1960) "Etude acoustique comparative des phonèmes vocaliques de la voix parlée et chantée", *Folia Phoniatrica*, vol. 12, n°3, pp. 188-196.

[2] Harmegnies, B., Landercy A. (1993) "Analyse spectrale et voix chantée. Contribution à une métrologie objective de la maîtrise du chant", *Revue de phonétique appliquée*, n°106, pp. 35-59.

[3] Leothaud, P. (1991) "L'acoustique vocale", *La voix dévoilée*, Paris : Romillat, pp. 71-92.

[4] Miller, R. (1990) *La structure du chant*, traduction de Gouëlou J.-M., Paris : IPMC.

[5] Scotto Di Carlo, N. (1976) "Etude acoustique et auditive des facteurs d'intelligibilité de la voix chantée", *Travaux de l'Institut de Phonétique d'Aix*, vol. 1, pp. 115-129.

[6] Scotto Di Carlo, N. (1991) "La voix chantée", *La Recherche*, n°235, vol. 22, pp. 1016-1025.

[7] Sundberg, J. (1973) "The acoustics of the singing voice", *Scientific American*, vol. 236, n°3, pp.

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

[9] Sundberg, J. (1991) "Phonatory vibrations in singers. A critical Review", *STL-QPSR*, vol. 1, Stockholm.

[10] Sundberg, J., Johansson, C., Wilbrand, H. (1982) "X-Ray Study of Articulation and Formant frequencies in two female Singers", *STL-QPSR*, vol. 4, Stockholm.

Table 1. Reference values for formants in speech given in Hz.

	F1	F2	F3
[i]	334	2627	3295
[e]	378	2317	3136
[ɛ]	635	2109	3080
[a]	732	1727	2980
[y]	329	2021	2688

Table 2. Frequency thresholds

	distinct acoustic composition	imprecise acoustic composition
[i]	up to A sharp 3	from A sharp 3 to E4
[y]	up to B3 - C4	from B3 - C4 to D4 - D sharp 4
[e]	up to G sharp 3 - A3	from A sharp 3 to D4
[ɛ]	up to A3	from A3 to C sharp 4
[a]	up to C sharp 4 - D4	from C sharp 4 - D4 to E4

Figure 1. Example of vocal exercises

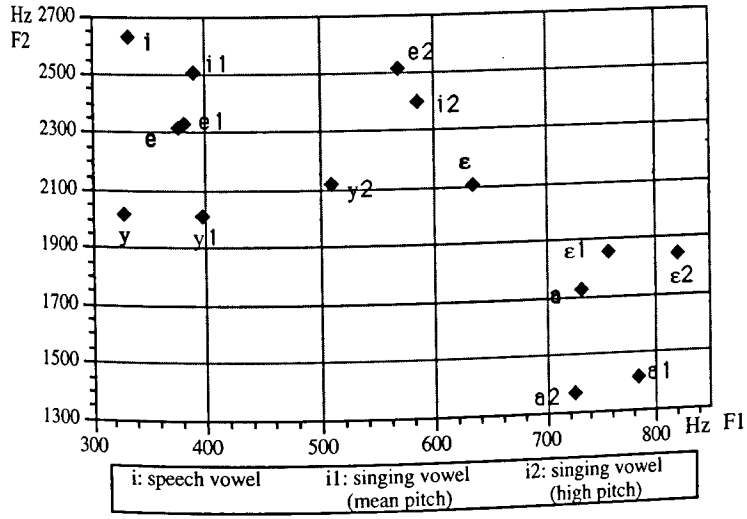


Figure 2. Formant values for speech and for singing voice.

VERTICAL AND SAGITTAL POSITION OF LARYNX IN SINGING

Pertti Hurme[1] and Aatto Sonninen[2]
Department of Communication, University of Jyväskylä, Jyväskylä, Finland[1]
Sibelius Academy, Helsinki, Finland[2]

ABSTRACT

Teachers of singing recommend a comfortably low position of the larynx in singing. Some studies have corroborated a low vertical larynx position (VLP) in singers, whereas others have observed that the larynx rises with pitch. In this study, VLP was measured from roentgenograms of singers producing a rising pitch series. The roentgenological method permits the measurement of several other variables in addition to VLP, such as the sagittal (anterior–posterior) movement and position (SLP) of the larynx when singing.

LARYNX POSITION

Several factors affect larynx position: anatomical differences, vital functions such as breathing and swallowing, habitual position and movements during speech and singing. It is well known that larynx position has an effect on vocal tract resonances and on the biomechanical properties of the vocal folds [1, 2].

Several studies on professional singers have concluded that during singing the larynx is in a low position irrespective of pitch. For instance, in a roentgenological study of singers of the Bolshoi ballet Dmitriev [3] observed relatively little variation in larynx height (VLP) as a function of pitch. He also observed a connection between larynx height (related to the cervical spine) and voice type. The observations of Shipp [4] are similar. On the other hand, Johansson, Sundberg and Wilbrand [5] and Pabst and Sundberg [6] report that the larynx rises with pitch at least in some subjects, especially at higher pitches. Measurements pertaining to this question will be presented in this paper.

Methodical limitations often confine the study of larynx movements to VLP measurements. However, it is well-known that the larynx can move in an anterior–posterior (sagittal) direction as well. In this paper we focus on biomechanical factors, the exterior forces affecting the larynx, which is a relatively elastic structure, mainly consisting of cartilages, muscles and connective tissue. Laryngeal joints are not rigidly hinged, but allow gliding in addition to rotation, e.g. in the cricothyroid joint [7]. We study the vertical (superior–inferior) and sagittal (anterior–posterior) position of the larynx in relation to the cervical spine and to the mandible; our subjects are singers producing a series of vowels spanning the musical range.

PROCEDURE

This study is a reanalysis of a data corpus collected by Aatto Sonninen [8]. The corpus consists of 12 singers (9 females [sopranos and mezzo-sopranos] and 3 males [tenors and baritones]) as well as a number of nonsingers (not reported here; see [9]. The singers were of high national or international level. Lateral spot roentgenograms were taken as the subjects sustained the vowel /a/ on an ascending scale. The distance of the posterior superior part of the cricoid cartilage (point d) was measured along x and y coordinates defined by a vertical line connecting the 2nd and 6th cervical vertebra (dy, vertical larynx position or VLP) and a horizontal line at the 6th cervical vertebra perpendicular to the vertical line (dx, sagittal larynx position, SLP). In addition, we measured the distance between the anterior–inferior part of the thyroid cartilage (point C), the anterior–inferior part of the hyoid bone (point B) and the mandible (point A; placed as anterior as possible, in each subject at a fixed distance from the 2nd cervical vertebra); the measurement points can be seen in Figure 3.

RESULTS

There was considerable interindividual variation in the vertical and sagittal measurements of singing a rising pitch series (see Figure 1): the medians for the subjects' VLP varied up to 40 mm and the medians of SLP up to 15 mm. Figure 1 also shows that the subjects differed in VLP as compared to the zero point (the 6th cervical vertebra): in some subjects VLP is clearly above zero (e.g. MV, JH), in some at or below zero (e.g. HN, AK). The subjects also differed in SLP: MH in particular has

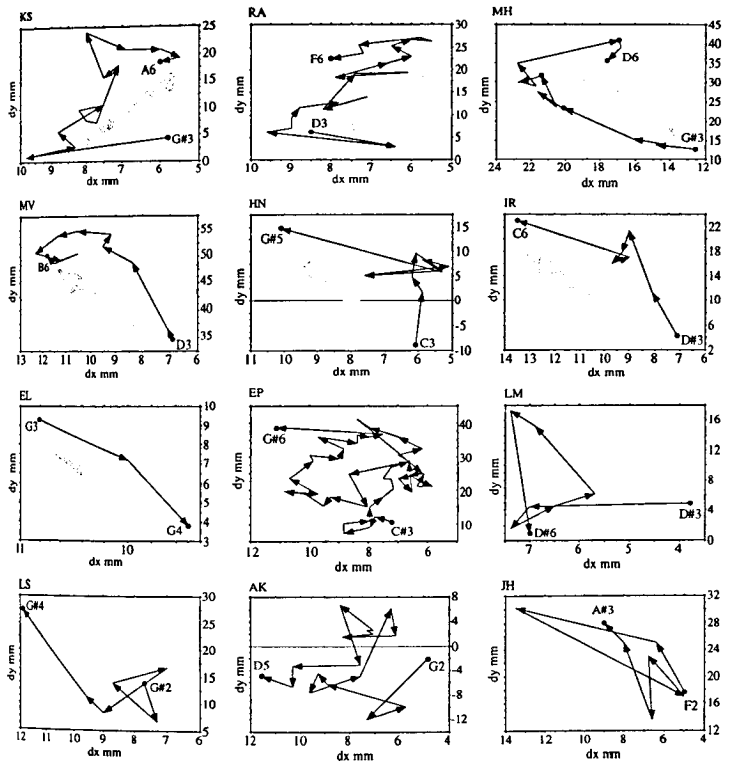


Figure 1. Vertical (dy) and sagittal (dx) position of larynx in female (three top rows) and male (bottom row) singers in a rising pitch series. Arrows indicate general direction of movement.

more anterior values than the others.

Figure 1 also shows the movement of the position of VLP and SLP as the subjects produced sustained phonation at various pitches (shown in the figure). The three top rows describe female singers, the bottom row male singers. The movements of the posterior superior part of the cricoid cartilage (point d) for female and male singers appear not to differ in any systematic way. The over-all movements can be described by 4 patterns (indicated with arrows in the figure): (1) movement in the posterior–superior direction (subjects KS, RA), (2) movement in the posterior–inferior direction (subject EL), (3) movement in the anterior–superior direction (subjects MV, HN, MH, IR, LS) and (4) complex zig-zagging movement (subjects AK, LM, JH, EP).

The results of the measurements of the distances between the thyroid cartilage, the hyoid bone and the mandible are not described in detail here. However, by means of selected examples (schematicized from the roentgenograms) Figure 2 shows the relation of the thyroid cartilage and the hyoid bone to each other. Case A is a textbook case. The other examples show that the hyoid bone and the thyroid cartilage can assume a wide variety of positions in relation to each other. Case F is very extreme: in relation to the thyroid cartilage the hyoid is very anterior and inferior. Case F (subject IR) is shown in more detail in Figure 3, showing the production of a vowel at D#3, D#4, D#5 and C6. With increasing pitch the hyoid bone moves in an anterior and inferior position: however, it can also be seen that the thyroid cartilage moves in a superior and slightly anterior direction. The extreme position is accomplished by moving both the hyoid bone (and the mandible) and the thyroid cartilage.

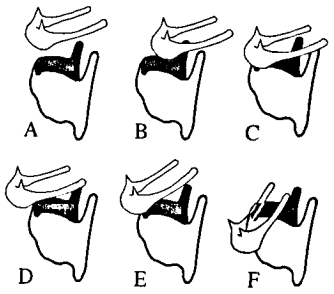


Figure 2. Various observed positions of hyoid bone and thyroid cartilage.

DISCUSSION

VLP and SLP in singers

Some voice pedagogues claim that in trained singers the larynx is not raised with pitch, implying that individuals without training or with insufficient training raise their larynx with pitch. In our subjects we see much variation between individuals: we have evidence both for larynx-raising with pitch and for nonraising (or even lowering) at high pitches. However, our data show that the larynx does not move only along one dimension, up/down. There is another dimension, forward and backward. Figure 1 shows that singers use 4 strategies in positioning the larynx when singing an ascending pitch series. One of these strategies is mixed; the others can be described as a result of three forces in competition, pulling the larynx in an anterior–superior (up and forward) direction, in a posterior–superior direction (up and backward) and in an inferior direction (down). If these forces acting on the larynx follow the principle of motor equivalence [10], each contributing to achieve a common goal, good singing results. Larynx position is determined by these three forces within the larger context of the singer's body posture and artistic expression.

VLP: Supported vs. Unsupported

We have recently published data on VLP and a number of other variables in singers producing supported and unsupported singing on a variety of tasks (low and high pitch, messa di voce, etc; see [11]. Our data showed that VLP was generally lower in supported voice as compared to unsupported voice. VLP was invariably lower in supported voice in 4 subjects out of 9, lower in the majority of cases in 4 subjects and higher in 4 tasks out of 5 in one subject. In all, VLP was lower in supported voice than in unsupported voice in 32 cases, about the same in 2 cases and higher in 11 (n=45: 9 subjects and 5 tasks). Thus, when asked to sing with support and without support, one of the means by which to differentiate voices is VLP.

Hyoid Bone

Our measurements of the position of the hyoid bone in relation to the thyroid cartilage and mandible show that the textbook conception (graph A in Figure 2) is limited and overly simplified. Real constellations of these structures show variation even to a surprising degree. From a biomechanical point of view such diverse behavior on the part of these structures is motivated: the extreme constellations guarantee the vital function of air flow when singing at extreme pitches (which require extreme maneuvers in the laryngeal region). A comparison of the laryngeal behavior of subject IR in Figures 1-3 shows that drastic measures have been to taken to secure air flow: Figure 1 shows that IR's larynx moves in an anterior–superior direction when singing an ascending pitch series, and graph F (=IR) in Figure 2 (which is a simplification of the rightmost drawing in Figure 3) shows that the hyoid bone is in an extremely anterior–inferior position when singing at C6. Thus, during singing

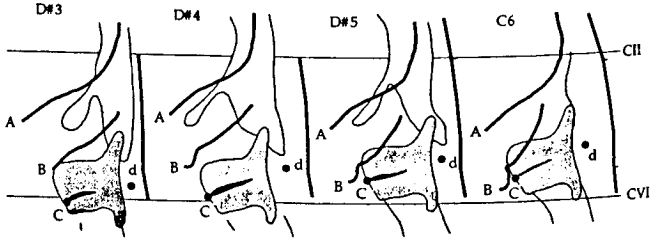


Figure 3. Schematicized roentgenograms of subject IR singing at various pitches.

the larynx can sometimes rise considerably, but it is still possible to sing (the air flow is not obstructed). This is apparently accomplished by the controlled balance of the three forces described above.

Methodical Considerations

The measurement procedure used in this study compares favorably with the Twin-Channel Electroglottograph of Rothenberg [6, 11]. Our roentgenological measurements give data that cannot be obtained by means of the Rothenberg method: we obtained data on the SLP in addition to the VLP. We have shown that there is more variation in VLP (up to 40 mm) than the amount that can be registered by the Rothenberg method (maximum 20-25 mm). Our method allows for inter-individual comparison, whereas the Rothenberg device needs individual calibration and thus does not really allow for interindividual comparison. On the other hand, it is true that the use of X-rays is limited and potentially dangerous, whereas the Rothenberg system is noninvasive (but does not work very well on fat necks). The Rothenberg system is also very accurate in registering time-varying data, whereas roentgenological measurements are necessarily more limited in the time domain.

CONCLUSION

In our data the singers exhibit very varied larynx positions when singing an ascending pitch series. Some singers raise the larynx with pitch, others do not – and even more complex patterns occur. In our opinion, many kinds of laryngeal maneuvers (including larynx raising) are possible when singing, as long as the forces affecting the vocal folds are kept in balance. Larynx position may be connected with the chest–falsetto transition in singing (to be discussed in a forthcoming article by the present authors and Erkki Vilkman).

It is inadequate to describe the position of the larynx by VLP alone. In addition to superior-inferior movement, the larynx also moves in an anterior–posterior direction. What is needed is an understanding of such complicated movements: to know when they are harmful, and when they are necessary for achieving a certain goal (in pitch or vocal quality).

Laryngeal tension – which in itself is a vague concept – is often regarded as a cause of poor performance [12]. To relieve such tension, it would be desirable to understand the forces operating on the larynx, both internal and external. It may be useful to massage the neck area in general, but a more detailed analysis of the contribution of external muscles to laryngeal movements would help in getting to the root of voice problems.

References

[1] Fink, B. (1975), *The human larynx.* New York: Raven.

[2] Sundberg, J. (1987), Vertical larynx position – Research findings and their relationship to singing (discussion). *Journal of Voice*, vol. 2, pp. 220-222.

[3] Dmitriev, L. (1962), *Golosoobrasovanie u pevtsov.* Moscow:

[4] Shipp, T. (1987), Vertical laryngeal position: Research findings and application for singers. *Journal of Voice*, vol. 1, pp. 217-222.

[5] Johansson, C., Sundberg, J., Wilbrand, H. (1985), X-ray study of articulation and formant frequencies in two female singers. In: *SMAC 83. Proceedings of the Stockholm International Music Acoustics Conference*, vol. 1, pp. 203-218.

[6] Pabst, F. & J. Sundberg (1992), Tracking multi-channel electroglottograph measurement of larynx height in singers. *QPSR* (RIT, Stockholm), 2-3/1992, pp. 67-78.

[7] Vilkman, E. (1987), *Studies on human voice production.* Acta universitatis Tamperensis: A232. Tampere.

[8] Sonninen, A. (1956), The role of the external laryngeal muscles in length-adjustment of the vocal cords in singing. *Acta Oto-laryngologica* (suppl. 130).

[9] Sonninen, A., Hurme, P., Vilkman, E. (1992), Roentgenological observations on vocal fold length-changes with special reference to register transition and open/covered voice. *Scandinavian Journal of Logopedics and Phoniatrics*, vol. 17, pp. 95-106.

[10] Abbs, J., Cole, K. (1986), Neural mechanisms of motor equivalence and goal achievement. *Speech Motor Control Laboratory Preprints*, University of Wisconsin, pp. 1-29.

[11] Sonninen, A., Hurme, P., Sundberg, J. (1994), Physiological and acoustic observations of support in singing. In: *SMAC 93, Proceedings of the Stockholm Music Acoustics Conference*, pp. 254-258. Stockholm.

[12] Hülse, M. (1991), Zervikale Dysphonie. *Folia phoniatrica*, vol. 43, pp. 181-196.

SPECTRAL MEASUREMENT OF VOICE QUALITY IN OPERA SINGERS : THE CASE OF GRUBEROVA

Tom Johnstone and Klaus Scherer
University of Geneva

ABSTRACT

Excerpts from different recordings of the cadenza in Ardi gli incensi from Donizetti's opera Lucia di Lammermoor as sung by Gruberova are acoustically analyzed to determine the nature of higher frequency energy and higher formant structure (in particular the presence or absence of a singer's formant) particular to this singer. In light of the results, the role played by these acoustic features in the expression of emotion in opera singing is discussed.

INTRODUCTION

In a recent study, Siegwart and Scherer (1995) acoustically analyzed two excerpts from the cadenza in Ardi gli incensi from Donizetti's opera Lucia di Lammermoor as sung by five famous sopranos (del Monte, Callas, Scotto, Sutherland, Gruberova). The acoustic parameters measured were correlated with preference and emotional expression judgments, based on pairwise comparisons, made by a group of experienced listener-judges. In addition to showing major differences in the voice quality of the five dive studied, the acoustic parameters suggested which vocal cues affect listener judgments. Two component scores, based on a dimensional analysis of the acoustic parameters, predicted 84% of the variance in the preference ratings.

The results showed that Gruberova's rendering of the cadenza was generally preferred and thought to express more "tender passion" and "sadness" than the other singers by the judges in this study. Acoustically, Gruberova's voice was markedly different from all the other singers, showing lower energy in a predicted singer's formant band and stronger high frequency energy components in the spectrum. Her voice was like that of Sutherland and Callas in being characterized by high energy in the F0 band and little variation in lower frequency peaks.

Reviewers of the above study raised the possibility that the strong higher frequency energy measured for Gruberova (which might have affected the judges ratings) could be due to the sound engineers' selective boosting of specific frequency bands. To study this possibility, we recorded the two excerpts of the cadenza live in Gruberova's dressing room before a performance of Lucia at the Zurich opera, having obtained the artist's cooperation for this study. In this paper we compare the acoustic analyses for this recording with several professional recordings - a new CD recording and a radio broadcast from a concert hall, in addition to the cassette recording used in the earlier study. The recording of Sutherland (1971 CD) from the original study is also included for comparative purposes.

This paper reports the acoustic results and reviews the role of the higher frequency spectral energy bands and the singer's formant for the expression of emotion in speech and singing.

METHOD

Gruberova's rendering of the two lines of the Lucia cadenza studied in this research were recorded in her dressing room using a Sony TCD-D3 DAT recorder. The prerecorded samples consisted of a 1984 EMI cassette, a 1992 Teldec CD, and a recent live radio recording. The recordings were digitized using a Kay CSL 4300B speech station at 20kHz sampling rate. An optimal recording level was chosen for each sound recording. This did not affect the subsequent analyses, as the spectral measurements of intensity were all normalized with respect to the total intensity of each recording.

RESULTS

Analysis of the digitized recordings paralleled that used in the original study [1]. A 128 point long term average spectrum was calculated for the full duration of each of the digitized recordings. This spectrum was used to calculate the relative amount of energy present in the two frequency bands measured in the original study (Table 1, rows 1 and 2). With the exception of the CD recording, the energy in the high frequency band (3500 to 10000 Hz.) is higher in the Gruberova recordings than in the recording of Sutherland (Figure 1).

The CSL LPC formant tracking program was then applied to each recording, yielding means and standard deviations of the frequencies of tracked formants (Table 1, rows 4 to 7). The figures for the fourth tracked formant are included for purposes of comparison with the data from the original study. The figures pertaining to this formant should be regarded with some caution however, as the formant was not consistently identified by the tracking algorithm and the number of valid samples varied substantially between the different recordings.

The formant figures also show general agreement with the original study. Specifically, the recordings of Gruberova consistently show a third formant located at a higher frequency than that of Sutherland. The recordings of Gruberova also have a higher standard deviation of the fourth formant than the Sutherland recording, as in [1]. Contrary to the previous study however, there was no significant difference between the two singers in the mean positions of the fourth formant.

An examination of the relationship between the different measured parameters can help determine the nature of the measured high frequency energy. Correlations were calculated between the two formant frequencies and the energy in the first two frequency bands. It was found that the frequency of the third peak correlates strongly with the amount of energy in the spectrum above 3500 Hz (Pearson's $r = 0.97$). Examination of the long term average spectrum for the five recordings indicates that the amplitude decreases sharply above 4500 Hz., indicating that energy in this region does not contribute substantially to the high frequency band measured. Thus it would seem that the large amount of high frequency energy in the Gruberova recordings is due to the higher frequency position of the third spectral peak. This was confirmed by measuring the energy in the frequency range from 3500 to 4500 Hz. (Table 1, row 3). All the Gruberova recordings were characterized by more energy in this band than the recording of Sutherland. Importantly, Sutherland's (and the other *Dive* studied in [1]) third formant is located under the 3500 Hz. cutoff for the two measured high frequency bands.

An examination of spectrograms can be used to better understand the nature of the formant structure in the different recordings. Figure 2 shows the spectrograms for the live recording of Gruberova and the recording of Sutherland (only one spectrogram of Gruberova is displayed here, although spectrograms of the other Gruberova recordings were very similar, showing concentrations of energy in the same regions). The spectrograms reveal quite a different spectral energy distribution for Gruberova as compared to Sutherland. Gruberova shows a concentration of energy in two closely spaced bands between 2900 and 4100 Hz., with relatively low energy in the 1500 to 2500 Hz. range. In comparison, Sutherland shows a more constant spectral slope, with three bands or reducing energy located between 1500 and 3900 Hz. It is also apparent that the automatic formant tracking program was not able to distinguish between the third and fourth formants of both singers, thus compounding the two into one measured formant track (the third formant as given in Table 1, rows 4 and 5).

DISCUSSION

The original purpose of this study was to determine whether or not the presence of more high frequency energy and higher frequency peaks in recordings of Gruberova was due to recording artifacts or manipulation by sound engineers. By analyzing three new recordings, including one taken directly in the singer's dressing room, it has been shown that the high frequency energy is indeed a characteristic of Gruberova's singing itself. More specifically, long term spectra of all Gruberova recordings displayed a high energy region between 2900 and 4100 Hz. This region appears due to the clustering of the third and fourth formants. In contrast, the recording of Sutherland lacks such a high energy region and the formants appear at lower frequencies.

An explanation of the high energy region in the recordings of Gruberova might be the presence of a singer's for-

Table 1. Acoustic analysis results for the five recordings . Rows 1-3 are given in dB, normalized with respect to overall intensity. Rows 3-7 are given in Hertz.

	Gruberova				Sutherland
	Cassette	Live	CD	Radio	CD
Singers Formant	-36	-46	-32	-36	-36
3500-10000 Hz.	-28	-32	-36	-27	-36
3500-4500 Hz.	-29	-32	-32	-27	-38
Peak 3 Mean	3779	3618	3546	3797	3443
Peak 3 Std. Dev.	236	399	321	229	355
Peak 4 Mean	4982	4826	5427	4934	5320
Peak 4 Std. Dev.	1640	1253	1981	1427	403

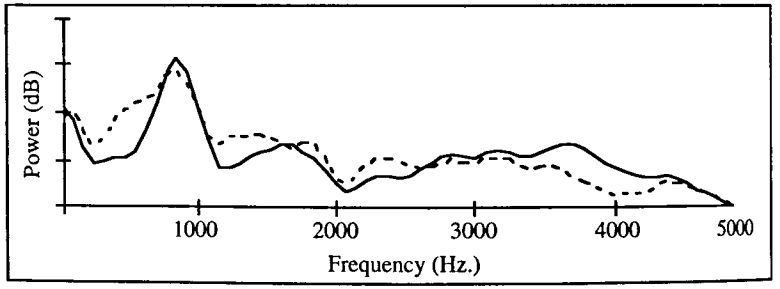


Figure 1. Average spectra for Sutherland (broken line) and live recording of Gruberova (solid line). The spectra are normailized to overall intensity.

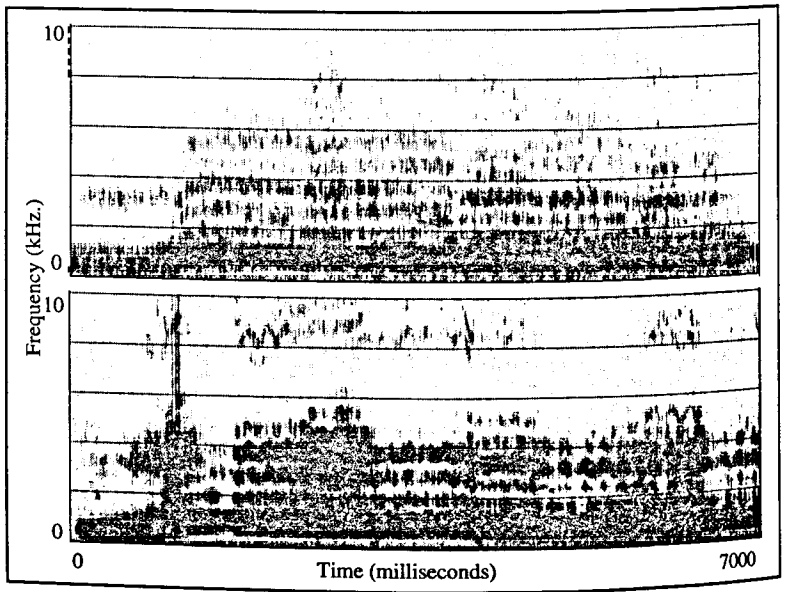


Figure 2. Normalized spectrograms of Sutherland (top) and Gruberova (bottom).

mant centred at about 3600 Hz. As discussed in [2], the singer's formant is not a single formant as such, but rather a clustering of formants around a predicted frequency of about 3000 Hz (in sopranos). When clustered sufficiently closely, individual formants tend to reinforce each other, leading to a spectral region with increased overall resonance. In the case of soprano singing, the partials are spaced widely apart, which makes the exact positions of the formants relative to the partials crucial.

Whilst most sopranos may be able to vary the formants to follow the positions of the harmonics, the way in which this is done may vary between singers. Thus some singers might raise the fourth formant in order to make it coincide with a harmonic, thus separating it from the lower formants, which typically might drop ([2], pp. 125-129). Such a separation of third and fourth formants, which would prevent the development of a singer's formant, would seem to be the case with Sutherland. In the recordings of Gruberova, however, the fourth formant drops along with the third formant, thus maintaining a close distance and allowing the formants to reinforce. The presence of a singer's formant will not necessarily ensure high energy in that region of the spectrum; the spectral drop-off of the harmonics must also be sufficiently gradual.

The question of whether sopranos possess a singers formant has been discussed recently by Berndtsson and Sundberg [3]. Berndtsson and Sundberg compared the classification by trained judges of synthesized soprano voices for various manipulated singer's formant positions. Also included in the study was one recording resynthesized using the formant positions from a professional soprano. The study found that perceived quality of the synthesized voices increased as the centre frequency of the singer's formant increased. The recording resynthesized from the professional soprano's formant positions was, however, judged as natural as the best of the synthesized recordings, despite its lack of a singer's formant.

The strong correlation found in [1] between the proportion of energy above 3500 Hz. and judgments of emotional expressivity may well be due to the presence of a singer's formant at about 3600Hz. This would fit well with the results using synthesized recordings in [3]. As that study only examined singer's formant positions up to 3500 Hertz, it was unclear whether even higher positions for the singer's formant might be judged even better. This study indicates that perceived quality and expressivity might continue to increase with an even higher singer's formant. The finding in [3] that the resynthesized recording with no singer's formant was judged to be as natural as the synthesized recordings might have been due to the more natural formant spacing, rather than the lack of a singer's formant per se. The relative positions of the formants in relation to the harmonic structure might be crucial to perceived quality, and thus the synthesized recordings using formant spacing taken from baritones might have suffered from their somewhat arbitrary relative formant positions. As admitted by the authors of that study, none of the recordings in their study were judged as being particularly natural.

Many of the conclusions drawn here concerning higher frequency spectral regions and the formant structure of sopranos remains speculative. In particular, the temporal changes to these features have not been examined. It is clear that much further empirical research is required in order to better understand the processes involved in emotional expression in singing.

ACKNOWLEDGMENT

The authors wish to express their gratitude to Editha Gruberova for her cooperation in this study.

REFERENCES

[1] Siegwart, H. and Scherer, K. R. (1995), "Acoustic concomitants of emotional expression in operatic singing: The case of Lucia in Ardi gli incensi," *Journal of Voice.*

[2] Sundberg, J. (1987), "The science of the singing voice," Dekalb, Il: Northern Illinois University Press.

[3] Berndtsson, G. and Sundberg, J. (1995), "Perceptual significance of the centre frequency of singer's formant," *STL-QPSR*, 4/1994, KTH, Stockholm, pp. 95-105.

EFFECT OF VOWEL MODIFICATION ON THE PHONEMIC ACCURACY OF VOWELS AND PALATALIZATION OF CONSONANTS IN RUSSIAN VOCALIZED SPEECH

Ekaterina Oussilova
Moscow State University, Moscow, Russia

ABSTRACT

The study analyses the effect of vowel modification, i.e. certain "colouring" of the vocal sound, on the phonemic accuracy of vowels and quantifies the formant changes it causes for Russian vowels. It is suggested that vowel modification may have an impact on palatalization of consonants. The results of detailed experimental research show that this hypothesis can be confirmed with a certain degree of confidence. However, the conclusion is that the situation is not straightforward and study of extensive additional material is required to help form a more grounded opinion.

INTRODUCTION

Vowel modification is generally agreed to be a key feature of professional operatic singing. It helps achieve an even quality of vocal sound throughout the singing registers, a certain intensity of sound and is a means of producing sounds with a certain colour. Words commonly used as synonyms of vowel modification are "covering" or "darkening". We believe, however, that vowel modification is not confined to covering, i.e. making sounds more closed. Very often in singing a vowel modification towards more open sound can be perceived.

Phonemic accuracy is pronunciation of a sound in such a way that it can be identified by listeners as a realization of the intended phoneme. Lack of phonemic accuracy is sometimes called "vowel migration" towards a different phoneme and in extreme cases can lead to vowel alteration. See [1].

The consonant system of the Russian language is characterized by a correlation of palatalized / non-palatalized consonants (otherwise known as soft and hard respectively). *Palatalization* is an auxiliary articulation in the production of consonants and is accomplished by raising the middle of the tongue towards the roof of the palate. Acoustically, palatalization is chiefly manifested within transition sectors between sounds by an [i]-like element. (See [2].) Vowels following palatalized consonants are more "closed".

An analysis of Russian speech in singing carried out by the author has identified a considerable number of changes in the quality of vowels, which can be described as vowel alteration, and cases of loss of palatalization in "consonant+vowel" syllables. We noticed that these phenomena coincide with vowel modification.

The purpose of this study was to investigate the relation between vowel modification and phonemic accuracy of vowels, to quantify the changes in migrating (altered) vowels and to analyse the effect of vowel modification on the palatalization of consonants.

DESCRIPTION OF THE STUDY

For the subject of the study we chose recordings of the singing of two celebrated Russian opera artists, bass Fyodor Shalyapin and mezzo-soprano

Yelena Obraztsova. The reasons for selecting these singers' phonations were the internationally recognised excellence of their performance and their (possibly only known locally) particular attention to the role of words in singing. It was also considered beneficial to select a male and a female singer in order to be able to cover a wider range of issues. We used recordings of Russian and West European songs and arias performed by Shalyapin as well as his recital of the Nadson's poem *Грёзы* (Dreams) selected from a collection of 12 recordings reissued by Melodia Records in 1980 and a recording *Yelena Obraztsova. Russian Songs and Romances* copyrighted Melodia 1982.

Although we are aware of the influence that recording techniques can have on sound, we believe the accuracy of conveying phonetic parameters to be sufficient for the purposes of this study.

For the identification of cases of lack of phonemic accuracy of vowels (first auditors' session) from the above described material, 90 samples were selected for auditors' and spectrographic analysis. Most of the samples were open syllables:

- 54 of them had the structure "consonant + vowel";
- 30 had the structure "a group of consonants + vowel";
- 6 were "consonant + vowel + consonant" syllables.

The stimuli included syllables where vowel alteration was expected and those containing accurately conceived vowels, or "pure" vowels presented in random order.

For the consonant palatalization study (second auditors' session) the samples had the structure "vowel + consonant / group of consonants + vowel".

All samples were presented to 3 groups of auditors with 4 people in each. All were native Russian speakers, 10 of them also spoke fluent English, 5 had a degree in linguistics, and almost none had any advanced training in music or singing.

In the first session, auditors were asked to write down in ordinary Russian letters what they heard. In the second the assignment was to circle the appropriate word on answer sheets.

The results can be summarized as follows.

Alterations of vowels were perceived in situations of vowel modification of both types ("more open" and "more closed"). Identification of alteration was considered valid when more than 60% of the auditors registered a similar alteration. "Closing" of vowels occurred throughout the singers' ranges and was used for the purposes of bridging registers and achieving the desired intensity and colour. "More open" vowels were used for giving the sounds a brighter colour. The musical and aesthetic sides of these phenomena are beyond the scope of this article.

Phonetically, alterations towards "more closed" sounds identified by the auditors can be grouped as follows.
(The phonetic symbols of Russian sounds we use are the same as those used by G Fant [3]):

[o] > [ou] or [u]

Shalyapin - [mnoi] > [mnoui] in the word combination *со мной* (with me);
- [fs'o] > [fs'ou] in the word *всё* (all);

Obraztsova - [t'ot'] > [t'out] in *цветёт* (blooms);
- [to] > [tou] in что (what).

['e] > ['i] / [e] > [i]
these alterations were found only in Obraztsova's singing:
- [l'e] > [l'i] in *прелестный* (charming);
- [sv'e] > [sv'i] in *светит* (shines);

- [ǯe] > [ǯi] in *же* (a Russian particle).

Alterations *[a] > [o]* were more rare. One of the examples is:

Shalyapin - [pa] > [po] in *покой* (quiet).

Production of vowels of "brighter", "more open" colour lead to the following alterations:

[o] > [a]

Shalyapin - [d'om] > [d'oam] in *идём* (we go);
- [mr'om] > [mr·oam] in *умрём* (we will die).

[u] > [o]

Obraztsova - [muz] > [moz] in *муж* (husband).

['o] > ['e]

Shalyapin - [jom] > [joem] in *поём* (we sing).

[e] > [a]

Obraztsova - [s'e] > [s'ea] in *сердцем* (heart).

As can be noticed from the above examples, alterations differ in their consistency. In some of them a vowel is simply replaced by another vowel, in other cases the quality changes within the duration of one vowel, thus constituting diphthongs.

The diagram below summarizes vowel alteration patterns:

"more closed"	*"more open"*
[a] > [o]	[o] > [a]
[o] > [u]	[u] > [o]
['e] > ['i]	['o] > ['e]
[e] > [i]	[e] > [a]

The second auditors' session showed that the loss of consonants' palatalization is more typical of Shalyapin's singing. Some examples are:

[r'e] > [re] in *ревность* (jealousy);
[v'e] > [ve] in *поверь* (believe).

Some of the auditors perceived in such cases as *серп* (sickle) a partial loss of palatalization, which in Russian linguistic tradition is called "semi-softness" and is indicated by the sign [·]: *[s'e] > [s·e]*

It was noticed that loss of palatalization of consonants occurred in a position before more open (but not altered) vowels.

Subsequently a spectrographic analysis of all the samples was undertaken, including syllables containing "pure" vowels, altered vowels, consonants retaining their palatalization and the "new hard" consonants.

The experiments were performed at the Research Centre of the Moscow State Conservatoire. The sound signals were analyzed with a Russian developed software package "Signal Viewer".

A brief overview of the results follows.

Production of more closed front vowels causes raising of the second formant (F2):
from 1500-1950 Hz for *['e]* to 1800-1950 Hz for *['i]*.

When front vowels are given more open colour, the second formant lowers: from 1500-1700 Hz for ['e] to 1350 Hz for ['a].

Additional lip rounding and covering, as in case of alteration of back vowels *[o] > [u]* results in lowering of the second formant: F2 = 650-800 Hz (F2 for a typical Russian [o] lies within the range of 800-1000 Hz), whereas lip spreading and increasing the degree of mouth opening has the opposite effect on the spectra of back vowels:
[o] > [a] - F2 = 1050 Hz
(as compared with F2 for [o] mentioned above).

The spectrograms of syllables, in which the auditors perceived loss of, or insufficient, palatalization, have shown that the [i]-transitions are present (with the exception of the syllable *[stre]* in Shalyapin's recital of the poem). However, these spectra are characterized by two other features that are of interest:

(a) the relative duration of transition to the whole vowel in sung syllables is much shorter than that in spoken ones (40msec / 570msec and 40msec / 280msec respectively);
(b) the vowels in the selected syllables are more open than the vowels that are found in speech after "soft" consonants.

It is possible to conclude from the above that the perceived loss of palatalization is caused by the interaction of the shorter than in speech relative duration of transitions and more open quality of the vowel caused by modification.

However, this cannot be considered the final solution. The lack of a transition sector that was registered in the spectrum of the *[stre]* syllable from Shalyapin's speech makes it possible that he simply had problems pronouncing a palatalized [r'] sound (70% of cases of loss of palatalization are with sound [r']). This difficulty can be attributed to the influence of a local dialect. Indeed, there are Russian dialects in which [r] sounds can only be hard. Such dialects are spoken near the borders of Byelarus and in some Siberian areas but Shalyapin appears to have had no exposure to those dialects. On the other hand, the remaining 30% of sounds that lost their softness should not be ignored. Moreover, loss of palatalization by soft consonants in pronunciation of opera singers has been described in the 1950s by A Reformatsky [4] who attributed it to singers' affected manners. Phonations of the examples he used cannot be obtained at present, which makes it difficult to argue with his conclusions. It appears that only a considerable amount of additional material will allow conclusions to be drawn with a high degree of certainty.

CONCLUSIONS

This study has confirmed that lack of phonemic accuracy of vowels in Russian vocalized speech (including their alteration) is indeed in some cases caused by vowel modification. As a result of this research exact changes of the formant structure of modified vowels have been registered. It is likely that vowel modification, together with the shorter relative duration of transition sectors of vowels in singing, create an auditory effect of loss of palatalization of preceding consonants. However, this is still largely a hypothesis and requires further investigation.

REFERENCES

[1] Appelman, D R (1967), *The Science of vocal pedagogy. Theory and application,* Bloomington: Indiana University Press, pp.222-234.

[2] Bondarko, L V (1967) "Structure of the syllable and characteristics of phonemes", Voprosy Yazykoznaniya, No 1, p. 37

[3] Fant, G (1960) Acoustic theory of speech production, Russian Translation (1964), Moscow.

[4] Reformatsky A A (1955), "Speech and music in singing", Voprosy Kultury Rechi, 1st issue, Moscow.

ACKNOWLEDGEMENT

I wish to thank *Coopers and Lybrand Moscow* for the financial support of this undertaking and its staff who provided a lot of assistance at all stages of the preparation of this paper. I would also like to express my gratitude to professor N K Pirogova and professor V P Morozov for their valuable advice.

FACIAL EXPRESSIONS IN SINGING
(A Pilot Study)

Nicole Scotto Di Carlo and Isabelle Guaïtella
Laboratoire Parole et Langage, U.R.A. 261 CNRS, Aix-en-Provence, France

ABSTRACT

An experiment dealing with the recognition of emotions in speech and in singing for two subject populations (opera amateurs and non-amateurs) based on visual, auditory, and audiovisual perception tests was used to assess the respective roles of the singer's voice and facial expressions in the perception of emotions produced by a professional soprano, and to determine how spectators decode these emotions.

INTRODUCTION

How do opera singers manage to produce the physiological or functional facial movements (required for emitting a given sound) at the same time as they produce the facial expressions (aimed at displaying the wide range of human feelings) in order to reflect the emotions they must transmit to the public? Do spectators use a specific strategy to decode the emotions expressed by lyrical artists? Do they ignore the functional part of a facial expression and focus their attention solely on the expressive part? We designed a pilot experiment that allowed us not only to assess the respective roles of the voice and the face in the perception of emotions produced by a professional soprano, but also to determine how spectators decode these emotions.

EXPERIMENTAL PROCEDURE

Corpus

The soprano recruited had to be both an excellent professional singer and a good actress. She was filmed in an anechoic chamber using two synchronized cameras so as to obtain simultaneous profile and front view videotapes of her face as she carried out a certain number of tasks of increasing complexity. For the purposes of the present study, the videofilms were used to extract a corpus of the vowel [a], spoken and sung on C_4 (262 Hz) for the lower register, C_5 (523 Hz) for the middle register, and C_6 (1047 Hz) for the upper register, with four basic emotions : joy, sorrow, fear, and anger (which taken two at a time, can be opposed along the activation/inhibition dimension). The best sequences were selected and the most characteristic part of the facial expression in each was photographed from the videotape. This gave us a corpus composed of both sound sequences and photographs.

Testing

Tests were administered to two populations of subjects : 20 opera lovers who regularly watch videotapes of lyric works and 20 subjects with no particular interest in the lyrical arts. Using a computer-driven projection system, the slides used in the visual tests were presented to the subjects for four seconds. Testing was done in three phases : (1) a solely visual phase where the subjects had to judge static images only (photographs), (2) a purely auditory phase where they had to judge sound sequences only, and (3) an audiovisual phase which combined the sound sequences and the static images.

- *Visual test.* The visual test consisted of two steps. In the first, the subjects had to identify the emotions expressed. In the second, they were told what the emotion would be and had to assess its intensity.

The first step (identification of emotions) included two series, one containing real images where each face corresponded to a given emotion and a given register (vowel [a] spoken and sung in the lower, middle, and upper registers) and one containing mixed images where the top and bottom of the face did not correspond to the same emotion. Joy was permuted with sorrow, and fear with anger, always within the same register.

The second step (assessment of emotion intensity) consisted of three phases. In the first, the subjects' task was to choose the most expressive face out of two simultaneously-presented real photographs, each representing a given emotion (of which the subjects were informed) expressed in the lower and upper registers. In the second phase, the two photographs represented the same emotion (again, of which the subjects were informed) expressed in the lower and upper registers, but this time the two halves of the faces were interchanged across the lower and upper registers. For example, a top half (lower register)/bottom half (upper register) face had to be compared with a top half (upper register)/bottom half (lower register) face. The third phase dealt solely with emotions expressed in the spoken voice. For each emotion, the subjects had to select the face they felt was the most expressive among three simultaneously presented photographs, one of a real face and two of reconstructed faces made from two right halves and two left halves.

- *Auditory test.* The subjects' task was to identify the emotions by listening to two series of randomly mounted sound sequences containing the vowel [a] spoken and sung in the three registers.

- *Audiovisual test.* The subjects had to identify the various emotions by looking at photographs of real faces on which each facial expression corresponded to a given emotion and a given register (vowel [a] spoken and sung in the lower, middle, and upper registers) while listening to the corresponding synchronized sound sequences.

RESULTS

Soprano's Morphological Features

- *Expressivity index.* The soprano who volunteered for this experiment has a relatively symmetrical face. Because of this, her right side was judged to be as expressive as her left side (36% and 37%, respectively) for all emotions.

It is interesting to note that, for all emotions pooled, the real face was judged to be expressive by only 24% of the subjects. It appears as though joining two right halves and two left halves reinforces the facial symmetry, making the mixed faces more expressive.

For sorrow, fear, and anger, the faces made up of two left halves were considered by the greatest number of subjects (49%) to be the most representative of the emotion in question (compared to the real face and the face made up of two right halves). These results are consistent with those obtained in an experiment by Sackheim and Gur [1], where the emotion in faces composed of two left halves was judged to be more intense than in faces composed of two right halves. For joy, the right side was judged to be more expressive by 50% of our subjects, a finding which is in line with Bruyer's [2] theory that the left hemisphere is involved in the expression of pleasant emotions.

- *Within-register salience index.* For all subjects and emotions pooled, when joy and sorrow or fear and anger were mixed within a given register, the bottom of the face clearly dominated in speech (B = 54% and T = 32%), whereas this effect was not as pronounced in singing (B = 37% and T = 32%). Moreover, the differences between the top and bottom of the face gradually decreased from the lower register to the upper register (lower : T/B = 37/43; middle : T/B = 33/37; upper : T/B = 27/31), probably because emotion information becomes increasingly difficult to perceive, making it necessary to use all available cues regardless of whether they are on the top or bottom of the face.

For all subjects and all registers pooled, the bottom of the face dominated for joy (T = 13%, B = 58%) and sorrow T = 16%, B = 51%), whereas the top dominated for fear (T = 36%, B = 18%) and anger (T = 61%, B = 46%). These results partially corroborate Bassili's [3] findings, which showed that the top part of the face is used to detect anger; the bottom, joy, sorrow, and disgust; and both parts, surprise and fear.

Between-medium Comparisons

Regardless of the medium used (sound, image, sound+image), the emotions expressed in speech were identified the best (83% vs. 56% for singing).

For all subjects, emotions, and registers pooled, sound was found to be the poorest conveyor of information about emotions (39%), whereas images attained 74% and sound accompanied by images, 76%. The combination of image and sound improved the score by 7% over images alone for the spoken voice (I = 88%, S+I = 95%) and by 4% over singing in the lower register (I = 72% and S+I = 76%). In contrast, the scores were virtually the same in the middle (I = 81% and S+I = 82%) and upper (I = 53% and S+I = 54%) regis-

ters. As the acoustic cues of emotion are partially destroyed in these registers, almost no additional information is provided by the sound. In the other registers, the acoustic cues are added to the visual cues and thereby contribute to improving recognition.

Among the four basic emotions, for all subjects, registers, and media combined, sorrow (73%) and joy (72%) were recognized the best. Then came fear (54%), with anger in last place (53%).

Comparison of Lower and Upper Registers

- *Between-register comparison of expressivity.* When subjects had to choose the most expressive of two real photographs representing the same emotion in the lower and upper registers, the lower register predominated at 50%, vs. 48% for the upper register. For the emotions taken separately, the upper register was judged to be more expressive for joy, sorrow, and fear, while the lower register was judged so for anger. This is no doubt due to the wide buccal opening in the upper register, which gives the impression of a more intense emotion, whether it be joy, sorrow, or fear. The small buccal opening in the lower register is more representative of anger, whose main visual feature is the clenching of the teeth.

- *Between-register comparison of salience.* When for a given emotion, the lower and upper registers were mixed to obtain L/U and U/L pairs, the lower register dominated in 55% of the cases. The faces judged to be the most representative of joy and sorrow were the U/L mixtures (lower register on bottom of face). The faces judged to be the most expressive of fear and anger were the L/U mixtures (lower register on top of face). This confirms our previous results. The dominance of the lower register can be explained by the fact that, in the upper register where all facial muscles are being used to produce the sound, achieving good vocal technique takes precedence over expressing emotions.

Auditory Identification of Emotions

At the auditory level, for all registers and subjects pooled, the recognition rate for sorrow was 56%, fear 38%, joy 33%, and anger 28%.

In the upper register, for all subjects pooled, joy was recognized the best (48%). In contrast, none of the subjects recognized sorrow (0%) : 45% thought it was joy, 30% fear, and 23% anger. This is most likely related to existing emotional stereotypes. Joy, fear, and anger are the three emotions which, when speaking, the soprano produced in the upper register (sometimes even in the upper-upper register). It seems that joy is associated with the upper register. Inversely, one can hypothesize that the lower register is associated with sorrow. If this hypothesis is valid, not only should sorrow be recognized better in the lower register (which was indeed the case : 73%) than in the upper register (0%), but joy should be poorly recognized in the lower register (which was also the case : 18%). Fear and anger seem to follow the sorrow pattern and are treated as "internalized" versions of the traditionally accepted emotional stereotypes, with recognition rates of 28% in the lower register and 8% in the upper register for fear, and 38% in the lower register and 13% in the upper register for anger.

CONCLUSION

The face of the soprano who volunteered for this experiment has a mixed salience index. This means that whether she is speaking or singing, the part of her face which determines the recognition of emotions depends on the emotion being expressed : the top of the face dominates for fear and anger, and the bottom dominates for joy and sorrow. Moreover, this soprano has a neutral expressivity index. In other words, the right side of her face is judged to be as expressive as the left, probably due to the regularity of her features.

Regardless of the medium, the emotions expressed in speech were identified better than those expressed in singing. The visual medium was the best, whereas sound was a poor vehicle of emotion information. The combination of sound and image led to a slight improvement in emotion recognition in speech, but did not turn out to be very effective in singing, where the partly destroyed acoustic emotion cues are unable to supply any additional information.

For the visual aspect of this study, the emotions expressed in the upper register appear as a whole to be overridden by those expressed in the lower register, judged to be more expressive. This finding is not surprising in that the entire facial musculature is involved in upper register emissions, which jeopardizes the expression of emotions. However, the upper register appears to be considered more expressive for joy, sorrow, and fear, most likely because of the wide buccal opening which reinforces the intensity of the emotion being expressed. The lower register in turn is judged to be more expressive for anger, no doubt due to the fact that the narrower buccal opening in the lower register corresponds more closely to this emotion, generally expressed by a set jaw.

For the auditory aspect of this study, the emotion identified the best was sorrow, although none of the subjects recognized it in the upper register where it was mistaken for joy. Due to the existence of strong emotional stereotypes, joy seems to be associated with the upper register, and sorrow, with the lower register. This accounts for the fact that sorrow was not recognized in the upper register and that joy was poorly identified in the lower register.

In summary, for singing, the identification of emotions and the assessment of their intensity appears to be influenced by parasitic phenomena related to their extreme production conditions. For instance, buccal opening plays a role in the assessment of emotional intensity, which is judged to be greater when the mouth is wide open (upper register) than when it is only slightly open (lower register). Likewise, the fact that unconsciously, most people associate joy with the upper register and sorrow with the lower register, has an impact on the recognition of these emotions in the outer registers.

Whereas in speech, the recognition of emotions is not subject to any particular production constraints, in singing the constraints are great and are manifested in the face by specific movements which interfere with the expression of emotions. Due to this fact, emotions are only correctly identified when the functional and expressive movements are compatible. In all other cases, the functional movement takes precedence over the emotional expression, because even when an opera singer is an excellent actor or actress, he or she cannot run the risk of jeopardizing the quality of the sound emitted in order to express an emotion or a feeling.

REFERENCES

[1] Sackheim, H.A. & Gur, R.C. (1982), Facial asymmetry and the Communication of Emotion, in *Social Psychophysiology*, Cacioppo J.T. & Petty R.E. (eds) Guilford Press, New York.

[2] Bruyer, R. (1980), Implication différentielle des hémisphères cérébraux dans les conduites émotionnelles, *Acta Psychiatrica Belgica*, 80, 266-284.

[3] Bassili, J.N. (1979), Emotion recognition : the role of facial movement and the relative importance of upper and lower areas of the face, *Journal of Personality and Social Psychology*, 37, 2049-2058.

ACKNOWLEDGMENTS

We would like express our sincere gratitude to all those who contributed to making this study come to be : Nathalie Labry lyrical artist, who so kindly and patiently complied with our scientific demands, Bernard Teston, C.N.R.S. research engineer, who designed part of the experimental equipment, Georges Bourrat and Louis Seimandi, University of Provence technicians, who skillfully ran the various phases of this extremely complex experiment, Robert Espesser, C.N.R.S. research engineer, who developed the software for the random auditory testing procedure, Claudette Bonnaud, head of the A.E.L, who lent us the projection equipment, Annie Poujol and Marie-Françoise Borel, in charge of the International Lyrical Arts Video Library in Aix-en-Provence, who with their usual friendliness and conscientiousness, allowed us to use the video libary auditorium and helped in recruiting part of the subjects, Vivian Waltz, who agreed to do the English translation in an amazingly short time, and finally, all the subjects, known and unknown, who were willing to devote some of their time to taking the various tests.

CO-MODULATION OF F_0 AND FORMANT FREQUENCIES IN SINGING

Leo-Geert van den Berg, Christel de Bruijn, Gerrit Bloothooft, Guus de Krom
Research Institute for Language and Speech
Utrecht University, The Netherlands

ABSTRACT

Vocal vibrato can be considered as the result of pulsations in muscular activity at the respiratory, laryngeal, and supra-laryngeal levels. Sometimes, the latter can be observed in singers as rhythmic movements of the entire larynx and the pharyngeal wall. This will influence formant frequencies and hence the timbre of the sound. We investigated whether modulations of formant frequencies can be observed in singing, and how they relate to modulations of the fundamental frequency.

INTRODUCTION

In singing, we often observe vibrato as a modulation of the fundamental frequency (F_0), with a rate between 5 and 7 Hz. Several researchers report a co-modulation of the activity of intrinsic laryngeal muscles with F_0, most notably perhaps the cricothyroid muscle [1]. As a second mechanism, rhythmic pulsations of the respiratory muscles resulting in modulations of the subglottal sound pressure have been mentioned [2]. In addition, a co-modulation between F_0 and the activity of several supralaryngeal muscles has been found [3]. This co-modulation, among others, influences laryngeal height, as can sometimes be seen in video-stroboscopic images of the pharyngeal cavity: the entire larynx, as well as the pharyngeal wall, can be in regular movement during singing. These findings indicate that vibrato is a complex neuromuscular phenomenon that affects all aspects of voice production. The precise coordination of the various pulsated muscular actions is still largely unknown.

In the present investigation, we concentrated on the acoustic effects of vibrato as originating from a modulation of the vibration frequency of the vocal folds, and from a modulation of the volume of the pharyngeal cavity. First, we investigated whether a regular variation of the volume of the pharyngeal cavity exists to an extent that it can be measured as a modulation of formant frequencies. Second, whether there is a co-modulation between formant frequencies and F_0.

In a stationary situation, relations between vertical larynx positioning and the singer's formant have been studied theoretically and experimentally [4]. The main effects of a raised larynx were: (1) a significant increase of F_2 in high front vowels, (2) a raise in F_1 and F_2 for open vowels, (3) a raise in F_3 and F_4. Similar effects should be found during rapid modulations of the larynx height.

METHODS

We used recordings of four professional male singers (with a classification between bass and baritone) who sang the vowels /a/, /i/, /u/, and /ɔ/ at F_0 = 98 Hz in three conditions: straight (none to little vibrato), normal, and exaggerated vibrato, yielding a total of 48 recordings. The recordings were digitized at 20 kHz. For these singers, we had the audio material available, but no information on larynx movements.

We chose the low F_0 value of 98 Hz to ensure a high accuracy of the F_0 measurement (peak picking algorithm in the time domain). For formant frequency measurements, we first downsampled the data from 20 kHz to 10 kHz before performing an LPC-12 analysis. This was done to increase the accuracy of the formant frequency measurement by focussing on the 0-5 kHz spectral region.

Still, LPC procedures tend to be sensitive to the distribution of harmonics (and hence to F_0). Because there are various factors that influence the value of computed formant frequencies, we give a formal description of these factors to facilitate the interpretation of the results.

MODELLING F_0 AND FORMANT FREQUENCY MODULATION

We first make the assumption that undulations in the activities of muscular structures that have an effect on sound production (respiratory, laryngeal, and supralaryngeal structures) are coordinated at a fairly central level, and have identical rates. The acoustic phenomena associated with these centrally coordinated movements may exhibit phase differences, however.

We combine the acoustic effects of the pulsated activity of respiratory muscles, influencing the subglottal pressure, and the internal laryngeal muscles, influencing the tension of the vocal folds, in a modulation of the fundamental frequency $F_0(t)$:

$$F_0(t) = F_0 + \Delta F_0 \sin(2\pi Vt) \qquad (1)$$

in which $F_0(t)$ is the time-varying value of the fundamental frequency, F_0 is its average value, ΔF_0 is the extent of the frequency variation, and V is the vibrato rate. This combination may be considered a simplification, because interactions between modulating subglottal pressure and modulating laryngeal muscular tension can be quite complex. On the other hand, $F_0(t)$ normally shows a regular pattern during vibrato.

If the laryngeal height modulates due to a vibratory activation of the supra-laryngeal muscles, this affects the length of the vocal tract as well as the volume of the pharyngeal cavity. As a consequence, the formant frequencies will also be modulated. There are two aspects to consider here. First, a one-tube model of the vocal tract would predict that a raised larynx will result in increased formant frequencies. However, the actual effect may be more complex. Second, there may be a phase difference between larynx height modulation and the modulation of the vibration frequency of the vocal folds. We combined both effects into one phase term, and described the modulation of the center frequency of a formant n as follows:

$$F_n(t) = F_n + \Delta F_n \sin(2\pi Vt + \varphi) \qquad (2)$$

in which F_n is the average formant frequency, and ΔF_n is the maximum deviation of F_n.

It follows from equations (1) and (2) that a co-modulation of $F_0(t)$ and $F_n(t)$ may be observed with an unknown phase difference.

Furthermore, we have to consider the possibility that the LPC procedure used to estimate the formants yielded measurement errors. Ideally, we would like to compute formant frequencies independent of the harmonic structure. However, this is not possible in practice. In an extreme situation, for instance for high-pitched vowels with $F_0 > 500$ Hz, LPC analyses will result in formant estimates that follow the frequencies of the separate harmonics. Although this artefact will be less outspoken for the low F_0 value of 98 Hz that we used, we cannot exclude the possibility that computed formant frequencies to some extent follow the fundamental frequency:

$$F_n(t) = F_n + \Delta F_n \sin(2\pi Vt) \qquad (3)$$

The resulting co-modulation of $F_0(t)$ and $F_n(t)$ should be in phase.

There is an additional interaction between F_0 and harmonics underlying a formant. That is that the *amplitude* of these harmonics will show a modulation too: in phase with F_0 modulation along the positive slope of a formant, in counter-phase along a negative slope, and with a rate that is twice the vibrato rate of F_0 if the harmonic varies symmetrically around a formant frequency [5]. If the LPC procedure does not model this variation properly, the estimated formant frequencies may show the following types of modulations:

$$\begin{aligned} F_n(t) &= F_n + \Delta F_n \sin(2\pi Vt) \\ F_n(t) &= F_n - \Delta F_n \sin(2\pi Vt) \\ F_n(t) &= F_n + \Delta F_n \sin(4\pi Vt) \end{aligned} \qquad (4)$$

The combined effects of larynx height variation and formant measurement artefacts related to F_0, the combination of Eqs. 2-4, may be rather complex. This will have to be taken into account in the interpretations of the results.

RESULTS AND DISCUSSION

Vibrato rates of the 48 recordings varied between 5 and 7 Hz. The average extent of Fo-modulation frequencies (maximal deviation from the average value) for straight, normal, and exaggerated vibrato was 1, 3, and 5.5%, respectively. These are typical values [6]. With exaggerated vibrato, however, a maximum extent up to 10% (about 2 semitones) could be measured.

In several cases modulations in formant frequencies were found, but the results were of a highly variable nature and difficult to generalize. There was a tendency, however, for modulations to be less present in the higher formants. For this reason, we present a few examples here to demonstrate several types of co-modulation between F_o and F_1.

Figure 1 gives a clear example of co-modulation of F_o and F_1. The extent of the modulation is 5.5% for F_o and 6.0 % for F_1. The correlation coefficient between F_o and F_1 is .80. The close correspondence between the two traces, with almost equal relative extent, may indicate an artefact of F_1 computation that seems to follow the fourth harmonic.

A complex pattern is shown in Figure 2. This is an example of exaggerated vibrato with an F_o-extent of 10%. Apart from modulations, there is a gradual increase in the value of F_1, which can be explained by some variation in articulation of /i/. The superimposed modulations in F_1 have and extent of 3%. First they are in counter phase, later they have a double rate. It can be seen that in the first part, for the high F_o values (104 Hz), the F_1 values meet the frequency of the third harmonic (312 Hz), while for the low F_o values (85 Hz), the F_1 values meet the fourth harmonic (340 Hz). This gives a strong suggestion of a computational effect. We do not have an acceptable interpretation for the last part with double rate of modulation of F_1.

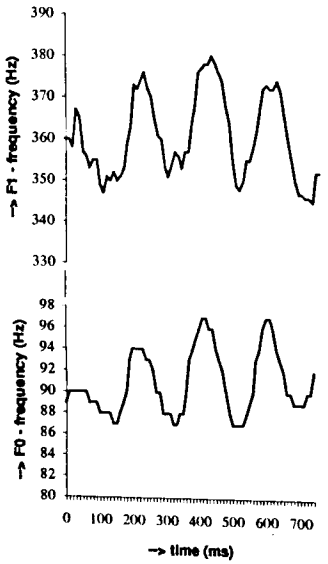


Figure 1. F_o and F_1 for the vowel /u/, normal vibrato.

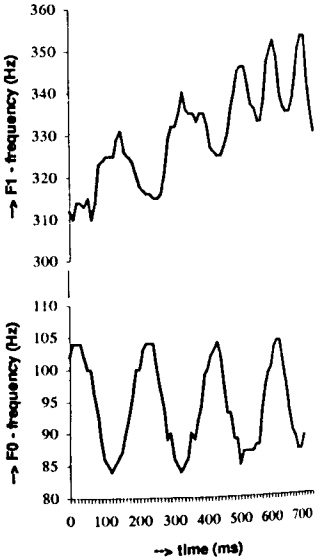


Figure 2. F_o and F_1 for the vowel /i/, exaggerated vibrato.

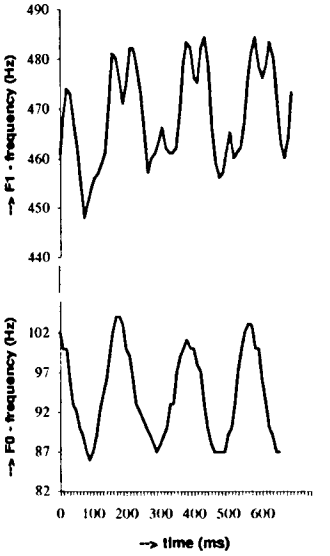


Figure 3. F_o and F_1 for the vowel /ɔ/, exaggerated vibrato.

Figure 3 again shows exaggerated vibrato, with an extent of 8.5%. The F_1 trace is complex and best described as two superimposed sinusoids, one with the same rate and phase as the F_o, and one with a double rate. There is no simple harmonic relationship between F_o and F_1 that may provide an explanation in terms of a computational effect for the entire trace. We hypothesize the following. The modulation with the same rate as F_o may be the result of variation in the height of the larynx. To explain the modulation at the double rate, we note that the fifth harmonic varies between 440 Hz and 500 Hz and thus passes the actual formant frequency in an almost symmetrical manner. Its amplitude will therefore vary with twice the F_o rate. As suggested earlier, the computation of F_1 may follow this rate.

CONCLUSION

Although co-modulation between F_o and F_1 can be measured in singing, an interpretation of these results is difficult. Both a movement of the larynx and the dependency on F_o of the LPC-based computations of F_1 can explain the results. More detailed investigations into the latter effect are needed. Measurement of larynx height with multichannel electro-glottography [7] may be of help in the interpretation of acoustic data.

REFERENCES

[1] Shipp, T., Doherty, T., and Haglund, S. (1990), "Physiologic factors in vocal vibrato production", *Journal of Voice*, vol.4, pp. 300 - 304.

[2] Sapir, S. and Larson, K. (1993), "Supra-laryngeal muscle activity during sustained vibrato in four sopranos: surface EMG findings", *Journal of Voice*, vol.7, pp. 213 - 218.

[3] Sundberg, J. and Nordstrom, P.E. (1976), "Raised and lowered larynx - the effect on vowel formant frequencies", *STL-QSPR*, vol.2, pp. 35-39.

[4] Lindblom, B.E.F. (1981), "Studies of articulation", *STL-QSPR*, vol.2, pp. 36-51.

[5] Horii, Y. (1989), "Acoustical analysis of vocal vibrato: a theoretical interpretation of data", *Journal of Voice*, vol.3, pp. 36-43.

[6] Sundberg, J. (1993), "Acoustic and Psychoacoustic Aspects of Vocal Vibrato", written version of paper for the COMETT conference, Utrecht.

[7] Rothenberg, M. (1992), "A multichannel electroglottograph", *Journal of Voice*, vol.6, pp. 36-43.

THE IMPORTANCE OF DISABLED PHONATION IN B. BRITTEN'S BILLY BUDD

Sibylle Vater

Centre de phonétique expérimentale
Université de Genève, Suisse

ABSTRACT

The purpose has been to show by acoustical analysis the essential function of Billy Budd's (BB) stammering in relation to the structure, the main characters and the intrinsic value of the opera. Fabrice Raviola, baritone, who has sung Donald's part in the late Genevan performance (March, 1994), has provided us with live singing of BB's stuttered articulation, free of orchestral interferences.

1. THE CORE OF THE OPERA

The very heart of the work is given by the instrumental leitmotif that appears in Vere's prologue, underlined by a sustained ppp trill. It becomes increasingly strong and insisting when the captain sings: There is always some flaw in it,...some stammer in the divine speech. Just on the extended final syllable of divine before Vere attacks speech, the clarinets play the theme that will illustrate all of BB's fits. The essence of Britten's drama is comprised in this cell of music: that is the ambiguity between BB's honesty, his helpfulness and the evil which assails him from outside. BB stammers whenever some outstanding emotion is overwhelming him.

As to the trill, it is much more extended than the short leitmotif. It indicates BB's acute sensitivity. When we hear it, his psychophysical coherence is splitting up, all his being is shaken. Any other character or happening is than subordinated to the hero's climaxes. While the tremolo sounds

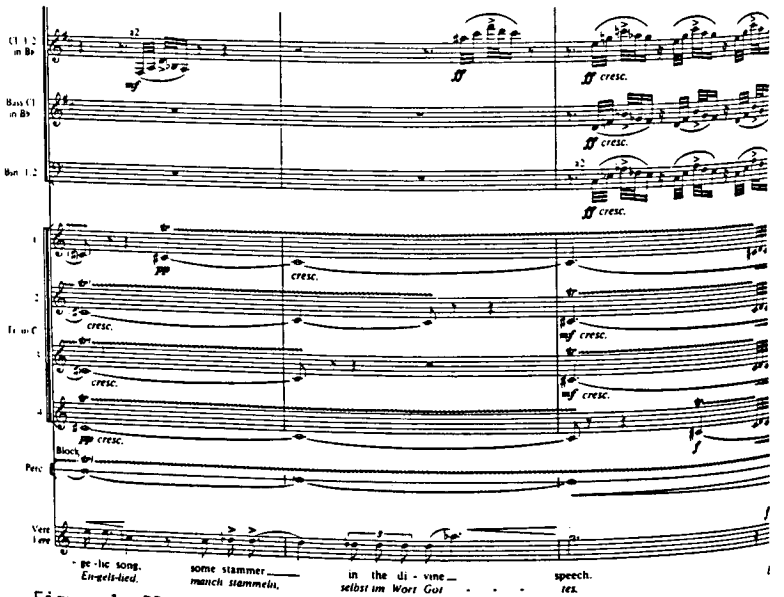


Figure 1. BB, act 1, prologue, Vere, [1]:5.

- even indirectly, for instance, when Claggert is acting his accusation against BB - everything gets uprooted and has to be reset. Any configuration becomes possible and at the same time these shivers push BB to his fatal issue.

2. BB's DISABILITY ARTICULATES THE OPERA

There are four crucial events of BB's stammering:

1. near the beginning, when the Master-at-Arms questions BB about his origin (I,1) and the latter sings: They/ say —/— I was a.../ was a ... a.../ a... a... a.../ a... a.../ ...a/ ...a/foundling /;

2. when BB sees Squeak rummaging in his bag (I,3): Hi! —/— You... a... a.../ a...-/ ... - a..—/— - a... a...-/- a...-/ Come out of that.-/;

3. when the Novice offers BB money to set up mutiny (I,3): Why for/ me? -/ - Why —/— d'ye think I'd.../ - Ah-.../ A... A... A...—/— A... A...- / Dansker old friend, glad - to/ see you.-/;

4. after Claggert's false accusation, just before BB carries out his fatal stroke (II,2): Vere: Defend yourself! BB: /— / a...—/— a... / - a... / a... a...-/ - - a.../ - a... -/ a...- a...—/— - a...——/ devil! - /.

By their placement, the introductory and the tragic final stammerings are particularly important and more elaborate than the other two. These intermediate ones occur in relative proximity, more or less half-way between the two outstanding stutterings and they are characterized by a lighter structure. Linking in a wide span BB's debut and his exit, they act as half-time trials.

3. PHONETIC AND MUSICAL SIMILARITIES OF THE FOUR STUTTERS

Besides the presence of the leitmotif in the orchestra, speech is blocked more or less on the same vowel [æ] or [a] during quadruple time (4/4). This one may be preceded (events 1 and 3) and followed (event 1) by 3/4 time but it is basicly necessary, especially for Britten's various handlings of syncopation, in order to shape the stammers to the utmost. Quadruple time offers the ideal frame to dislocate syllables by contra tempi into jerks until speech is rehabilitated and accompanied by action. The lower and the upper formants of [æ/a] are "frozen" in a compact display (Fig.3,4,6). Their neutral acoustical profile, largely rendered by quavers and semiquavers, assures optimal flexibility for rhythmical effects, crescendos and decrescendos as well as subtle instrumental and external vocal interspersions (s. second event).

4. PARTICULARITIES OF THE FOUR EVENTS OF INHIBITED SPEECH

The First Event (I,1)

BB has just joyfully proclaimed: But I can sing! The timpani underline the significant final word with a foretelling tremolo. Three-four time rules the first measures of stammering. In the middle of the verb say, the tremolo sounds again and a considerable extension of the syllable begins to unbalance BB's speech. Then, during a syncopated pp stammer, 3/4 time changes into 4/4 time. There follow three decrescendo arsis quaver stutters and another syncopated one. The rests lengthen. The next two stammers are contiguous, marking despair by redundancy: the one is a quaver, the other moves into a contra tempo, stopping dead like a hiccup. After a last crescendo arsis stammer, 3/4 time is reset and a quickly fired discharge on foundling (sf [[) brings back fluent speech.

As to the coherence of the opera, it is remarkable that at the break of the antepenultimate stutter, the Second Mate starts singing Vere's statement There is always some flaw in them. (slight variation). There is to be noticed, too, that the recovery of speech progresses by rebounds. BB sings foundling four times, unfolding the

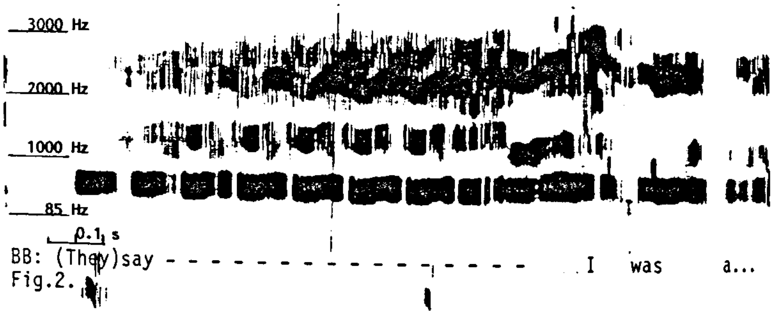


Fig.2.

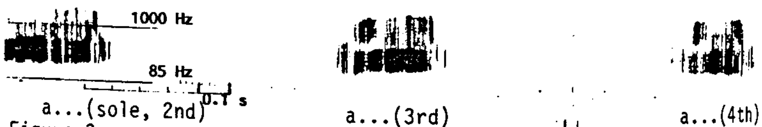


Figure 3.

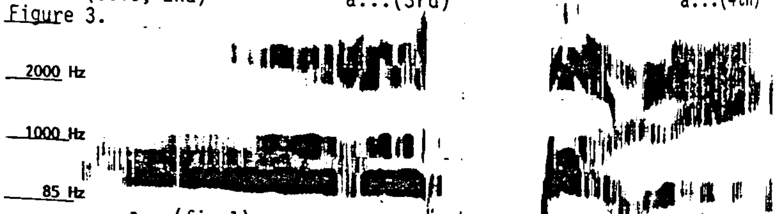


Figure 4.

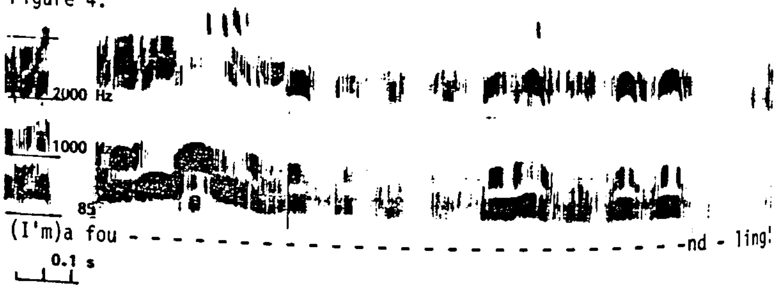


Figure 5.

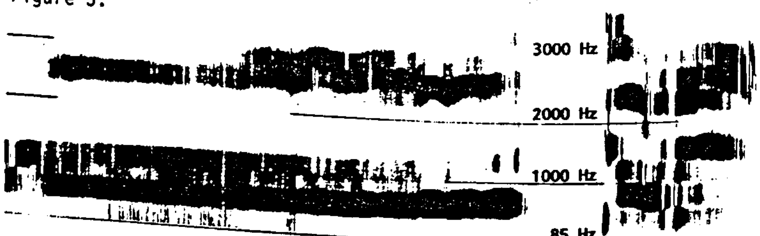


Figure 6.

notion like a precious find (Fig.5).

The Second Event (I,3)

For lack of space, we just want to compare the shortest and most isolated stammer of the previous happening (Fig.7) to the frontal stammer reduction and its effect of delay we find during the second event (Fig.8). Between this reduced stutter and the preceding one, Red Whiskers and Dansker insert the comment: He's a stammer!

The Third Event (I,3)

It is announced by a striking expansion of the pronoun why which breaches the usual syllabic rates. Psychic and dramatic requirements justify the singing voice in transgressing phonetic boundaries.

As to the stammering itself, compared with the second happening, it comprises similar frontal stutter reduction by a preset rest, associated with syncopation.

The Fatal Event (II,2)

Falsely accused, BB is dumbfounded: his stammer rises from complete muteness. And immediately after his fatal stroke, he is silent again. The stutter is projected straightforward without any preliminary or subsequent swell of syllables and words. Moreover, it is enacted in an extremely sharp-cut way: the longest stammer touches directly the shortest and strongest discharge (Fig.9).

There is no further stutter in the opera. The night before being hanged, BB sings a highly melodious part which, as it progresses, becomes a testimony. It ends with the repetition of brief phrases, a sort of phrasal stammer, a far echo of BB's severe inhibitions. Finally, there is the inarticulated singing of the choir.

5. CONCLUSION

BB's fits of aphasia are the fundamental knots of the musical drama. They link the characters by their acoustical, rhythmical and psychic constituents and give the opera its architecture and its symbolic dimension.

6. REFERENCES

[1] Britten, B., Forster, E.M., Crozier, E. (1985), Billy Budd, an opera in two acts, opus 50, revised version 1961, Londo, etc.: Boosey & Hawkes.
[2] Boukobza, J.-F. et al. (1994), "Billy Budd, Benjamin Britten", L'Avant-scène Opéra, no 158.

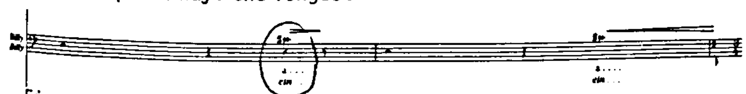


Figure 7. BB, I,1, first event, [1]:56.

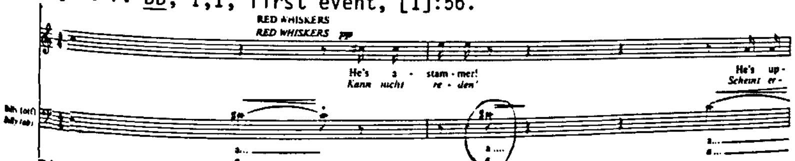


Figure 8. BB, I,3, second event, [1]:186.

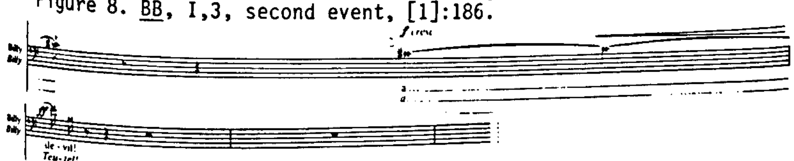


Figure 9. BB, II,2, fourth event, [1]:414-415.

HOW DOES STUDYING INFLUENCE ONE'S VOICE QUALITY?

Allan Vurma and Tarmo Pajusaar
Estonian Academy of Music, Tallinn, Estonia
Einar Meister
Institute of Cybernetics, Tallinn, Estonia
Jaan Ross
Institute of Estonian Language, Tallinn, Estonia

ABSTRACT

Excerpts from voices of 42 students at the Estonian Academy of Music, with different length of their study, were digitally recorded and acoustically analyzed with respect to fundamental and formant frequencies. It was found that objective characteristics of a student's voice often does not meet accepted standards of a good-quality opera voice. The level of the singer's formant and the amplitude of frequency vibrato tend to increase systematically with the number of years studied.

INTRODUCTION

Formants are peaks of spectral envelope caused by the vocal tract resonances. While the first two formants are mostly responsible for differentiation of different vowels, the function of higher formants is believed to influence the voice color. The so-called singer's formant is a spectral peak at the frequency region between 2500 and 3000 Hz, observed predominantly in spectra of male voices. It enables an opera singer to be heard "over" the big symphony orchestra because the spectral envelope of orchestral sound decreases more rapidly than that of a trained singer and also because of higher sensitivity of the human auditory system in the region specified above.

The singer's formant is considered to be an acquired skill. Its articulatory origin is believed to be in lowering the larynx as well as widening the pharynx and the Morgan cavities inside the larynx box which results in clustering together the 3rd and the 4th formants [1]. The singer's formant is an important factor in the perception of a singer's voice category. Higher voices tend to have the singer's formant at higher frequencies [2].

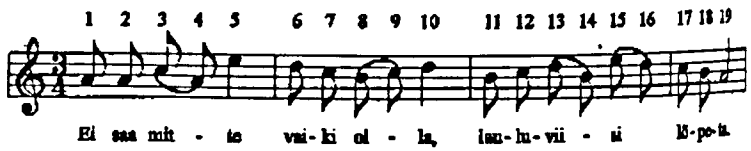


Figure 1. The phrase from a song by Estonian composer Miina Härma, "Ei saa mitte vaiki olla".

MATERIAL AND MEASUREMENTS

A seven-word phrase from the beginning of a well-known melody by Estonian composer Miina Härma (see Fig. 1) was recorded three times: as sung in E minor or A minor, or as spoken, by 42 voice students (16 male and 26 female) from the Estonian Academy of Music. The duration of years studied before the recording was different for individual students, and varied from 1 to 9 years. The students have been instructed by different professors (total of 12). Recordings were made with a DAT recorder (SONY TCD-D3) and a microphone ML-19 in a room with small reverberation.

All recordings were subsequently rated by 4 experts (3 professional singers and 1 musicologist) on two scales. Tone quality and intonation purity were both rated on a five-point scale. In addition, experts were asked to determine the category of voice (tenor, baritone, or bass for males, soprano or mezzo-soprano for females) a singer should belong to, according to the recording.

Measurements were made with a Kay Elemetrics Company Computerized Speech Laboratory (CSL), with a sampling rate of 16 kHz. For each singer, long-term average spectra (Hamming window, 1024 points) were computed for the whole phrases, and LPC (filter order of 16, frame length 15 ms, autocorrelation method) formant histories as well as fundamental frequency fluctuations for the separate syllables were found.

RESULTS

For the voices studied, it was possible to distinguish four different shapes of the spectral envelope in the region of the singer's formant, i.e. around 2.7 kHz (see Fig. 2): (1) a clearly pronounced triangle, (2) rise of the envelope which may result either in a rather flat maximum or in two equally strong peaks, (3) a small rise of the envelope with 3 to 5 equally strong peaks at the maximum (occurs mostly in female singers), and (4) no clear level increase. Table 1 presents the following average data separately for 2 sexes, 2 tonalities, and 5 voice categories: singer's formant frequency, its level, and the normalized frequency distance between the 3rd and the 4th formants.

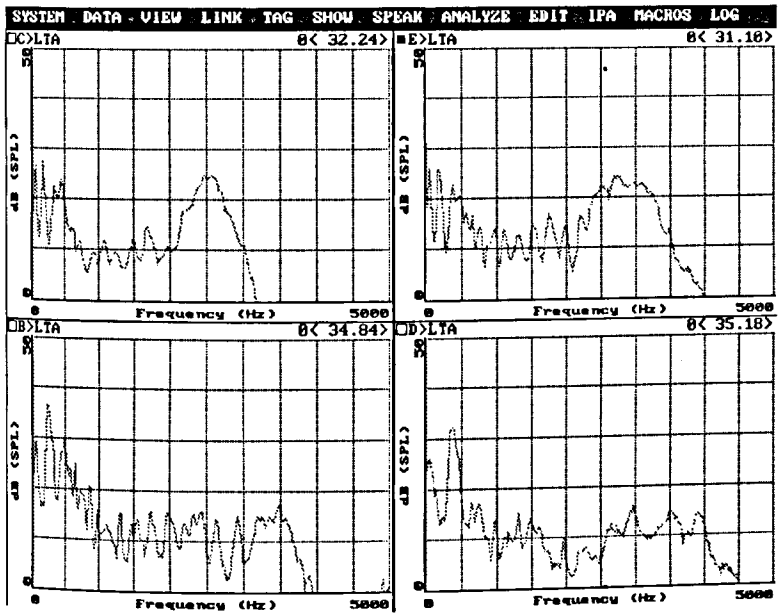


Figure 2. Four different shapes of the singer's formant: a triangle (top left), a flat maximum (top right), with no clear level increase (bottom left), and with a couple of smaller peaks (bottom right).

Analysis of variance demonstrates influence on the level of the singer's formant by (1) the normalized frequency distance between the 3rd and the 4th

formants ($p<0.001$), (2) the duration of a singer's vocal studies ($p<0.01$), and (3) the singer's sex ($p<0.001$). Greater distance between the 3rd and the 4th formants corresponds to smaller level of the singer's formant. The level increases with the number of years studied: e.g., the average for female singers of 1 to 2 years of study is -20.5 dB with respect to the highest peak in the spectrum, in comparison to -15.6 dB for those of 7 and more years of study. The level is generally 10 dB higher in male than in female voices. Intrasex differences in the level are not significant.

The frequency of the singer's formant is significantly different in male and female voices ($p<0.05$). The singer's formant tends to be higher in frequency, higher is the voice, providing that a mezzosoprano follows a tenor in ascending rank order. Intrasex differences in the frequency, however, are not significant. The expert ratings of the voice quality significantly correlates neither with the level nor the frequency of the singer's formant.

Table 1. Averaged data on the singer's formant frequency (fm2, Hz), level (sform, dB), and the normalized distance between the 3rd and the 4th formants [(F4-F3)/F3], with standard deviations (σ).

	key	Σ	fm2	σ	sform	σ	(F4-F3)/F3	σ
fem	E	26	3337	569	-18.4	3.5	38	6.2
fem	A	26	3065	560	-18.4	3.7	43	6.3
male	E	16	2717	338	-8.4	4.9	26	5.5
male	A	13	2732	304	-6.1	8.9	25	5.5
sopr	E	18	3456	572	-17.7	3.4	39	6.7
sopr	A	18	3165	592	-17.9	3.7	44	6.2
mezzo	E	8	3127	555	-20.0	3.4	36	4.8
mezzo	A	8	2840	429	-19.5	3.7	39	4.5
tenor	E	6	2612	339	-10.1	7.5	27	6.3
tenor	A	6	2914	359	-6.2	12.4	26	6.6
barit	E	6	2734	359	-7.2	2.4	24	6.3
barit	A	5	2585	154	-5.0	5.6	24	5.4
basso	E	4	2848	344	-7.8	2.4	23	2.2
basso	A	3	2552	33	-7.5	6.3	23	0.3

LPC formant history analysis was used predominantly in order to compare differences between voice production results for speech and singing. However, because of high fundamental frequency and resulting big frequency distances between harmonics, the LPC analysis results for lower formants tend to coincide with the harmonic frequencies. The results therefore may be used for higher formant frequencies only.

In singing, all formant frequencies are less in value than in speech. The most pronounced differences between different voice categories seem to be in F3 for male speech ($p<0.05$) and in F5 for female singing voices ($p<0.05$).

For the frequency vibrato, its rate and amplitude have so far been measured for two notes (syllables), nos. 5 and 19 (see Fig. 1). The vibrato rate varies from 4.8 to 9.9 Hz for the investigated singers. This variation across the singers is considered too wide, as the normally accepted limits of the vibrato rate are between 5.5 and 7.5 Hz [3]. Its large variation for our subjects can be understood as evidence about the untrained nature of their voices. The amplitude of vibrato was between 9 to 113 cents which, according to [3], is within acceptable limits of up to 200 cents. There is a significant tendency for the vibrato rate to increase with the higher voice category ($p<0.01$) as well as a significant tendency for the vibrato amplitude to increase with the duration of study ($p<0.001$). It is interesting to notice that in comparing vibrato of voice students at 4 different professors, those studying at LT had the vibrato amplitude almost twice as much as the rest of the students ($p<0.01$).

As to experts' recording-based decisions about the students' voice category, the analysis of variance shows significant dependence between the two ($p<0.001$). Confusions may occur when one and the same singer performs in the lower (E minor) and the higher tonality (A minor), and in the former case is classified as a mezzosoprano but in the latter case as a soprano.

DISCUSSION AND CONCLUSIONS

The present data show a tendency for the singer's formant to increase in level as one's vocal studies proceed. At the same time, however, we did not find correlation between experts' ratings of the tone quality and the level of the singer's formant. In many cases, the quality of a student's voice did not meet the standards of a good opera voice. The diffuse peak in the spectrum corresponding to the singer's formant may indicate certain inconsistency in voice production technique.

It must be pointed out that male and female singers must undertake different amount of effort to concentrate energy at the frequency of the singer's formant. Male voices tend to possess weaker fundamental in their glottal signal than female [3]. While in female singing voices the fundamental tends to be matched to the first formant, in male voices there is a match rather between the 2nd or the 3rd harmonic and the first formant. Consequently, there may be problems of matching the singer's formant to a certain harmonic for female voices, since the frequency distance between the harmonics is more sparse there. The strong fundamental may be characteristic to female voices in general: it has been claimed [4] that tenors deliberately bypass the production of a strong fundamental in order to avoid the female quality of their voice.

REFERENCES

[1] Sundberg, J. (1974), "Articulatory interpretation of the "singing formant"," *J. Acoust. Soc. Am.*, vol. 55, pp. 838-844.

[2] Dmitriev, L. and Kiselev, A. (1979), "Relationship between the formant structure of different types of singing voices and the dimension of supraglottal cavities," *Folia Phoniatrica*, vol. 31, pp. 238-241.

[3] Sundberg, J. (1987), *The Science of the Singing Voice*, Illinois: DeKalb.

[4] Titze, I. R., Mapes, S., and Story, B. (1994), "Acoustics of the tenor high voice," *J. Acoust. Soc. Am.*, vol. 95, pp. 1133-1142.

REPRESENTATION OF PROSODIC AND EMOTIONAL FEATURES IN A SPOKEN LANGUAGE DATABASE

Peter Greasley, Jane Setter†*, Mitch Waterman*, Carol Sherrard*, Peter Roach†, Simon Arnfield† and David Horton**

**Department of Psychology, University of Leeds, Leeds, LS2 9JT, UK*

†Department of Linguistic Science, University of Reading, Reading, RG6 2AA, UK

ABSTRACT

This article reports on research in progress on the Emotion in Speech project, funded by the UK ESRC and the Ministry of Defence (project No. R000235285). The ToBI transcription system is used to represent prosodic information. Additional layers have been created to code emotional speech according to the emotional lexicon, affective valence and cognitive appraisals. The aim is to produce a database of fully labelled emotional speech.

INTRODUCTION: THE PROJECT

The study of prosodic function in relation to grammar and syntax is a well-developed field, enjoying formalisms to represent linguistic information which are widely agreed. The prosodic characteristics of *emotional* speech are less tractable, and there is little agreement on the best method to analyse emotional speech or, indeed, how best to represent its specific features.

The Emotion In Speech project will transcribe 20 hours of naturally occurring emotional speech, eventually to be available as a CD-ROM database. The transcription system used to represent prosody is ToBI, devised by Silverman et. al. [1]. As ToBI is a multi-tiered system, we have created additional tiers to incorporate emotional information. The first four layers represent tones (tier one), words (tier two), break indices (tier three) and miscellaneous items, e.g., laughter (tier four). The fourth tier can be used to represent a broader range of prosodic and paralinguistic features than has been used in ToBI so far. Our coding system will be based on the work of Crystal [2]. The remaining four layers contain the emotional labels devised for the project. The aim is to collate multiple characterisations of the emotional response, i.e., prosodic features, judged emotion, lexical valency, appraisal category and cognitive antecedents; analysis should be consistent across levels. The resulting multi-layered database will allow more formal description of emotional speech, and improved understanding of the relationship between situation and communication.

EMOTION IN SPEECH

Due to the practical difficulties in obtaining records of naturally occurring emotional expression, the majority of data on emotion and its vocal correlates comes from studies which have either used actors [3], or have manipulated speech in some way using a computer [4, 5]. While using actors to simulate emotion may have methodological advantages, this approach does raise questions as to the ecological validity of the data. Actor portrayals of emotions may reproduce stereotypes [6] which stress the obvious vocal cues but "miss more subtle cues that further differentiate discrete emotions in natural expression" [7]. Williams & Stevens [8] compared contrived and natural speech and found the overall range of F_0 to be wider in contrived speech.

This approach also suffers from a simplistic and ambiguous labelling method, e.g., 'say x as if you are feeling angry/sad/happy etc.' which 1) neglects individual differences in the interpretation of verbal labels of emotion (the 'basic emotions' like *anger, joy* or *fear* can take different forms depending on the situation [7]), and 2) assumes that emotions are experienced as discrete psychological states whereas, in reality, they predominantly occur in complex blends [9].

It is important, then, that research on the vocal expression of emotion uses samples of speech which are as close to natural speech as possible. In our research, the primary source of data is television and radio (e.g., documentary programmes, talk/debate shows and sports commentary), though we are also seeking other sources. It is also clear that the coding will require a more sophisticated model to represent the varieties of emotional expression.

CODING EMOTIONAL SPEECH

In the psychological literature it is possible to discern three distinct approaches to the categorisation of emotion: 1) emotion labels (the affective lexicon), e.g., anger, rage, grief; 2) abstract dimensions of affect: i) valency (pleasant/unpleasant feelings); ii) potency (strong/weak feelings); iii) level of activity or arousal; 3) cognitive appraisals/antecedents which produce the emotional experience. Each of these approaches is utilised in our coding system.

The first layer (tier five in the modified ToBI system) is being used to record the judgments of listeners. Judges will be given a list of emotion labels from which to select those most appropriate and accurate in describing the emotion(s) expressed. (A number of recent publications provide taxonomies of the emotion lexicon [10]).

The second layer (tier six) is being used to code lexical valence. Osgood, Saporta & Nunnally [11] use a technique referred to as 'evaluative assertion analysis' which involves rating particular words or 'common meaning phrases' according to their valency and intensity. Extremes of valency should increase with emotionality of the source. More recently, Anderson & McMaster [12] and Bestgen [13] have coded the lexical valency of words and phrasal units to determine the emotional tone of literary documents. The occurrence of lexically valenced words, along with their frequency and intensity, provides a measure of the emotional force of the content.

Emotion labels can be notoriously imprecise in representing emotional states. It has been argued (Scherer [7]) that they are "rather unsuitable for the scientific description of affective states," and that "future work should use a conceptual system to describe emotional states and their antecedents that is more systematic than the natural language taxonomy of emotions ... the state referred to should be specified in terms of its underlying process" (p.146). In this respect, Ortony et. al. [14] provide a model of emotional experience based upon valenced reactions to events (pleased/displeased), actions of agents (approve/disapprove), or objects (like/dislike). Whether or not the valenced reaction is actually experienced as an emotion depends upon how intense the reactions are.

This model forms the basis of our third and fourth coding layers (tiers seven and eight). For the third layer we are using this model to identify the general reference category of cognitive appraisals, based on reactions to events, agents or objects. For example, if an event has occurred which the individual is displeased about, we have 'distress emotions'; if the individual disapproves of the actions of another person, we have 'reproach emotions'.

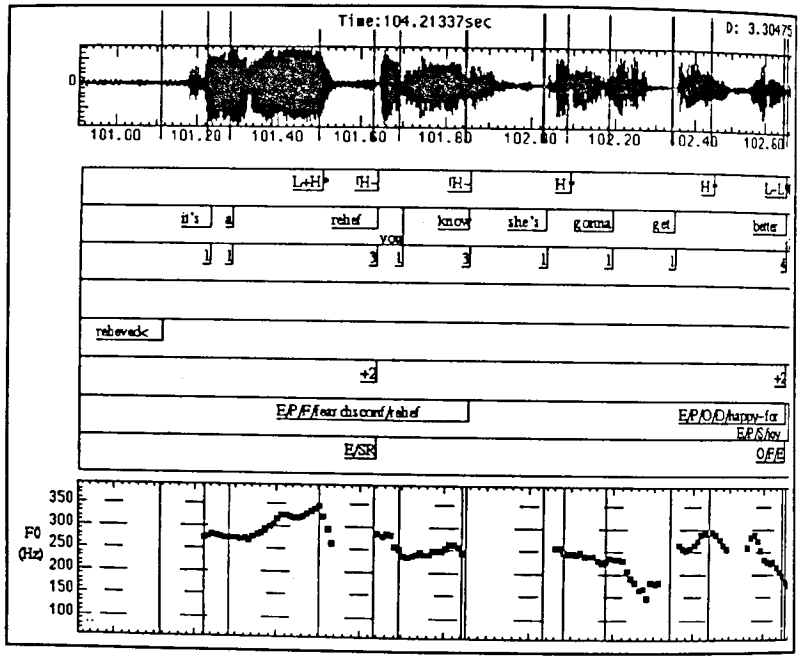


Figure1. Example of the coding system used in the Emotion in Speech project.

The fourth layer (tier eight) is adapted from Waterman's detailed qualitative analysis of interview data [15, 16], based on the theoretical position of Ortony et. al. [14]. From self-reports of emotional responses to emotionally loaded musical extracts, Waterman has devised a classification system to code cognitive antecedents of emotional responses. It can be applied to appraisal statements (i.e., that imply some reference to internalised representations) that need not necessarily be emotional. The purpose of the system is to inform deduction about the type of emotion being experienced by the speaker, and it provides greater detail than the layer above (the general reference category). In practice, application of the coding scheme is exhaustive for all possible appraisal possibilities in a statement. The system does not indicate actual emotion type, only the possible emotions, given the statement under examination.

An example of the coding system is provided in Figure 1. In this example we can see how the four coding tiers combine with the ToBI system to provide a multi-layered representation of emotion in speech. The first coding layer (tier five) provides emotion labels from the affective lexicon provided by judges who have listened to the speech. On the second layer we have two positive word valences: *relief, get better.* The third layer represents the speaker's cognitive appraisals (reactions that elicit the emotional expression), inferred from the speech content. In this case we have the *potential* emotional complex of: 'joy emotions', i.e., pleased about an event (lexical tokens include: delighted, glad, happy, pleased); 'happy-for emotions', i.e., pleased about an event desirable for someone else; and 'relief emotions', i.e., pleased about the disconfirmation of the prospect of an undesirable event. Finally, the fourth layer records every evidenced appraisal contained within the speech content (and thus provides a more detailed and comprehensive record of appraisals than the third, which simply targets particular valenced reactions within the text).

The coding system will, then, provide us with accurate emotion labels, the valenced direction of the emotion along with its intensity, and a sophisticated record of the possible emotions being experienced by the speaker, inferred from the appraisals evident within the verbal content.

REFERENCES

[1] Silverman, L., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J. (1992) ToBI: A standard for labelling English prosody, *Proceedings of the International Conference in Speech and Language Processing*, Alberta.

[2] Crystal, D (1969) *Prosodic Systems and Intonation in English*, Cambridge: Cambridge University Press.

[3] Scherer, K. R., Banse, R., Wallbott, H. G. & Goldbeck, T.(1991) Vocal cues in emotion ecoding and decoding, *Motivation & Emotion*, 15(2), 123-148.

[4] Uldall, E. (1961) Dimensions of meaning in intonation, in Bolinger, D. (ed) (1973) *Intonation*, Harmondsworth: Pengin, 250-259.

[5] Lieberman, P. & Michaels, S. B. (1962) Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech, *Journal of the Acoustical Society of America* 34(7), 922-927.

[6] Kramer, E. (1963) Judgment of personal characteristics and emotions from nonverbal properties of speech, *Psychological Bulletin* 60(4), 408-420.

[7] Scherer, K. R. (1986) Vocal affect expression: A review and a model for future research, *Psychological Bulletin*, 99, 143-165.

[8] Williams, C. E. & Stevens, K. N. (1972) Emotions and speech: some acoustical correlates, *Journal of the Acoustical Society of America* 52, 1238-1250.

[9] Ellsworth, P. C. & Smith, C. A. (1988) From appraisal to emotion: Differences among unpleasant feelings, *Motivation and Emotion*, 12(3), 271-302.

[10] Shaver, P. , Schwartz, J., Kirson, D. & O'Connor, C. (1987) Emotion knowledge: further exploration of a prototype approach, *Journal of Personality and Social Psychology*, 52, 1061-1086.

[11] Osgood, C., Saporta, S., & Nunnally, J. (1956) Evaluative Assertion Analysis, *Litera*, 3, 47-102.

[12] Anderson, C. W., & McMaster, G. E. (1982) Computer assisted modeling of affective tone in written documents, *Computers and the Humanities*, 16, 1-9.

[13] Bestgen, Y. (1994) Can emotional valence in stories be determined from words? *Cognition and Emotion*, 8(1), 21-36.

[14] Ortony, A., Clore, G. & Collins, A. (1988) *The Cognitive Structure of Emotions*, Cambridge: Cambridge University Press.

[15] Waterman, M. G. (1992) Emotion in Music: Towards a new methodology for the investigation of appreciation. *International Journal of Psychology*, 27 (3/4) 189.

[16] Waterman, M. G. (1994) Emotion in Music: Implicit and explicit effects in Listeners and Performers. In I. Deliege (Ed.) *Proceedings of the 3rd International Conference for Music Perception and Cognition.* Liege: ESCOM Publications.

ON THE PERCEPTION OF EMOTIONAL CONTENT IN SPEECH

Anne-Maria Laukkanen, Erkki Vilkman**, Paavo Alku*** and Hanna Oksanen*****

** Institute of Speech Communication and Voice Research, University of Tampere,*

***Department of Otolaryngology and Phoniatrics, University of Oulu,*

****Acoustics Laboratory, Helsinki University of Technology,*

***** Department of Health, Faculty of Medicine, University of Tampere, Finland.*

ABSTRACT

From a nonsense utterance produced expressing neutral state, surprise, sadness, enthusiasm and anger the first 200 ms of the main stress carrying syllable were played at equal loudness to the listeners. Surprise, sadness and anger were especially correctly identified. As the differences in F0 level were artificially eliminated, neutrality, surprise and anger were still identified due to differences in intrasyllabic F0 change and glottal waveform.

INTRODUCTION

The aim of this experiment was to study the perceptual relevance of various speech variables, mainly that of glottal variation in speech, although the elimination of all other variables is in practice very difficult. The speech material was obtained from an earlier study by Laukkanen et al. [1]. There one male and two female subjects produced a nonsense utterance "paappa paappa paappa" simulating five emotional states: neutral, surprise, sadness, enthusiasm and anger. The main stress was given to the underlined syllable. The utterances were 64 - 100 % correctly identified in a listening test. This utterance was chosen since oral pressure during /p/ was used as an estimate of subglottic pressure. Production of emphatic sentence stress simulating various emotional states was regarded as an ideal context to study glottal variation in speech.

The results of the previous study showed that the stressed syllable always had a higher F0 level than the other syllables in the utterance. The emotional states differed from each other in terms of F0 and SP level and change in these parameters as the first and the third (main stress carrying) syllable of the utterance were compared to each other. There was also significant variation in the time based parameters SQ and QOQ [2 and 3, respectively] derived from the acoustically inverse filtered signal. The results from a variance analysis (GLIM) revealed that the glottal variation was more dependent on emotional state than on F0 and SP level alone. Thus the perceptual role of these various parameters needed to be studied further.

MATERIALS AND METHODS

Test 1. The role of syllable length, SPL and intersyllabic F0 and SPL change was eliminated by cutting only the first 200 ms of the main stress carrying (underlined) syllable of the utterance "paappa paappa paappa" to be evaluated. Test 2. The role of F0 level in the samples used in Test 1 was eliminated by artificial pitch modification. In both tests the samples were evaluated by five students of speech science. They answered in a forced choice test, whether the syllables expressed neutrality, surprise, sadness, enthusiasm or anger.

Pitch was artificially modified using a device specially designed for that purpose [4]. The function of the device is based on a digital circuitry for time compression/expansion of the voiced speech segments. The frequency of the voiced segments of a signal is estimated and changed in real time without affecting the signal length. The formant frequencies are changed simultaneously.

Intrasyllabic F0 and A0 (period amplitude) change and the success of pitch modification were studied by calculating F0- and A0 curves from the samples with a microcomputer based signal analysis system ISA (Intelligent Speech Analyser).

Glottal airflow waveform was estimated from the acoustic signal by the IAIF (Iterative Adaptive Inverse Filtering) method [5-6]. Using both synthetic and natural speech the method has been found to yield fairly reliable estimates for voice sources of different fundamental frequencies and phonation types [6].

RESULTS AND DISCUSSION

Figure 1 shows intrasyllabic F0 and A0 changes of the original samples and Figure 2 the results of pitch modification. Figure 3 shows examples of inverse filtered signal waveforms estimated from the part of the syllables where F0 had reached its maximum value.

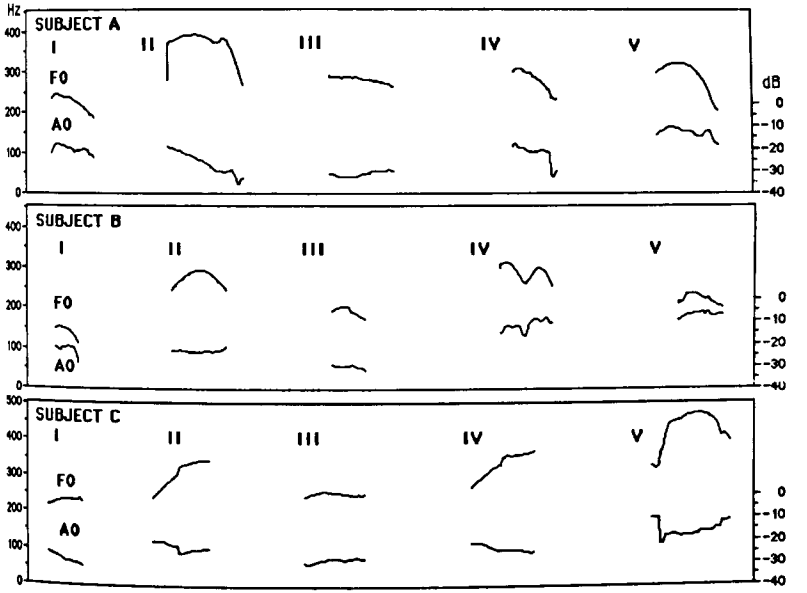


Figure 1. Changes in F0 and A0 during the first 200 ms of the main stress carrying syllable /pa:/. I = neutral state, II = surprise, III = sadness, IV = enthusiasm, V = anger. Subject B is male. Time scale is the same in every sample; differences in the length of the F0/A0-curves is due to the function of the F0 analysis program used.

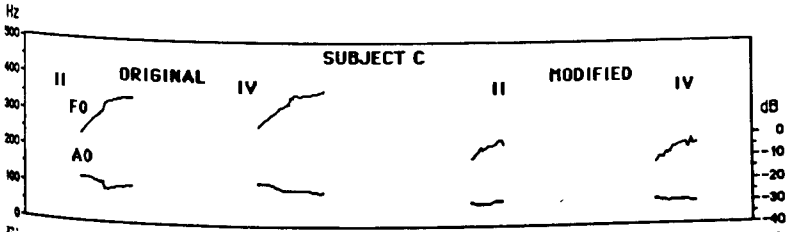


Figure 2. F0 and A0 curves from the original and pitch modified syllables of Subject C (female) expressing (II) surprise and (IV) enthusiasm.

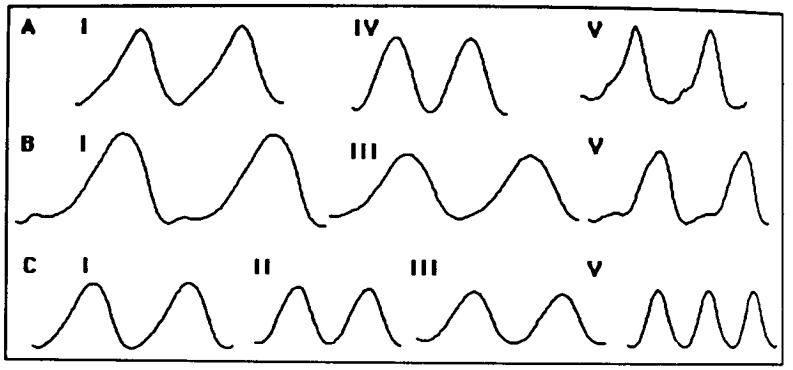


Figure 3. Inverse filtered signal in different samples of Subjects A, B (male) and C. Horizontal axis: time, vertical axis: flow (on an arbitrary scale).

Table 1 shows the results of the listening tests.

Table 1. Number of correct identifications in two listening tests (1. and 2.). Number of listeners = 5. In the first test the first 200 ms of the main stress carrying syllable /paa/ was used. For the second test the same samples were artificially modified in pitch so that all F0 differences between the samples were eliminated. I = neutral state, II = surprise, III = sadness, IV = enthusiasm, V = anger. - = the sample could not be modified.

	Subject		
emotion/test	A	B (male)	C
I 1.	4/5	2/5	3/5
2.	4/5	0	4/5
II 1.	5/5	5/5	4/5
2.	1/5	0	5/5
III 1.	1/5	4/5	4/5
2.	0	3/5	1/5
IV 1.	3/5	0	0
2.	1/5	0	0
V 1.	5/5	5/5	3/5
2.	3/5	5/5	-

In the first listening test the misidentifications were mainly due to the fact that the neutral and sad types were confused, likewise surprise and enthusiasm or enthusiasm and anger. The misidentifications might be based on F0 level, which was lowest in neutral and sad samples and highest in the other samples (Fig. 1). In the second test the number of correct identifications naturally dropped. The neutral samples and those expressing anger were identified best. The misidentifications seemed to emanate from similarities in F0 contour: For Subject A neutral, sad and enthusiastic samples with a falling F0 contour were all identified as neutral. Surprised samples of Subjects A and B as well as the enthusiastic sample of Subject B were misidentified as angry. This may be attributable to the strongly and quickly rising-falling F0 contours in the samples. For Subject B the sad sample was misidentified as surprised, possibly because of the slightly rising contour at the beginning of the sample. This, in turn, may be related to the fact that Subject B seemed to express more compassion than sorrow. For Subject C the enthusiastic sample was misidentified as surprised, most probably because of the similarly rising F0 contours. For Subject C the angry sample could not be pitch modified because of the very high F0 and very pressed voice quality in the original sample.

The subjects differed from each other in their strategies for expression of emotions. For example the fact that in the case of Subjects A and B the number of correct identifications for surprise dropped drastically through pitch modification but in the case of Subject C contrastively the sample was 100 % correctly identified despite the modification suggests that Subjects A and B expressed surprise especially through F0 level but Subject C in contrast through intrasyllabic F0 contour.

A0 changes did not seem to have any independent role in expression, they merely reflected changes in F0.

The most serious disadvantage in pitch modification using this procedure is the fact that in some cases the voice quality becomes unnatural, since the formant frequencies become changed together with F0 and since the digital time manipulation causes some discontinuity of the signal sometimes leading to a slightly clattering sound. However, the results obtained in this experiment were logical suggesting that the quality in the samples was not too much distorted. Furthermore, the pitch modification related formant alteration can be regarded as positive since it eliminates the possible role of formant frequencies in the expression of emotions.

From the bases of this experiment no conclusions can be drawn on the role of the pure glottal airflow waveform in the identification of emotions in speech. The intrasyllabic F0 contour seems to have great perceptual relevance; however, F0 contour naturally also includes dynamic glottal waveform changes. The observations made in this experiment would suggest that voice quality in terms of glottal waveform also had perceptual relevance: In the case of Subject A the enthusiastic sample was by some listeners misidentified as sad, which may be due to the very breathy voice quality in that sample. In the case of Subject C the sad sample was confused almost without exception with the neutral sample, which may be due to the very similar waveforms in both cases (in the period of F0 maximum). In the case of Subject B the F0 contours in the sad and angry samples were fairly similar; however the samples were not perceptually confused, which most likely can be explained by the very different waveforms in them (Fig. 3). However, since the glottal waveform was studied only at F0 maximum, it remains uncertain, whether the possible relevance of voice quality in these samples was related to the glottal waveform in general or to a pitch synchronous change in it. Anyway, the results suggest that also in terms of glottal variation there are individual differences in the strategies for expressing emotions: In these samples Subject C seemed to use less glottal variation and express more through F0 than Subjects A and B (see Figure 3). The role of the glottal waveform in conveying emotions needs to be studied further.

REFERENCES

[1] Laukkanen, A.-M., Vilkman, E., Alku, P., Oksanen, H. (1995). A preliminary study on stress production related physical variations in utterances signalling different emotional states. Submitted for publication.

[2] Timcke, R., von Leden, H. & Moore, P. (1958) Laryngeal vibrations: measurements of the glottic wave. *AMA Arch Otolaryngol.*, 68, 1-19.

[3] Hacki, T. (1989) Klassifizierung von Glottisdysfunktionen mit Hilfe der Elektroglottographie. *Folia phoniatrica* 41, 43-48.

[4] Viitanen J (1982) [Puheäänen korkeuden muuttaminen reaaliajassa]. In Finnish. (Real-time pitch modifying of human voice. Abstract in English). *Folia fennistica & Linguistica*, X Fonetiikan päivät Tampereella 20.-21.3.1981 (Papers from the Xth Meeting of Finnish Phoneticians). Tampereen yliopiston suomen kielen ja yleisen kielitieteen laitoksen julkaisuja 7, Tampere, pp. 323-328.

[5] Alku, P., Vilkman, E. & Laine, U. K. (1991) Analysis of glottal waveform in different phonation types using the new IAIF-method. *Proc. 12th International Congress of Phonetic Sciences, Vol. 4, Aix-en-Provence, 19-24 Aug. 1991*, pp. 362-365.

[6] Alku, P. (1992) Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering. *Speech Communication* , 11, 109-118.

PROSODIC SIGNS OF EMOTION IN SPEECH: PRELIMINARY RESULTS FROM A NEW TECHNIQUE FOR AUTOMATIC STATISTICAL ANALYSIS

S. McGilloway, R. Cowie & E. Douglas-Cowie
School of Psychology / School of English, Queen's University, Belfast, UK

ABSTRACT

Emotional expression has been studied using a technique in which automatic preprocessing extracts key speech features and statistical descriptions of them are generated. Five broad types of marker are found to distinguish among passages - spectral balance, range of pitch movement, timing of pitch movement, timing of intensity changes, and intensity distribution.

INTRODUCTION

There are many reasons for trying to analyse the vocal expression of emotion in objective and quantitative terms. It is a natural extension of research on machine speech to explore methods of recognising and generating speech which is not emotionally neutral [1]. Social and behavioural studies could benefit from reliable methods of measuring vocal signs related to emotion. Clinical applications also exist. Our own interest in the area stems from one of these. Lack of emotional expression in speech is diagnostically important in schizophrenia [2], but the absence of formal measures in the area impedes refinement and evaluation of diagnostic practices, as well as leaving critical decisions to depend on the ear of someone who need have no particular aptitude for phonetics. We studied markers of emotion in normal speech as a necessary part of a project concerned with that clinical problem.

The study uses a system called ASSESS which is described more fully elsewhere in these Proceedings [3]. It extends earlier work on the statistical description of speech [4]. The main innovation is that statistical description is preceded by preprocessing which extracts key features of the speech signal and simple units associated with them. ASSESS then generates an array of approximately 400 statistics specifying the attributes and variations of these features.

METHOD

Speech sample

Passages were constructed to suggest four emotions - fear, anger, sadness, and happiness. A fifth, emotionally neutral passage was used as a baseline. The passages were of comparable lengths, taking about 25-30 seconds each to read.

Speakers were 40 volunteers from the Belfast area, 20 male and 20 female, aged between 18 and 69. There was a broad distribution of social status, and accents represented a range of local types.

Subjects familiarised themselves with the passages first and then read them aloud using the emotional expression they felt was appropriate. They were presented in computer-generated random orders.

Recordings were digitised using a CED 1401 signal capture system. Sampling was at 20kHz, after low pass filtering at 10kHz. System limitations meant that files had to be entered in sections of 7 seconds or less and rejoined at a later stage of processing. Splits were placed by hand within substantial pauses.

Acoustic analysis

ASSESS is based on standard descriptors: the speech spectrum; the intensity contour; and the pitch contour. It breaks these up into significant units using techniques chosen for robustness rather than elegance or precision, otherwise hand correction would be essential. Contours are smoothed before finding inflections, i.e. points at which volume or pitch stops rising and starts falling, or vice versa. Rises, falls, and plateaux (periods of relatively flat pitch or intensity flanking an inflection) are then found. Spectral information is used to identify transitions which mark at least roughly natural units. Four types are considered - silences, sound blocks, tunes, and fricative bursts. Sound blocks are defined by the way intensity rises after a silence, peaks, and falls to the next silence. Tunes are defined by the way pitch rises and falls between silences long enough to be considered pauses. Fricative bursts are defined by the distribution and amount of energy high in the spectrum. Subspectra are formed for special types of episode, fricative busts and peaks in the intensity contour.

Properties and relationships of these units are summarised in a battery of statistics, primarily measures of midpoint and spread, generally in parametric and nonparametric forms (the latter are less sensitive to occasional erratic data points). Lines and curves are also fitted to specify the shape of tunes and spectra.

ASSESS can estimate absolute intensity by using a calibration signal with a known dB level. However in this study no absolute referent was available, and level was normalised by treating the first file in a passage as a referent and setting its median intensity at 60dB. This seems unlikely to confound the results.

Statistical analysis

The basic statistical procedure was analysis of variance followed by post hoc comparison of group means using Duncan's range test. Characteristics were considered distinctive only if the overall analysis of variance was significant with $p<0.05$ and the emotional passage in question contrasted with the neutral passage with $p<0.05$. Passage was treated as a between groups variable for both tests. This is conservative, i.e. it is likely to conceal real differences rather than generate spurious ones. Sex was considered as a third variable in analyses involving pitch height, because otherwise variance is inflated by sex differences and effects of emotion are swamped.

RESULTS

There was wide range of differences between passages - over 1/3 of the measures considered yielded significant differences. The challenge is to reduce these to a manageable set.

The largest set of differences reflect an effect that distinguishes two broad groups of passages: afraid, angry and happy on one hand, sad and neutral on the other. They involve intensity contrasts. It seems apt to call the groups intensity marked and intensity unmarked respectively.

Table 1 shows the main features of the effect. Measures are in bold face if they are significantly different from the neutral passage. The first two columns show intensity measures for all points outside pauses. These global measures are higher for fear, anger and happiness than for sad and neutral passages. However, intensity marking is not a simple matter of loudness. ASSESS reveals two types of structure in it.

Table 1 : Selected intensity contrasts between groups (normalised scale).

	mean	median	peaks	troughs
Anger	**64.11**	**61.57**	**66.87**	**59.97**
Fear	**63.64**	**61.51**	**66.45**	59.57
Happiness	**63.38**	**61.59**	**66.07**	59.52
Sadness	62.42	60.32	65.10	59.12
Neutral	62.33	60.73	64.87	58.83
p	0.000	0.003	0.000	0.095

First, note that intensity is normalised. Hence the first two columns do not mean that the first three emotions are associated with louder speech: it means that intensity rises after the first few phrases. This may be called a crescendo effect.

Second, note that the effect is more marked with means than with medians. That suggests it involves stretching in the top end of the intensity distribution rather than just a global upward shift. The inference is confirmed by the last two columns. The contrast in the level of peaks in the intensity contour is even more marked than the contrast in overall mean. However, there is much less contrast in the level of troughs (that is, minima). It is not significant overall, and the Duncan test shows only anger differs significantly from the neutral passage.

There is a trend for silences to be longer in the intensity marked passages, which is just short of significance ($p=0.051$) and most marked in anger. This is consistent with the general pattern of heightened dynamic contrast in the intensity marked passages.

Several other features distinguish intensity marked passages from the neutral passage, and to a greater or lesser extent distinguish them from each other.

Properties involving the duration of intensity features may tend to signal negative emotions: they do not affect happiness, and they may affect sadness. Table 2 summarises the data.

The durations of amplitude movements distinguish fear and anger from the neutral passage again. Both have longer median durations for both falls and rises. But in contrast to the crescendo and intensity stretching effects, this effect is stronger in fear than anger. Protracted intensity falls also characterise sadness. The durations of tunes show a similar pattern. Also broadly similar is a property of intensity plateaux. The interquartile range of their duration increases markedly in fear and sadness, and less so in anger.

Table 2: Aspects of duration associated with negative emotions (times in ms).

	rises median	falls median	tunes mean	plateau IQR
Fear	**82.35**	**84.8**	**1265**	**10.8**
Anger	**81.66**	**80.5**	**1252**	10.2
Happiness	78.03	77.4	1404	8.2
Neutral	78.50	77.2	1452	8.4
Sadness	77.28	**81.4**	**1179**	**11.0**
p	0.000	0.000	0.001	0.006

The passages differ in the distribution of energy across the spectrum, but few of the effects are easy to interpret.

Most straightforwardly, all the emotions are characterised by greater variability in the duration of fricative bursts (as measured by the standard deviation) than the neutral passage.

A second clear effect involves anger. Here the average spectrum for non-fricative portions of speech has a high midpoint. This is not surprising: it parallels a well-known effect of tension on spectral balance [5]. Conversely, the sad passage gives a significantly lower spectral midpoint than any of the intensity marked passages - it is lower even than the neutral passage.

Fricative bursts are associated with a number of effects which seem paradoxical at first sight. Anger is associated with high average energy in fricative bursts, but the average spectrum for slices classed as fricative has a low mean and a markedly negative slope. The implication appears to be that the intensity associated with frication is not rising as fast as the intensity associated with the lower spectrum. Fear and happiness are distinctive in terms of the subspectrum which shows variability in slices classed as fricative. These too show markedly negative slopes, indicating relatively low variability in the regions associated with frication. The effects may be less to do with frication than with raised variability in the lower spectrum.

Two aspects of the pitch contour show differences - the distribution of pitch height and the timing of pitch movement.

Passages do not differ significantly in pitch height per se. However, they do differ in its distribution. Again, the differences which are clearly significant fall into an orderly pattern. All of them involve interquartile intervals, which can be thought of as measures of the range a measure usually occupies. When all pitch inflections are considered together, the passage difference in interquartile interval just reaches significance ($F\ 4, 185 = 2.91$, $p=0.023$). Separating maxima and minima shows a weak passage effect for minima ($F4, 185=2.74$, $p=0.03$) and a much stronger one for maxima ($F4, 185 = 3.76$, $p=0.006$). In all three cases, range is widest for happiness and nearly as wide for anger, with the lowest range in the neutral passage.

All the distinctive pitch duration features are associated with happiness. Pitch plateaux are shorter in the happy passages than elsewhere, and their durations generally lie within a narrower range (as measured by the inter quartile range). Conversely, pitch falls last longer in the happy passage than in the neutral one. This feature is shared with the sad passage. Pitch rises are also significantly faster in the happy passage than in the neutral passage. The overall picture is that happiness involves pitch movement which is not only wide, but constant.

The outline of the findings can be summarised in a table. This shows that each of the passages is distinguished in several ways from any other.

Table 3: Summary of distinctions among passages

	Afraid	Angry	Happy	Sad
Intensity				
• marking	+	+	+	
• duration	+	+		+
Spectrum				-
• midpt & slope		+		
Pitch movement				
• range		+	+	
• timing			+	+

DISCUSSION

It has been pointed out that the corpus of research on emotion in speech is not large, and studies tend to agree only partially among themselves [6]. Our main claim for this study is that it demonstrates the potential of an approach which may be relevant to those problems.

One reason for not drawing stronger conclusions lies in our speech sample. The passages do generally convey the emotions that they are meant to, but they are of simulations of emotion, made by people who have no expertise in simulating it. An obvious need is to obtain samples of genuinely emotional speech. That presents both practical and ethical difficulties, and it would also aggravate a problem which is present in this study, which is to distinguish effects due to linguistic content from effects due to emotion. It seems unlikely that the strongest features we have noted are due to linguistic content, but the possibility should be acknowledged.

With that qualification, our data make a simple point. Statistics which can be extracted automatically and conveniently do seem to distinguish emotional speech episodes. They include statistics of a higher order than the global measures of mean and range which have been in use for half a century [7],[8], and it seems likely that distinctions can be sharpened by using these higher order measures.

A significant theoretical attraction of reducing description at this level to numbers routinely extracted is that it frees us to explore pattern at a different level. We have noted that our passages are distinguished by feature combinations, and share features with each other. It is a natural extension to conjecture that different expressions of the same emotion bear similar relationships, with some features in common, but not all. This suggests a geometric picture which is familiar in research on automatic classification: emotions may be thought of as regions in a multidimensional space where points (corresponding to episodes of speech) are positioned by the strength of various attributes.

Theory apart, our approach has an obvious practical attraction. It points quite directly towards automatic methods of recognising emotion in speech. It seems clear that there are rather complex linguistic cues to emotion in speech [9], and capturing them automatically remains a long term project. However, using essentially simple statistical techniques seems a reasonably immediate prospect.

ACKNOWLEDGEMENT is due to Drs D Sykes and C Cooper for statistical advice.

REFERENCES

[1] Murray, I., Arnott, J. and Newell, F. (1988), "Hamlet: simulating emotion in synthetic speech", *Proc. 7th FASE Symposium*, Edinburgh, pp. 1217-1223.

[2] Andreasen, N., Alphert, M. and merrill, J. (1981), "Acoustic analysis: an objective measure of affective flattening", *Arch. Gen. Psychiatry*, vol 38, 281-285.

[3] Cowie, R., Sawey, M. and Douglas-Cowie, E. (1995), "A new speech analysis system: ASSESS (Automatic Statistical Summary of Elementary Speech Structures)", *Proc.13th International Congress of Phonetic Sciences*, Stockholm.

[4] Cowie, R., Douglas-Cowie, E. and Rahilly, J. (1991), "Instrumental measures of abnormalities in deafened speech", *Proc 12th International Congress of Phonetic Sciences*, Aix-en-Provence, pp. 350-353.

[5] Hammarberg, B., Fritzell, B., Gauffin, J., Sundberg, J., & Wedin, L. "Perceptual and acoustic correlates of voice qualities" *Acta Otolaryngologica* vol. 90, pp. 441-451.

[6] Murray, I. and Arnott, J. (1993), "Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion", *J. Acoust. Soc. Am.*, vol. 93 (2), pp. 1097-1108.

[7] Skinner, E. (1935), " A calibrated recording and analysis of the pitch force and quality of vocal tones expressing happiness and sadness", *Speech Monog.*, vol. 2, pp. 81-137.

[8] Fairbanks, G. and Pronovost, W. (1939), An experimental study of the pitch characteristics of the voice during the expression of emotion", *Speech Monog.*, vol. 6, pp. 87-104.

[9] Ladd, R., Silverman, K., Tolkmitt, F., Bergmann, G. and Scherer, K. (1985), "Evidence for the independent function of intonation contour type, voice quality, and F0 range in signalling speaker affect", *J. Acoust. Soc. Am.*, vol. 78 (2), pp. 435-443.

PRAGMATIC PHONETICS: ACOUSTIC CORRELATES

Katherine Morton
University of Essex, Colchester, U.K.

ABSTRACT

Within the general theory of linguistics, pragmatics is concerned with describing the intentions, attitudes and beliefs of the speaker. Pragmatic phonetics itself is about how speech production interprets the requirement to communicate pragmatically determined effects, and about how the perceptual system is triggered by the acoustic signal to invoke the appropriate reaction in the listener. This paper examines the role of the acoustic signal in this complex chain of processes, and discusses how the system might usefully be modelled.

SPEECH PRODUCTION THEORY

Phonetics has been primarily concerned with modelling the physical processes of speech production. Speech production is usually associated with motor, aerodynamic and acoustic processes. But phonetics also models speech perception, involving physical and cognitive processes.

Modern theories of speech production blur the distinction between cognitive and physical processes [1] [2]. For example, they do this in different ways: Articulatory Phonology uses the gestural score formalism to represent requirements in both planning and execution: Cognitive Phonetics introduces cognitively driven supervision of motor processes to explain how some of the universal physical effects of speech vary in a linguistically sensitive way.

The rigid distinction between phonology and phonetics made it difficult to understand the effects of these phenomena. Integration of cognitive and physical descriptions at both the production and perception levels of speech modelling is essential if we want to examine pragmatic phenomena in speech.

PRAGMATIC PHONETICS

Pragmatics can be thought of as an extension of semantics [3]; linguistic semantics describes meaning and within semantics, pragmatics attempts to explain the interpretation of meaning in terms of attitudes and belief structures. Pragmatic phonetics models how the set of beliefs and intentions available to human beings become part of spoken language. It is concerned with the expression and interpretation (or production and perception) of intention, attitude and belief, where these properties of the language are not directly expressed by choice of words or word order in sentences, but by *how the utterance is said.*

This claim means that there must be different ways of speaking a particular sentence, and that the resulting acoustic signals will trigger in the listener an awareness of the emotion, attitude or belief which the speaker may be communicating [4].

Additionally this communication may not be voluntary, that is, under the conscious control of the speaker. For example, a speaker might be so angry as to be unable to suppress communicating that anger though tone of voice, or so happy that that emotion cannot be suppressed.

Pragmatic phonetics is therefore about triggering a listener response to stimuli over and above the usual phonological and phonetic content of utterances [5] [6].

TONE OF VOICE

Pragmatic Phonetics is being developed for two reasons:

- to characterize and explain pragmatically derived effects in speech production and perception, and
- to simulate the acoustic effects using speech synthesis.

It will be useful to have synthetic speech able to convey an added dimension of naturalness [7], but the simulation is also being developed to test the model itself. This rests on the assumption that it is possible to capture the tone of voice which triggers effects in the perceiver, and that the information is in the acoustic signal. Two consequences of this model are

- we can link semantics and phonetics;
- we can model the humanness of dialogue.

Perceived variations in tone of voice should be obvious and detectable as departures in the acoustic waveform from an expected norm. It should be possible in principle to identify and quantify these changes. But there is considerable variability in speech waveforms and it has proved difficult to separate out the variations associated with conveying pragmatic effects from other variability present in the waveform [8], introduced by properties of the vocal tract and articulators.

It is essential now to develop a computationally oriented model for synthesis. In modelling spoken communication for dialogue systems, it is useful to distinguish between two types of independently varying and independently sourced tone which produce different types of pragmatic effect. This choice enables the explicit execution in speech of pragmatic markers in speech. For a computational model to operate, this arrangement requires a hierarchical rather than linear organization.

1. Global tone of voice

Tone of voice at the global level characterizes what is appropriate for the overall dialogue situation. Here are some examples from human/machine dialogue situations:

- In an inquiry system about the weather the informant would ideally sound friendly and confident of the facts.
- In a situation warning of emergency the speaker needs to be simultaneously firm, confident and reassuring.
- In a dialogue as part of a computer assisted teaching programme, the listener should be made to feel that the speaker is being sympathetic as well as instructive.
- In an aircraft cockpit dialogue information system the pilot would expect the synthetic speech heard to be confident, clear and sometimes urgent, but never sympathetic or admonishing.

Global tones form the background tier for the pragmatic phonetic model. At this point global tones such as those expressing anger or happiness can be modelled. But more subtle attitudes, such as firmness and confidence, might well need to be modelled as the dialogue unfolds. In human dialogue there is a requirement to respond to pragmatic changes; the listener's perception varies as the context develops.

These changes can be characterized in another level superimposed on the global tone of voice, called local tone of voice.

2. Local tone of voice

Local tone of voice varies according to specific short term requirements during the unfolding dialogue. A speaker might be aware of a listener's changing levels of understanding while something is being explained and react accordingly with short term changes of style. As a specific example:

- A teacher needs to sound firm yet patient during the short term explanation of some point within a wider context.
- In the aircraft cockpit the computer's global firm and confident tone might be modified by encouraging and patient instructions if the pilot fails to understand an explanation or course of action.

3. The overall model

Tone of voice execution is modelled as a layered process. Execution begins with a neutral tone which might never be acoustically realized — an abstract representation of tone. This is the tone of 'neutral' phonology or the tone of a synthesis system implementing only a basic

language model, and is intended only for conveying plain messages.

Global tone is a specific long term modification of the abstract neutral tone. It is contextually determined by general pragmatic considerations deriving from the speaker.

Local tone comprizes specific short term *overlays* on global tone. It is contextually determined either by the changing nature of the semantics or pragmatics being communicated or by feedback concerning listener reaction. Local tone is superimposed on the global tone as the dialogue develops.

Both types of tone are generated by markers arising within the language model framework. Global markers are generated initially are only exceptionally updated, whereas local markers are repeatedly generated, updated or changed.

This framework is intended to relate observations of pragmatic effects in speech, to provide for a source for these effects (the pragmatic component in the language model), and will eventually set out the production and perceptual processes involved.

THE ACOUSTIC DATA

The acoustic data relating to pragmatic effects in speech is extremely difficult to obtain. It is not the intention of this paper to list acoustic correlates of particular pragmatic effects, but under this heading to account for some of the difficulties researchers face.

The biggest problem facing analysis of the acoustic signal is noise, that is, unwanted speech signal — not background noise against which the waveform is heard. The point here is that the variations imposed on the speech signal by pragmatic effects are buried in the natural variability associated with speech signals. Unfortunately it is not obvious which particular variation on any one occasion derives from the pragmatic marker — variability from many different sources is a basic characterisation of speech.

The problem is knowing what aspects of the variability are generated by pragmatically derived intentions and what aspects result from other sources. It was for this reason, for example, that data reduction was attempted using an artificial neural network paradigm [9] and a two-parameter (f0 and syllable durations)model to determine the associative relationship between pragmatic markers, abstract prosodic representations and an acoustic signal judged by listeners to evoke the required perceptual response. Despite the fact that neural networks are particularly good at tasks of this kind the results were disappointing: variations between speakers still became a confusing element in describing the acoustic signal.

Eskenazi [4] used a traditional technique of firstly selecting eight acoustic parameters (overall intensity, f0 maximum, dynamic range of f0, number of pauses, speaking rate, amount of phonological changes, F1/F2 shift and the amount of stop bursts) and then measuring them, but concluded that individual speakers expressed speech styles in different ways, and that not all parameters were equally used by all speakers. This phenomenon is also commented on by O'Shaughnessy [10], who emphasized that the mapping between physical acoustics and perceived prosody is not one to one.

Many researchers have tried hard to determine the acoustic correlates of pragmatic effects, and some have attempted to incorporate this information in their synthesis. So far, though, there has been little success in adequately unambiguously capturing subtle global effects like 'firmness' or providing sufficient contextual information to enable the automatic triggering of local effects.

CONCLUSION

The model most of us currently use assumes that the problem is to generalize acoustic cues from the waveform information. The listener is seen as responding to the cues. But if looked at from the listener's point of view he/she is supplying information in order to decode the signal. The interpretation of the signal will vary, but so will the acoustic cues responsible for triggering the percept.

If we wish to simulate the dialogue context either for practical purposes or to test the model we might well look more closely at modelling the listener, and how the listener may anticipate cues from knowledge of the dialogue context as it unfolds.

If the acoustic signal has within it cues for triggering an appropriate perceptual response to the pragmatics of the spoken utterance then that signal shows large variability and useful information is buried in variable noise.

Rather than model the process as heavily dependent on these cues, it is becoming necessary to shift the focus of the model away from the cues, leaving them as minimal and variable, and move toward a compensatorily oriented peceptual model. In traditional terms, we would assume a knowledge based system very heavily dependent on a full and accurate representation of the effects of various types of variability in the acoustic signal.

For the moment we cannot do this, but it is to be hoped that the adopting a model *framework* along the lines of what is suggested here might advance the situation. At the moment it is discouraging to measure acoustic data without some improvements to the framework within which that measurement is carried out.

REFERENCES

[1] Browman, C.P. and Goldstein, L. (1986), 'Towards an articulatory phonology', in C. Ewan and J. Anderson (eds.), *Phonology Yearbook 3*, Cambridge: Cambridge University Press, pp. 219–252.

[2] Tatham, M.A.A. (1986), 'Towards a cognitive phonetics', *Journal of Phonetics*, Vol. 12, pp. 37–47.

[3] Levinson, S.C. (1983), *Pragmatics*, Cambridge: Cambridge University Press.

[4] Eskenazi, M. (1992), 'Changing speech styles: strategies in read speech and casual and careful spontaneous speech', *Proceedings of the International Conference on Spoken Language Processing*, Banff, pp. 755–759.

[5] Bladon, A., Carlson, R., Granstrom, B., Hunnicutt, S. and Karlsson, I. (1987), 'A text-to-speech system for British English, and issues of dialect and style', *Proceedings of the European Conference on Speech Technology*, Edinburgh, pp. 55–58.

[6] Morton, K. (1993), 'Speech synthesis in dialogue systems', *Proceedings of Eurospeech '93*, Vol. 2, Berlin, pp. 905–908

[7] Granstrom, B. and Nord, L. (1992), 'Neglected dimensions in speech synthesis', *Speech Communication*, Vol. 11, pp. 347–356.

[8] Engstrand, O. (1992), 'Systemacity of phonetic variation in natural discourse', *Speech Communication*, Vol. 11, pp. 337–346.

[9] Morton, K. (1992), 'Pragmatic phonetics', in W.A. Ainsworth (ed.) *Advances in Speech, Hearing and Language Processing*, Vol. 2, London: JAI Press, pp. 17–53.

[10] O'Shaughnessy, D. (1987), *Speech Communication: Human and Machine*, Reading, Mass.: Addison-Wesley.

INTONATIONAL UNIVERSALS IN TEXTUAL CONTEXT

E. A. Nushikyan
Odessa State University, Odessa, Ukraine

ABSTRACT

This paper reports the results of an experimental research of universals in emotional speech. Intonational universals were studied at a textual level in reference to general emotional colouring and expression of particular emotions. Intonational universals were also analysed in a new aspect-communicative orientation of emotional texts.

INTRODUCTION

For some decades contrastive studies have gained much attention in linguistics. Contrastive studies of systems and functions are particularly useful when languages with different structures are compared. The results reported in literature show that on the one hand languages can differ from each other without limit and in unpredictable ways, and on the other hand, "The existence of deep seated formal universals, ... implies that all languages are cut to the same pattern" (Chomsky N., 1965). Thus among its other goals, contrastive typology is largely concerned with revealing linguistic universals. Various proposals have been put forward as to what constitutes universals. Many of them have taken the form and function of the rules of grammar.

The universals that have been studied best in phonetics refer to phone-mic systems. Studies of prosodic universals are relatively scarce, although intonation systems manifest more universal features than other linguistic categories.

Available descriptions of intonational typology disregard its emotional aspect, despite the fact emotionally coloured speech contains more universals than neutral speech; this evidently should be explained by the common biological nature of emotions. "I hypothesized that of the parts of the human vocal system that are used linguistically intonation respond more closely than any other to states of organism. ... there are tendencies in the repetition of intonational forms in the most widely separated languages" (Bolinger D., 1980).

The aim of this paper is to reveal universals in emotional speech intonation at a textual level.

SPEECH MATERIAL AND SUBJECTS

An experimental study of textual prosody was carried out on emotional and corresponding neutral texts recorded by 30 native speakers of English, German, French, Russian and Ukrainian. Whereas our previous investigation of emotional speech (Nushikyan,1987) was based on 3 languages - English, Russian and Ukrainian, this one involves 2 more languages - German and French. Here attention has been spread from the single utterance to the longer units - texts. These texts expressed 16 positive and negative emotions (Nushikyan,1986). The speech signal was instrumentally analysed by Visi-Pitch, IBM speech program, which enabled graphical presentation of fundamental frequency and intensity. Spectrograms were made on Sona-Graph of the Kay Elemetrics Corporation using the wide band filter (300 Hz).

Emotional speech prosody is described as a complex of acoustic features that includes features of melody, intensity, duration and spectrum. The description was made both for a whole text and a separate utterance. Investigation of textual prosody provides a deeper analysis of its intonation structure and reveals differences in integration of emotional tension.

Prosodic features and intonation patterns of utterances in a text were analysed and compared on a set of structurally important syllables: the unstressed preceding the first stressed, the first stressed, the nuclear, and the unstressed post-nuclear syllables. All acoustic features were analysed and compared in relative units in order to level individual differences and put together data obtained from 30 different speakers.

DATA ANALYSIS AND RESULTS

Intonational universals in emotional speech were studied in reference to general emotional colouring and expression of particular emotions.

Comparison between emotional and neutral texts revealed universal features of general emotional colouring. Thus, in all the five languages emotional texts differed from neutral in variations of fundamental frequency intervals and range, velocity of fundamental frequency changes, energy of the whole phrase and of its nuclear syllable, mean syllable duration, and, more conspicuously, nuclear syllable duration.

Speech prosody of particular emotions in the five languages was also characterized by some common features. They were: higher fundamental frequency, greater intensity and longer duration, along with wider F_2 , F_3 and F_4 bands and a more complicated structure of their harmonics, the greater role of the high frequency noise regions of consonants in texts expressing **anger, indignation** and **threat** ; wider formant bands of F_2 ,F_3 , F_4, the shift of the intensity of formant frequencies of semantically important words into higher regions, smaller formant energy of unstressed syllables, greater role of high frequency noise regions of affricates, longer duration and greater intensity of key and thematic words, higher fundamental frequency pitch in all structural parts of communicatively strong utterances of texts expressing **delight, joy, admiration** ; longer duration of all structural parts of an utterance, lower intensity level, smaller formant energy of unstressed syllables, and wider F_3 band, greater role of high frequency noise regions of affricates, lower frequency pitch in texts expressing **sadness** ; higher fundamental frequency pitch of all the structural parts of an utterance and the decrease of their intensity, smaller formant energy of F_1 and F_2 , lower frequencies of F_3 in texts involving **surprise** ; etc.

Some common features in speech prosody of particular emotions can also be observed through the study of their intensity.

The acoustic analysis of the intensity difference between emotional and neutral texts reveals a significantly greater total energy of most emotional texts.

The ratio of the total energy of emotional and neutral texts proves to be universal in the prosodic structuring of emotional speech. The quantity of this ratio depends on the type of emotion expressed in the text (see Table 1).

Table 1. The ratio of total energy of emotional and neutral texts.

Emotions expressed in the text	Ratio of total energy				
	English	German	French	Russian	Ukrainian
joy	1,65	1,71	1,58	1,26	1.37
sorrow	0,86	0,92	0,83	0,76	0,87
anger	1,78	1,86	1,65	1,68	1,49
fear	0,75	0,91	0,74	0,86	0,89
despair	1,19	1,31	1,29	1,28	1,26
threat	2,23	1,96	1,84	1,81	1,69
surprise	1,11	1,19	1,09	1,12	1,14
shame	0,95	0,98	0,84	0,79	0,81
offence	1,55	1,51	1,31	1,24	1,16
contempt	1,83	1,87	1,61	1,42	1,36
suspicion	1,56	1,58	1,41	1,02	1,09
irony	1,56	1,65	1,54	1,42	1,51

The data of the table show that such emotions as **anger, contempt, threat, irony** greatly increase the total textual energy in all the studied languages. In English and German a significant increase of energy is observed in texts expressing offence and suspicion. In Russian and Ukrainian the increase of total energy is not so great. The decrease of

textual energy is observed in texts expressing **sorrow, fear, shame**.

Research based on acoustic data presents evidence that for languages under study variations in speech rate are mainly due to variations of pauses and the type of speech and emotion that is being used.

Identical and non-identical elements of prosodic systems of different languages can also be established from phonetic division of texts pronounced emotionally and neutrally.

Extensive experimental data from the five languages shows that phonetic division boundaries mostly coincide with syntactic division boundaries both in neutral and emotional texts, and so this feature may be considered universal.

Another regularity in phonetic division is that pauses are more frequent in emotional texts than in neutral. An analysis of pauses in identical texts in English, German, French, Russian and Ukrainian provides their uniformity, which indicates the universal character of this prosodic feature.

The results of the experiment are interesting from the point of view of speech communication which is a continuous and complex process of transmitting not only ideas but also emotions, attitudes. Communicative approach to the study of emotions reveals that emotional information is often organized in larger suprasentential units - texts. Each text presents a speech act with a concrete pragmatic aim. The components of speech act are described in a well-known R. Jackobson's scheme-addressor (speaker) - message - addressee, recipient (listner).

The study of communicative orientation of emotional texts show that such emotions like **tenderness, sorrow, offence, shame** are connected with the **addressor** as they express the emotional state of the speaker; **anger, indignation, threat**, reproach are directed to the recipient-to the listner; new information in the message arouse **surprise, delight**; absence or delay of message leads to **fear** and **despair** ; **contempt, irony, suspicion** are connected with the character of the received message.

The data of experimental study show that the same communicative orientation of emotional text in different languages leads to common acoustic parameters in them.(see Table 2).

Table 2. The comparison of acoustic parameters of emotional and neutral texts.

Components of speech acts	Emotions expressed in speech acts	Acoustic parameters F_0	I	t
addressor (speaker)	sorrow tenderness offence shame	$F_e < F_n$	$I_e < I_n$	$t_e > t_n$
addressee recipient (listner)	anger indignation threat reproach	$F_e > F_n$	$I_e > I_n$	$t_e > < t_n$
message: receipt	delight surprise	$F_e > F_n$	$I_e < I_n$	$t_e > < t_n$
message: absence delay	fear despair	$F_e > < F_n$	$I_e > < I_n$	$t_e > < t_n$
message: character	contempt irony suspicion	$F_e m > F_n m$	$I_e m > I_n m$	$t_e n > t_n n$

Where $F_x m = F_x max$, $I_x m = I_x max$
$t_x n = t_x$ nuclear syllable

The data in the table prove that texts expressing emotional states of the speaker are characterized by the decrease of fundamental frequency and intensity and longer duration. Emotional texts directed to the listner basically increase all the acoustic parameters. Emotional texts connected with the receipt of the message are characterized by a sharp increase of fundamental frequency, decrease of intensity and great variations of their temporal structures.

Absence or delay of message leads to great variations of all the acoustic parameters. Texts, which are reactions to the character of the message received increase all the acoustic parameters of semantically important words in them.

CONCLUSIONS

The study has attempted to elucidate one aspect of prosody in the textual function-universal character of emotional expression. It shows that there are large areas of overlap between even the most diverse languages in the use of common features of prosodic patterning. A contrastive analysis of emotional and neutral text intonation in English, German, French, Russian and Ukrainian has revealed common means of emotional expression. Emotions are expressed in speech signal through a complex of acoustic features. The particular set of the features, however, depends of the type of emotion and degree of emotional colouring.

REFERENCES

[1] Chomsky, N. (1965) *Aspects of the theory of syntax*, Cambridge, Mass., Mit-Press.

[2] Bolinger, D. (1980) "Intonation across languages", In Universals of Human Languages, London, Longman, pp. 471-524.

[3] Nushikyan, E. (1987) "The typological analysis of emotional speech prosody", Proceedings XI th ICPhS, Vol. 3, pp. 210-213.

[4] Nushikyan, E. (1986) *Typologia intonatsii emotsionaloj rechi*, Kiev-Odessa.

[5]Jackobson, R. (1963) *Essais de linguitique général*, Paris.

PRODUCTION AND PERCEPTION OF THE SINGING FORMANT

C. Pillot
Institut de phonétique de Paris, France

ABSTRACT

Microphonic and glottographic records of 7 male opera singers confirmed the acoustic properties of the singing formant (SF). MRI measurements and the use of a vocal-tract acoustic simulation program allowed to quantify the effect of the lowering of the larynx. Perceptual experiments reveal that SF influences a number of perceptual dimensions of voice, the phonetic quality of the vowels and the personal vocal quality.

INTRODUCTION

The acoustic characteristics of the singing formant SF are well known [1, 2, 4]. The purpose of this study is to investigate further the acoustic (spectrographic measurements and use of Maeda's vocal tract acoustic simulation program), physiological (MRI) and perceptual aspects of SF for French vocal productions.

I. ACOUSTIC DESCRIPTION

We perform simultaneous microphonic and glottographic records of 4 male professional and 3 untrained singers singing isolated sustained vowels [a], [i], [o] and [u], a given sentence extracted from a familiar piece of music and a melody they have freely chosen.

In accordance with Bartholomew [1] and Sundberg [2], our results show that there is a significant difference of amplitude of the region around SF between the untrained and trained singers (t_7=7.57; $p<0.02$) in favour of the latter.

Furthermore, for a given singer, SF frequency doesn't vary whatever the production sung. Finally, SF frequency varies significantly according to the vocal category of the trained singers: for example: about 2620 Hz for the bass singer, 2800 Hz for the baritone and 3406 Hz for the tenor (Fig. 1).

The comparison of the source spectra on the one hand, and of the acoustic spectra on the other hand allowed us to formulate a few hypotheses about the origin of SF:

The intensities of the source spectrum harmonics at the frequencies of the SF are higher for the professional singers.

Furthermore, the bandwidth of SF is about equal to the double ($p<0.0001$) of the theoretical bandwidth of the formants at the same frequencies, Fant [3]: SF doesn't come presumably from one, but from at least 2 formants, as asserted by Sundberg [2].

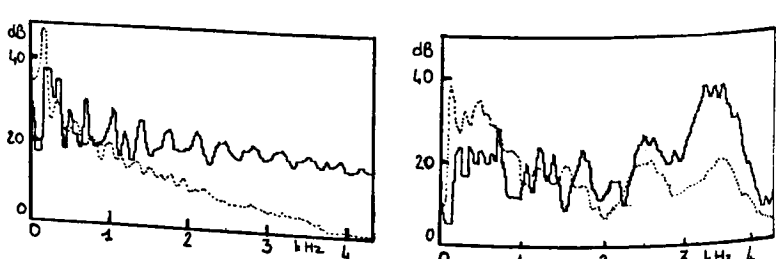


Figure 1. Average spectra of the sentence sung by a professional tenor singer (solid lines) and by an untrained one (dotted lines) at the same tempo and at the same frequency (spectra averaged on 8 seconds). Left: source spectra; Right: acoustic spectra.

II. PRODUCTION

Wave study

We have just noticed that the source produces more energy around 3KHz among the singers who have SF. The influence of the glottal flow parameters on the higher formants in the French vowels [i], [a] and [u] was studied using the Klatt's synthetizer type named Compost (Bailly).

The open quotient (OQ) is the ratio of the opening and closing times to the total duration of the cordal vibratory cycle. The disymmetry quotient (DQ) is the ratio of the opening time to the vocal cords closing time. Our results show that the reduction of OQ and the increase of DQ allow the spectra of the resulting sounds to have more intense high harmonics whatever the vowel sung.

Articulatory study

The preceding source phenomena are not sufficient to explain the emergence of a *peak* like SF.

Maeda's Model

What sort of articulation can generate SF? The hypothesis of a sinus Morgagni's resonance combined with a laryngeal lowering, Sundberg [4], is checked with the Maeda's vocal tract acoustic simulation program [5]. Laryngeal lowering simulation with the 2 tube model entails an emergence of F3 and F4 which come closer, as the F5 amplitude decreases. The variation of the ratio of the laryngeal to the pharyngeal section by modification of the laryngeal section confirms Sundberg's theory according to which the larynx is active as a resonator only if the pharyngeal section is at least six times wider than the laryngeal one.

Magnetic Resonance Imaging

The area functions corresponding to the French vowels [i], [a] and [u] spoken and sung by a professional bass singer (Fig. 2) were estimated from Magnetic Resonance Imaging (MRI) midsagittal images (Magnetic field: 0,5T; Gyrex V machine; ET: 15ms; RT: 33ms; head coil; acquisition time: 8s). Fig. 2 (bottom) shows for [a] the creation of SF (merging and reinforcement of F3 and F4) and Fig. 2 (top) suggests a significant lowering of the larynx (arrow 1) and the jaw (arrow2), a slight labial protrusion and a lingual posteriorisation (arrow 3).

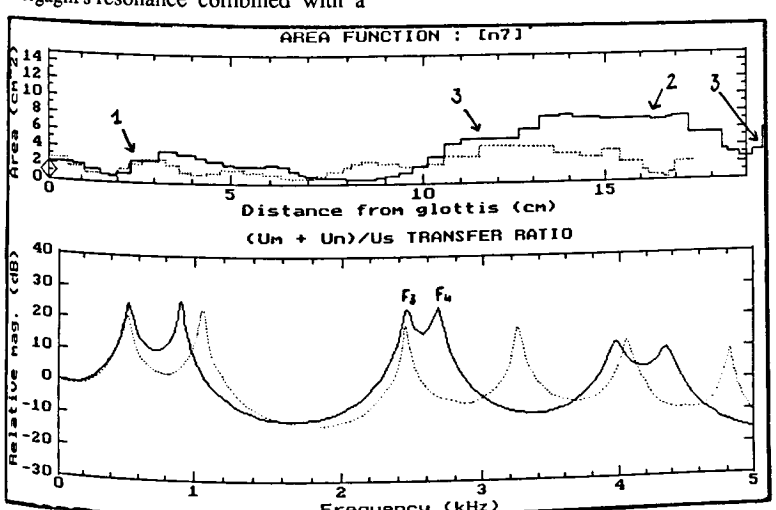


Figure 2. Area function and transfer function of the vowel [a] spoken (dotted lines) and sung (solid lines) by a professional bass singer (same frequency: 100Hz). Data coming from the Maeda's vocal tract acoustic simulation program [5].

III. PERCEPTION

To study SF perceptive significance, 2 comparative tests of pairs of words and orchestral extracts were performed with naive auditors.

Words

The word "solitaire" was sung by 7 singers. The SF of the 4 professional singers was filtered, while the corresponding SF frequency zone of the untrained ones was amplified. In a discrimination task, all of the 22 auditors perceive a difference between words with and without SF. In a forced decision task ("Quel est le mot le plus riche?"), the word considered as "plus riche" have the SF for the professional singers (70%). The artificially created SF for the untrained singers was not as perceptually effective as the original SF (40%).

Musical sentence (orchestral context)

We choose as stimuli an excerpt of a CD-record (The opera Faust by Gounod) during 15 seconds. The SF of the tenor voice was then filtered. 54 auditors were asked to freely judge the perceived difference between the original and the filtered versions.

Table 1. Distribution of the responses for the comparison of 2 musical extracts among 54 auditors.

Sort of responses	Result in %	Sort of responses	Result in %
Timbre	31,9	Pitch	7,34
Aesthetics	13,07	Others	6,62
articulation	12,4	/orchestra	4,36
Emotion	9,17	Duration	3,44
Intensity	8,49	Distance	3,21

Our results show that the SF is first perceived in terms of timbre ("clair, brillant, riche"). The intensity only intervenes in the fifth position ("forte"). The voice of the singer who has the SF is more "articulée", "belle" (aesthetics), "proche" (distance), "courte" (duration) and "plus aigue" (pitch). The interpretation of the results is complex.

Vocalic identification

22 naive auditors were asked to identify the original vowels of the professional singers, and their filtered versions, and the original vowels of untrained singers and their SF added versions. The singer productions were [i], [a] and [u] vowels, and the listeners had an open choice (15 French vowels).

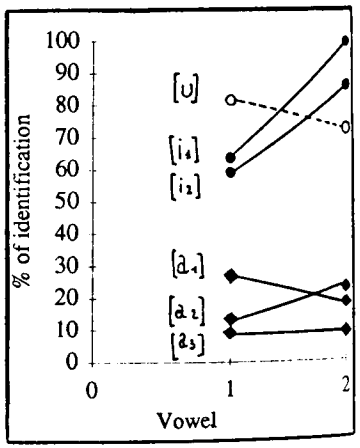


Figure 3. Average percentages of correct identification of vowels sung by the professional singers. 1: without SF; 2: with SF.

As shown in Fig. 3, the percentages of identification for the professional singers are statistically smaller for the vowel [a] than for [i] and [u] (t_{22}=-9.7; $p<0.0001$). Furthermore, SF has more influence on the identification scores of [i] and [u] than [a] (t_3=-9.07, $p<0.005$). The suppression of SF (on the left) worsens the identification of [i], and allows a better recognition of [u]. [i] and [u] of the untrained subjects are better identified (100 %), without (original) or with added SF.

The filtered vowel [i] of the professional singers is mainly misperceived as [y] (77.2 % of the errors), while original [u] (with SF) is often perceived like [o] (91.67 % of the errors). [a] without SF is perceived like [ɔ].

DISCUSSION

The acoustic analysis of sung productions allows to formulate a few hypotheses about the origin of the SF, which have to been confronted with physiological data.

The study of Magnetic Resonance Imaging (MRI) of sustained vowels confirms the *compensatory articulation* hypothesis of Sundberg & al [6]: MRI shows us "supershapes" of the sung vowels; for example, a lowered larynx and jaw position in spite of a raised tongue can be interpreted as "superpalatalization" for the vowel [i], according to Sundberg [6]. Singing requires indeed a vocal tract free of any constriction and, in the same time, a high degree of flexibility of phonatory muscles. This investigation entails methodological problems due to the supine position of the subject or to the width-to-area conversion.

The study of SF shows also the importance of its perceptive relevance: it's located in a frequency region where the auditory sensitivity is maximal. (The threshold of audibility is minimal at this frequency) [7] This allows the singer's voice not to be masked by an orchestral accompaniment and this without pathogenic vocal effort. Trumpets have the same peak of intensity at 3 KHz, and this instrument merges in an orchestra. The spontaneous qualification of two musical extracts (identical except presence or absence of SF) showed us a great variability in the sort of obtained responses: indeed, a physical value like the SF affects a multiplicity of perceptual values, including the timbre.

Musicians have started to be interested by the timbre of their vocalizations only in the nineteenth century, and expansion of orchestra have obliged singers to develop new vocal strategies to make their voices more audible, at a time where electric amplification didn't exist.

CONCLUSION

SF, which is the real signature of the occidental operating singing, probably exists in other vocal and sound productions. Its origin is multifactorial : it's produced by a singer who have learned to face the orchestral accompaniment by means of phonatory and articulatory modifications.

SF has an effect non only on the vocalic quality, but also on the personal vocal quality of the singer.

ACKNOWLEDGEMENTS

This work was greatly aided by the careful experimental assistance of D. Rostolland and M. Elmaleh. We want to express our gratitude to J. Vaissière and S. Maeda for their support and comment on an earlier draft of the paper.

REFERENCES

[1] Bartholomew, W.T. (1934), "A physical definition of good voice quality in the male voice", *JASA*, vol.VI; pp 25-33.

[2] Sundberg, J. (1987), *The science of the singing voice*, Dekalb Illinois, Northern Illinois University Press, 216p.

[3] Fant, G. (1970), *Acoustic theory of speech production*, The Hague, Mouton & Co., 2nd edition, 328p.

[4] Sundberg, J. (1974), "Articulatory interpretation of the singing-formant", *JASA*, vol.55,n°4, pp. 838-844.

[5] Maeda, S. (1992), "Modélisation articulatoire du conduit vocal", *Journal de Physique IV*, vol.2, pp. 307-314.

[6] Sundberg, J. & Lindblom, B. (1971), "Acoustical consequences of lip, tongue, jaw and larynx movement", *JASA*, vol.4, pp. 1166-1179.

[7] Castellengo, M. (1973), "Considérations sur la voix des chanteurs professionnels", *Bulletin d'acoustique musicale*, n°67, pp 9-14.

AN ACOUSTIC AND PERCEPTUAL STUDY ON THE EMOTIVE SPEECH IN KOREAN AND FRENCH

CHUNG Soo-Jin
Institut de Phonétique, 19 rue de Bernardins Paris, France

ABSTRACT

This study presents the acoustic and perceptual analysis of the emotive speech loaded with anger, joy, sorrow, or tenderness, in Korean and French. Statistic analysis found the factors which affected the identification of emotions, such as emotion-type, comprehension level of a given language, modality of phrase, etc. Based on the acoustic similarities, we regrouped the studied emotions into active vs. passive emotion group and positive vs. negative emotion group. The perceptual confusion of the emotions in a same group was explained mainly by the similarity of the activity aspect. We reported also the acoustic filtering experience which had an effect on the identification of emotions in relation to the mother tongue of listeners.

1. INTRODUCTION

The emotion is a complex phenomenon ; a given emotion is considered as a result of the interaction between acoustic, physiological, and psychological features.

In the speech analysis, the personal or emotional aspect was somewhat ignored in contrast with the rich literature of the lexical or grammatical aspect. In this paper, we review briefly the general prosody and the notion of emotion, and make a acoustic and perceptual study of the vocal expression of emotions, such as anger, joy, sorrow, and tenderness, using neutral sentences as a reference.

2. GENERAL CONCEPT

Every prosodic feature seems to have a motivated origin, ethological, physiological or psychological, and this paralinguistic origin may explain the similar prosodic tendencies through non-related languages, for example, Korean and French.

In general, the biological necessity to breathe at regular interval creates the pause which can be also present for marking different degrees of frontier. The downward pitch contour is due to the depression of subglottal pressure and the gesture for a sentence or a meaning group can be described as a set of tension and relaxation of articulators.

Concerning the perception, one does not hear directly physical variations. Fraisse (1974) suggested two perceptual organizations, "Accentual rhythm" of initial segment and "Temporal rhythm" related to the final lengthening. The coexistence of two rhythms exercises contradictory forces on the interpretation of rhythm.

These general tendencies can be modified by semantic or emotional emphasis.

Many of terms used to describe emotions, not being clearly defined in the literature. Because it is impossible to quantify the emotion and there is no objective rules to define emotive terms.

According to Scherer (1981), the emotion is "the organism's interface to the world outside", having three principal functions ; "a) they reflect the evaluation of the relevance and significance of particular stimuli, b) they physiologically and psychologically prepare the organism for appropriate action, c) they communicate the organism's state and behavioral intentions to other organisms in the surroundings". He noted also that emotion is not a steady state condition, but a process of events, which arise in rapid succession following a stimulus event ; a given emotion is the result of a series of "stimulus evaluation checks" in the "component patterning model".

Many emotion theories use the concept of "basic" emotions but there is few agreement as to what constitutes a "basic emotion". In the combinational emotion theory called a "palette theory" by Scherer, new emotions are produced by mixing the primary basic emotions together. In terms of universality, the basic emotion responses are cross cultural, while responses to nonbasic emotions are learned, and hence culture dependent.

Three aspects can be determined as dimensions of the emotion ; "Strength" ranging from contempt to fear or surprise, "Valence" ranging from love or happiness to anger, and "Activity" ranging from sleep to tension.

Davitz (1964) found that the activity aspect of emotional meaning is carried by the relatively simpler elements of the vocal symbol, such as pitch and loudness, while both valence and strength are communicated by subtler and more complex vocal patterns of inflection, rhythm, etc. His study reported also that where erroneous judgments were made, it was very often in favor of another emotion with a similar activity level rather than a similarity in terms of valence or strength.

3. EXPERIMENTAL ANALYSIS

3.1. Procedure

The recordings were made in a recording studio using a high quality microphone and a DAT recorder. A male amateur actor and a male professional actor, about thirty years old, were recorded in separated session, each speaking the same sentences, five for each emotion, anger, joy, sorrow, or tenderness, and neutral as an unmarked expression. The actors were presented with a description of a situation causing a given emotion and asked to repeat the sentences six times. This procedure was same for the Korean and French recordings. The meaning of the sentences were : "Jean invited you at his party. Will you go there? You have met him last night, it is so often, isn't it? Don't go there, today. Let's prepare the dinner for my friends coming tonight."

In order to validate the expressivity of the recorded sentences and to know how the emotions are identified, we proceeded the tests of identification. Listeners were asked to identify the intended emotion among five ones, after listening a stimulus phrase ; the questionnaire consisted of thirty stimulus phrases presented once at five seconds interval. Four tests were carried out by the combination of two kind of corpus, Korean and French, and two groups of listeners, eighty-four Koreans and eighty-five French's.

In the result, the score of correct responses of the emotions was always in this order ; Anger > Sorrow > Tenderness > Neutral > Joy. The amateur expression of emotions and the professional were not significantly different in the listener's perception, even though the latter was more typical and obvious.

As to establish the Korean and French corpus to be studied, we selected ten sentences for each emotion with the high score identification.

3.2. Statistic Analysis

On account of limited space, we present mainly the result and the discussion of the imperative sentences of the corpus which were most representative to characterize each emotion ; the phonetic transcription of the Korean phrase is [onirin kaʒima] and the French [nivapa oʒurdyi].

By means of the analysis of variance (ANOVA) of the result, we found that the factors influencing the identification of emotions, such as emotion-type, comprehension level of a given language, and modality of the phrase, but the length of the phrase did not. The most identification was guaranteed when listeners heard a imperative phrase of anger in his mother tongue.

3.3. Acoustic Analysis

We studied three principal parameters, pitch , duration, and intensity, not missing the voice quality.

The pitch contour seemed to carry a large part of emotional information in both languages ; especially, anger was characterized by the abrupt final chute. The wide pitch range with dynamic movements of anger or joy was located higher than the narrow range of sorrow or tenderness. These differences were emphasized on the important words, for example, the negative morpheme of the imperative negative phrase, located at the final and the beginning of the Korean and the French phrase respectively.

The duration of the last syllable differed greatly according to the emotions : it was very short in anger and long in joy and tenderness, while the speech rate was high for anger and joy but slow for sorrow and tenderness.

Emotions	Active	Passive		Voice Quality
Positive	Joy	Tenderness	===>	resonant
Negative	Anger	Sorrow	===>	Aspirated / devoiced
	⇓	⇓		
	High	Low		Register
	Wide	Narrow		Pitch Range
	High	Low		Speech Rate
	High	Low		Intensity

Figure 1. Regrouping of the emotions according to their characteristics

The vowel lengthening occurred at the final part of every emotion : the last syllable was longest for tenderness and shortest for anger. However, the variation of consonantal duration was less regular ; usually consonants lengthened on important words.

The intensity was high for anger or joy and low for sorrow or tenderness, as expected.

The voice quality determined on the valance dimension was quite different depending on the positive or negative emotion : the joyful and the tender voice were more resonant than that of anger and sorrow which were more aspirated.

According to the characteristics common to the Korean and French sentences, we regrouped the studied emotions in active vs. passive emotions on the one hand, and positive vs. negative emotions on the other hand as described in the figure 1. Neutral sentences were closer to the passive emotion sentences than to the active. Similar experiences and results were reported by [1] and [2].

3.4. Perceptual Analysis

The previous tests of identification of emotions showed how the emotions are identified and confused. By reason of the complex paralinguistic features of the emotion, which are not coded systematically as linguistic features, it happens often the disjunction between speaker's coding and listener's decoding, interpreted as the confusion.

Pakosz (1983) noted five points of special interest from the literature :

"a) Speakers vary markedly in their ability to express emotive meaning vocally in controlled situations.

b) Listener's recognition and interpretation of emotions from recorded speech varies substantially.

c) Some emotions are more readily expressed and identified than others.

d) Misidentifications seem to follow a regular pattern whereby similarity on the activation dimension between two emotions leads to confusing one for the other.

e) Recognition of emotions is possible under conditions of reduced information concerning pitch variation."

These notes were revised and validated in our analysis. For instance, anger was quite well expressed in the imperative phrase and easily identified, even by foreigners, then less confused with other emotions. While joy was often confused, especially with anger having high intensity at high register in the activity dimension. In the same way, tenderness was confused with sorrow by their low intensity at the low register.

The previous regrouping is efficient to explain the fact of confusion : a given emotion is more often confused with emotions in the same group than in the other group, especially in the activity dimension. By the way, in view of the direction of confusion illustrated in the figure 2, it seems that listeners tend to choose a negative emotion when they are not sure to decide a positive emotion in the same group.

In the supplement test, we asked listeners to write adjectives evoked by the stimulus. Diverse adjectives were written for a given emotion and some of them were written again for the other emotions.

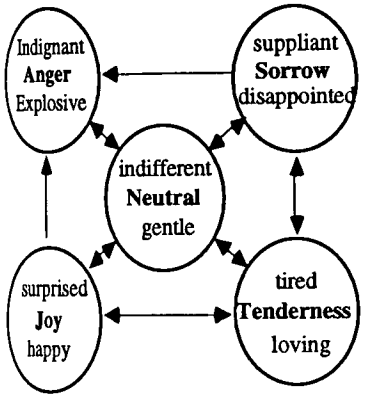


Figure 2. Direction of the confusion between emotions and Adjectives most frequently written in the supplement test

This figure shows the complex nature and the similarity of emotions. In a word, the perceptual confusion is caused by acoustic similarities of the emotions.

3.5. Filtering Effect

For the filtering experience, we selected the sentences of neutral and two emotions obviously contrasted, anger and tenderness, from the original recorded corpus.

In order to deprive the corpus of intelligibility, we eliminated the frequencies above 250 Hz. Apart from our intention, this operation affected the first formant of open vowels such as [a], then the syllabic regularity of intensity was disturbed. And it affected also some part of the F0 of anger exceeding 250 Hz, so destroyed its fine structure of pitch contour, while tenderness preserved its original pitch contour.

Next, we had listeners pass the test of identification of emotions with original and filtered corpus, of his mother tongue and of a foreign language, respectively ; thereby four tests were carried out.

As the result, the filtering had an effect on the identification of emotions when listeners heard the filtered corpus of his mother tongue. In most cases with unfiltered corpus, a listener identified better anger than tenderness, either of his mother tongue or of a foreign language. However, with the filtered, then unintelligible corpus, the pattern of the identification of emotions changed : the listener identified still better anger than tenderness in the filtered corpus of a foreign language but he identified better tenderness than anger in the filtered corpus of his mother tongue.

In conclusion, it seems that the fine structure of pitch contour and intensity play a important role in the identification of emotions and that listeners rely on different principal criterion, depending on the intelligibility of stimulus:

if the stimulus is intelligible with all information, whether he knows the language or not, he identified best anger having great intensity as well as fine pitch contour,

if the information is so reduced to make stimulus of a foreign language unintelligible, he identified still better anger than tenderness, based entirely on the striking intensity of anger

however, if he hears the unintelligible corpus of his mother tongue, his melodic intuition prevents him from the decision of anger which lost the fine prosodic structures of pitch and intensity, and he identifies better tenderness preserving the fine structure of pitch contour.

4. CONCLUSION

So far, we reported the analysis of the emotive speech in Korean and French, concerning the universality of emotions and the problem of perception.

This information could be used as the basis for a set of rules to control a high quality speech synthesizer with simulated emotion effects in the output speech. As emotion forms such an important part of human speech, its incorporation into speech synthesis systems is surely imminent.

REFERENCES

[1] Murray, I. R. and Arnott, J. L. (1993), *Toward the simulation of emotion in synthetic speech : A review of the literature on human vocal emotion*, J. Acoust. Soc. Am. 93 (2), pp. 1097-1108.

[2] Abadjieva, E., Murray, I. R., and Arnott, J. L. (1993), *Applying Analysis of Human Emotional Speech to Enhance Synthetic Speech*, Eurospeech 93, pp. 909-911.

DELAY IN ORAL PRODUCTION AND PRONUNCIATION ACHIEVEMENT IN A FOREIGN LANGUAGE

Darío Barrera Pardo

Department of English Philology, University of Vigo, Spain

ABSTRACT

This paper reports on experimental research based on the assumption that a methodological focus on well-founded receptive skills is essential for the eventual development of productive competence, and that this is especially relevant in the learning of foreign language (FL) pronunciation. In order to test this hypothesis, an experiment was designed in which a control and an experimental group (N = 9 for each group) followed the same program of instruction (L1 = Spanish, L2 = English), differing only in the method of instruction (perception-only for the experimental group, perception and production for the control group). The results of this experiment indicate an advantage for the control group.

1.INTRODUCTION

Research into the learning of foreign or second language pronunciation is relatively scarce. Although progress has been made in recent years, we are for the most part lacking substantiated answers to questions such as which strategies learners employ to approximate their speech to the target models, or which teaching methods best contribute to pronunciation achievement.

In this study I tested the hypothesis that the development of adequate speech models in perception facilitates eventual attaintment in pronunciation production. It has been argued by several researchers, notably Postovsky [11], and Gary [4] that a comprehension-first approach to foreign language learning has a number of advantages over methods that require immediate production. The methodological construct by which this comprehension approach is implemented in the teaching practice is what is known as "silent period". Specifically, these authors claim that the incorporation of a silent period in the beginning stages of instruction will enhance the learners' acquisition of the foreign language.As a general teaching procedure, the silent period has been fostered mainly by Krashen and his associates (e.g. [2]). In the area of pronunciation teaching, some authors have explored the effects that a silent phase in instruction have for phonological acquisition, with seemingly positive results. It is perhaps the study carried out by Neufeld [9] the one that shows a stronger correlation between an initial silent period and ultimate acquisition of the target sounds. Other researchers [12] include an initial silent phase in their pronunciation training programs. In addition, FL methods such as *Total Physical Response* incorporate a silent period in their instruction program.

There is some rather inconclusive evidence in support of this theory coming from child second language studies in natural settings [3], and more recently other researchers have included a perception-only stage in their pronunciation training studies [10].

2. PERCEPTION/PRODUCTION IN FL PHONOLOGY

That FL learners need to construct adequate target speech models before making attempts at production is an idea that has gained recognition in the field (e.g [7]). In fact, one explanation that has been advanced for the foreign accent of learners is that some FL speech approaches are based on the wrong assumption that "phonological representation should be easily, if not automatically, determined by second language learners within a phonemic model of phonology" ([1] p.247). Rather, learners face the taxing task of constructing their own representations, a goal they attain in many cases at best only partially.

On the other hand, there is evidence that points toward a perception/production split in FL speech competence. Neufeld [10] found that his subjects performed on a native-like level in listening and phonological discrimination tasks, although in terms of speech production they were rated as "poor articulators" by native judges. These subjects' sound knowledge of L2 phonology therefore was not matched by equivalent productive skills.

3. EXPERIMENT

3.1. Subjects

18 monolingual Spanish-speaking university students who volunteered to participate in the experiment were assigned to a control group and an experimental group (each group N = 9). These subjects had no knowledge of English, and were told that they would acquire basic "survival" English skills.

3.2. Method

The control and experimental groups were presented with the same input, with a focus on the following English phonemes: /i: ɪ ɛ æ t d/, which are typically problematic for Spanish-speaking learners for the reasons that follow:

1) maintaining the quality-quantity distinction in the pair /i: ɪ/, normally merged to the Spanish high front vowel /i/.

2) maintaining the quality distinction between /ɛ æ/, which may be realized as either /e/ or /a/ by Spanish learners.

3) realizing /t d/ as alveolar plosives, rather than dentalizing them as is the norm in Spanish; adjusting the Spanish VOTs for the durational values of these English consonants (e.g. the aspiration of [t^h]).

The rather restricted input of the experiment presented learners with numerous instances of words containing these phonemes (a minimum of 20 words for each target phoneme). All main allophonic variants of the target phonemes were represented in the input words (for example, [t^h] [t˺] [t] for /t/; [i:] [i˙] [i] for /i:/). Both groups met with the instructor two hours per week during a four-week period, following the same training in the type of language situations in which these sounds were presented, that centered around topics such as learning about foods to order in the US and some simple routines regarding the pragmatics of this linguistic task.

3.3. Procedure

The control group was instructed with a traditional approach, in which speaking on the part of the learners was encouraged from the beginning of the training. *Listen and repeat* activities were frequent in the classroom interaction, and active oral production from the part of the learners was also encouraged.

It must be noted, however, that a focus on correct pronunciation was not an aim here. In the experimental group, on the other hand, subjects were not required to respond *orally* or to talk with any regularity until the last week of instruction. To adapt their training to this less orthodox methodology, *active listening* was encouraged. According to this orientation,the subjects did not limit themselves to receive language input (in particular oral input) passively, since they had to answer to questions and instructions directed to them (identifying words, pictures, performing physical and gestural activities, among other modes of response).

3.4. Analysis

Once the four-week training program was completed, the 18 subjects of the study were presented with a list of 30 words to be read and tape-recorded in the language lab (5 words for each of the 6 target sounds that were the focus of the experiment).

This procedure yielded 30 recorded tokens for each subject. The test words for each focus sound represent a selection of contextual and positional variants of the sound being tested. Thus, for the focus sound /i:/, the first test word "tea" [t^hi:] is a token of the open-syllable variant [i:]; "beats" [bi˙ts] and "eat" [i˙t] both contain the voiceless consonant syllable-closed [i˙], in which the otherwise long vowel is significantly shortened; finally, "cheese" [tʃi:z] and "bean" [bi:n] are instances of the opposite case, that is, a voiced consonant

syllable-closed [i:]. All the test word recurred frequently in the input to which the subjects were exposed.

These recordings were next presented to three volunteering native-speaking judges (American students with little knowledge of Spanish), who rated the recorded test words according to a scale measuring degree of *accented speech* with mid-points for ease of evaluation (see Figure 1). Judges listened to the recordings in a random order in the language lab. Therefore each of the 30 words was given a value from 0 to 5 according to how *accented* it was rated by each judge.

3.5. Results and discussion

The results of the experiment are plotted on Table 1. These show a slight advantage for the control group (4,48 % less accented than the experimental group), thus refuting the perception-first hypothesis for FL phonological learning. This performance difference is significant [t = 1.90 (one-tailed), $p < 0.05$].

Our results therefore contrast with those obtained in previous studies [4, 9, 10, 11, 12] in which the group that was treated with a delay in oral production showed more improvement in pronunciation than learners who were instructed with a production approach since the beginning of training. A few reasons come to mind that may explain the divergent results obtained in our experiment:

1) the subjects of our instruction program were not encouraged in any explicit way to concentrate their learning efforts on approximating their speech to the target models (this, for instance, is an important difference with experiments such as [12]).

2) the material used in the program of instruction [5] and its *methodological approach* had a marked communicative orientation, and learners probably attended more to meaning than to phonological form (i.e. subphonemic aspects of the input such as aspiration of /t/ may have been largely ignored). This rationale is in consonance with some explanations advanced for poor pronunciation in second language acquisition [6].

3) it may in principle seem logical that 8 hours of instruction (2 hours/week X 4 weeks) should of necessity produce little result in terms of language acquisition and all the more in the domain of phonological acquisition, but it must be reminded that those FL phonology studies in which a training period is incorporated, it has typically a short duration. (cf. [8]).

The hypothesis that this study has explored is nevertheless worth being followed up in future research. As a factor in phonological acquisition, the perception/production dichotomy and its correspondingly different learning strategies from the part of learners is an issue that has attracted the interest of researchers in the field from time to time; it will undoubtedly be a center of interest in the future as well.

4. CONCLUSION

The purpose of our study was to asses the effect that an oral delay in production has for FL pronunciation ultimate attainment. The present results show a more positive effect for an approach that involves students in both perception and production since the early stages of instruction.

REFERENCES

[1] Ard, J. (1989), "A constructivist perspective on non-native phonology". In S. Gass and J. Schachter (Eds.), *Linguistic perspectives on second language acquisition*, New York: Cambridge University Press, pp. 243-259.

[2] Dulay, H., Burt, M., and Krashen, S. (1982), *Language Two*, Oxford: Oxford University Press.

[3] Ervin-Tripp, E. (1974), "Is second language like the first?", *TESOL Quarterly*, vol.8, pp. 111-127.

[4] Gary, J. (1975), "Delayed oral practice in initial stages of second language learning". In M.Burt and H. Dulay (Eds.), *On TESOL '75: New Directions in L2 Learning, Teaching and Bilingual Education*, Washington, D.C.: TESOL, pp.89-95.

[5] Hecht, E., and Ryan, G. (1979), *Survival pronunciation* (Student workbook), Englewood Cliffs: Prentice Hall.

[6] Krashen, S., and Terrell, T. (1983), *The natural approach: Language acquisition in the classroom*, Hayward: Alemany Press.

[7] Leather, J., and James, A. (1991), "The acquisition of second language speech", *Studies in Second Language Acquisition*, vol. 13, pp. 305-341.

[8] Macdonald, D.,Yule, G. and Powers, M. (1994), "Attempts to improve English L2 pronunciation: The variable effects of different types of instruction", *Language Learning*, vol. 44, pp. 75-100.

[9] Neufeld, G. (1978), "On the acquisition of prosodic and articulatory features in adult language learning",*The Canadian Modern Language Review*, vol. 34, pp. 168-194.

[10] Neufeld, G. (1988), "Phonological asymmetry in second language learning and performance", *Language Learning*, vol.38, pp. 531-559.

[11] Postovsky, V. (1974), "Effects of delay in oral practice at the beginning of second language learning", *Modern Language Journal*, vol. 58, pp. 229-239.

[12] Scheiderman, E., Bourdages, J., and Champagne, C. (1988), "Second-language accent: Relationship between discrimination and perception in acquisition", *Language Learning*, vol. 38, pp. 1-19.

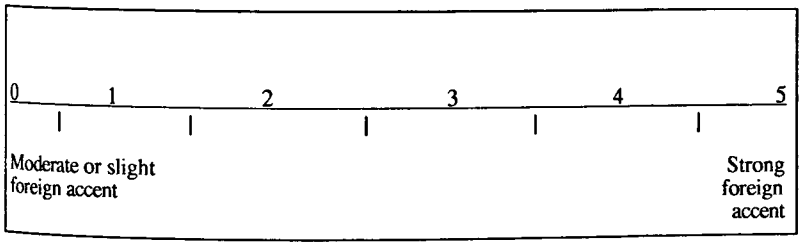


Figure 1. Scale measuring degree of accented speech used by the experiment judges

Table 1. Assessment of degree of accented speech

	N	M	SD	df	t	degree of foreign accent
Control group	9	293,9	17.62	16	1.90	65,3 %
Experimental group	9	314,2	26.76			69,8 %

$p > .05$

DO L1 TRACES HELP LISTENERS IN L2?

Z.S. Bond and Thomas J. Moore
Ohio University, Athens, Ohio; Wright-Patterson AFB, Ohio, U.S.A.

ABSTRACT

Four proficient Chinese speakers of English as a second language read intelligibility test materials to three groups of listeners. Under all testing conditions, American listeners performed best. Sharing the same native language did not always prove to be an advantage to listeners in L2.

INTRODUCTION

In the extensive literature dealing with the intelligibility of non-native talkers, there is evidence that L2 speakers understand speakers from their own language background somewhat better than speakers with less familiar speaking patterns.

This study was designed to examine intelligibility of proficient non-native talkers using controlled speaking materials. Talkers were tested in quiet and noisy speaking conditions.

The specific questions of interest concern the intelligibility of accented speech to proficient listeners from different language backgrounds; consistency of differences among listener groups across test materials and listening conditions; differences among talkers; and listener evaluations of talkers and test conditions.

METHOD

Four talkers participated in the experiment. All were male native speakers of Mandarin Chinese. The talkers had lived in the United States from 1 1/2 to 5 years. All four were highly fluent in English.

Materials

Each talker recorded two Modified Rhyme Test word lists and 20 sentences developed by Pisoni, et al. [1]. The MRT uses one-syllable test words in a carrier sentence. The test sentences were six words long. In responding to the MRT, listeners identify a spoken word from a group of alternatives; consequently, the MRT does not require much memory or linguistic knowledge. Rather, it assesses clarity of pronunciation.

The sentence test requires listeners to understand a statement and to judge it as true or false. Because the sentence task demands linguistic knowledge as well as an acquaintance with real-world cultural background, it may be representative of communications situations.

Listeners heard the recordings either in quiet or mixed with pink noise at S/N 3 dB.

Listeners

All listeners were students at Ohio University. There were three groups, Americans, native speakers of Chinese, and other students from East Asia, primarily from Korea, Japan, and Thailand. The number of listeners in each condition is given in Table 1.

Table 1. Listeners

	Clear	Noise
Americans	12	20
Chinese	20	20
Other E Asian	22	20

Listeners were tested in small groups in a language laboratory. They heard both the MRT and the sentences over headphones in one listening condition. After hearing each talker, they evaluated the talker on a 5-point rating scale; 1 was defined as 'Easy' and 5 as 'Difficult.'

Data Analysis

The MRT and sentence scores were the dependent variables submitted to ANOVA, treating between-group and within-group effects separately. The independent variables were talkers, language background of listeners, listening conditions, and tests. Talker and test were within-group factors; listening condition and language background were between-group factors. Post-hoc analyses of interactions used Cicchetti tests. In addition, correlations were calculated between subjective ratings and intelligibility scores.

RESULTS

As expected, clear speech was always more intelligible than speech mixed with noise. In addition, listeners from different language back grounds performed differently, and the four talkers were not equally intelligible.

Listening Conditions

Fig. 1 shows the percent correct responses to the MRT in both the clear and the noise listening conditions. In the clear condition, American listener performance was best, followed by the Chinese and by the other East Asians. In the noise condition, all groups performed considerably less well, and differences among the three groups were not significant. From the clear to the noise listening condition, American listeners showed the greatest performance decrement, 48%.

Fig. 1 also shows the percent correct responses to the sentence test in both clear and noise listening conditions. The expected drop in performance when listening in noise is present. The Americans performed extremely well in the clear condition, responding correctly to almost all sentences, 95% correct. The difference between the Americans and the other two groups was significant and greater than in the MRT, suggesting that both knowledge of language and cultural background were probably helpful.

The performance of the other two groups was similar to their performance on the MRT, and not significantly different. Listening in noise not only caused a drop in overall performance but also a reduction in the differences among the three groups. The interaction of language background by listening condition was significant ($F = 8.86$, $p < .01$).

In the MRT, differences among the

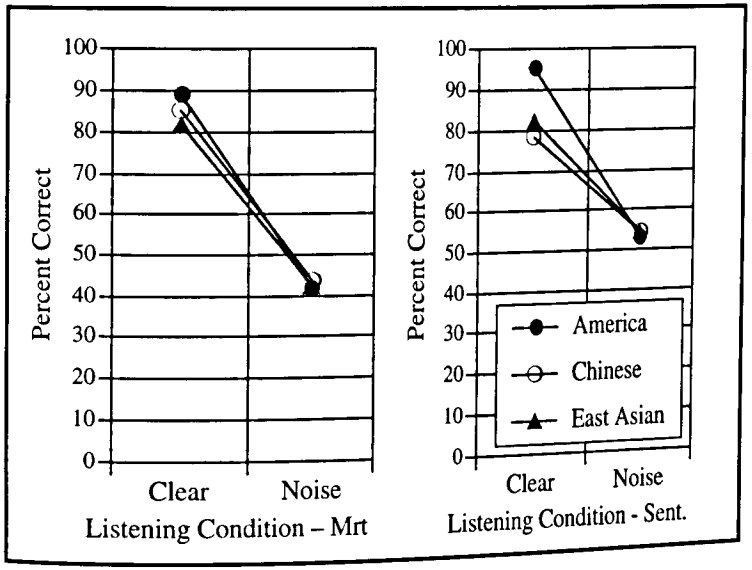


Figure 1.

three groups were relatively small and decreased in noise. Differences among the groups were greater in the sentence test, and the performance of the Americans decreased more than that of the other two groups when listening in noise. Language background, therefore, favored the American listeners for all test materials. The Chinese listeners performed better than the other East Asians only on the MRT, though it is possible that knowledge of language and culture were sufficiently different between the two groups to account for the differences in performance on sentences. When listening to speech in noise, differences between the groups were much reduced. For all three listener groups, the MRT was more difficult than the sentence test.

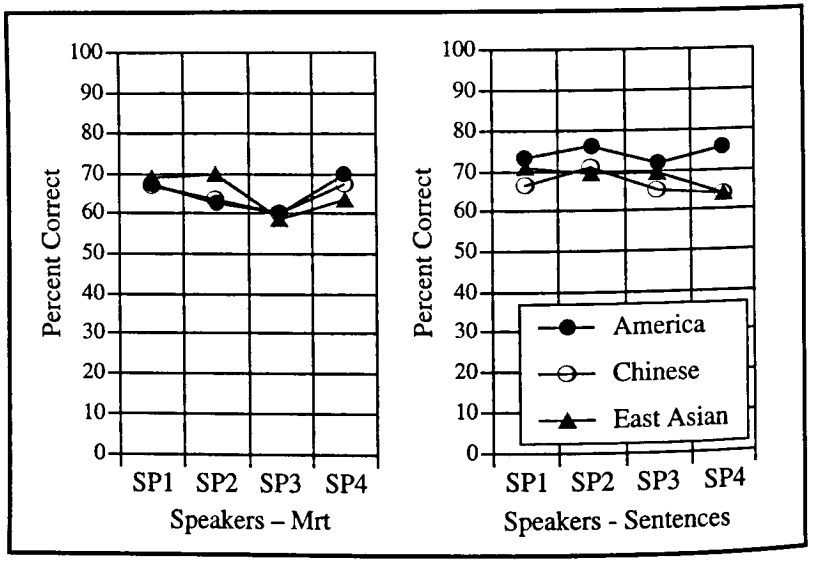


Figure 2.

Talkers

The percent correct responses to both the MRT and to the sentences for each talker are given in Fig. 2. In the MRT, Talkers 1 and 4 are some what easier to understand than Talkers 2 and 3. Differences between native language groups are greatest in responding to Talker 4 and small otherwise. Talker 3 was generally the least intelligible.

In responding to sentences, Americans are clearly superior to the other two groups. Differences between talkers were greater for the sentence test than in the MRT. Americans in particular found Talker 4 to be the most easy to understand. The language by talker interaction was significant ($F=2.88$, $p < .05$) as was the speaking condition by talker interaction ($F= 10.56$, $p<.01$). The three-way interaction was not significant.

We can conclude that talkers tend to vary in intelligibility somewhat, depending on the exact nature of an intelligibility test. However, a talker who is intelligible with one set of materials and in one speaking condition will remain relatively intelligible in other conditions.

Further, a talker who is relatively intelligible to one group of listeners will tend to be intelligible to other groups of listeners. These findings are in essential agreement with [2].

Ratings

Subjective ratings of intelligibility agreed with the test results. When ratings were correlated with intelligibility scores, the correlations tended to be high and significant. The exact values of the correlation coefficient were affected by the test itself, language back ground, and listening condition. For Americans, the correlations between ratings and test scores ranged between .52 to .92. For Chinese listeners, the range was .48 to .84. The ratings provided by the other East Asians were lower, ranging from .25 to .78.

The Americans tended to be most critical, judging the talkers relatively difficult to under stand, an overall rating of 3.6. The East Asians provided an over all rating almost as unfavorable as the Americans, 3.5. The Chinese rated the talkers the highest, 3.1 overall.

DISCUSSION

Proficient listeners from different language backgrounds differ in their ability to understand accented speech. Americans, in spite of little familiarity with the target accent, scored better than the other two groups, particularly when responses to the test required knowledge of language and culture. Sharing language background, as the Chinese listeners did, was not invariably an advantage.

Differences among listener groups were not consistent across listening conditions. The superior performance of the Americans decreased substantially when the listening conditions deteriorated both in the MRT and in the sentence test.

Differences among talkers were not perfectly consistent across tests materials and listening conditions. Americans, in particular, found some talkers much more intelligible than others. All three groups found talker 3 difficult to understand. Differences in intelligibility among talkers were relatively modest, though affected somewhat both by test and listening conditions.

Within language background groups, listener reactions to talkers and test conditions correlated quite well with their test performance. That is, listeners could make relative judgments of intelligibility with high reliability. The three groups of listeners differed in their overall evaluation of the talkers. Even though the performance of the Americans was better than that of the other two groups, they were most critical. The Chinese listeners were least critical, even though their performance was not nearly as good as the Americans.' The other East Asians were almost as critical of the talkers as the Americans.

The surprising finding in this study was the difficulty Americans experienced understanding non-native talkers in noisy listening conditions. Previous work has almost invariably found that non-native listeners experience much more difficulty than native listeners when presented with speech mixed with noise (see [3] and references cited there). In this previous work, talkers have always been native speakers of English.

Before the finding that native listeners have proportionately greater difficulty under standing non-native talkers in noisy conditions is accepted, this study must be replicated. It is possible that inadvertent differences in methodology, such as test item selection, test preparation, or test administration, are responsible for the decrement in the performance of Americans. To our knowledge, no directly comparable experiments are reported in the literature.

REFERENCES

[1] Pisoni, D. B., Manous, L. M. and Dedina, M. J. (1987), Comprehension of natural and synthetic speech: effects of predictability on the verification of sentences controlled for intelligibility, Computer Speech and Language, 2, 303-320.

[2] Bond, Z.S. and Moore, T. J. (1994), A note on the acoustic-phonetic characteristics of inadvertently clear speech, Speech Communication 14, 325-337.

[3] Bond, Z.S., Moore, T. J. and Gable, B. (1994), Listening in L2., ms., Ohio University.

NATIVE AND NON-NATIVE PERCEPTION OF DIALECTAL VARIATION IN SWEDISH

Una Cunningham-Andersson
Dept. of Linguistics, Stockholm University, Sweden

ABSTRACT

This study examines a single aspect of native speaker competence. The questions addressed here are: how well can a given non-native speaker perceive differences between dialects of Swedish? How well can native speakers of Swedish perceive this kind of variation? Does a long period of residence in Sweden and an apparently excellent command of the Swedish language imply that an immigrant's ability to place native speakers geographically approaches native standard? Is there an upper limit for how good non-native listeners can be, or can they approach native standard?

METHOD

Fifteen native speakers and thirty-three non-native speakers served as listeners. The listeners were divided into 5 groups: *NN0-1* (9 individuals) were non-native listeners who had spent one year or less in Sweden; *NN3-9* (9 listeners) had spent between three and nine years in Sweden; *NN12-17* (9), twelve to seventeen years in Sweden; *NN23-25* (6) twenty-three to twenty-five years in Sweden and *native* were the 15 native Swedish speakers.

Their standard of Swedish was judged on three dimensions: how well the listeners could express themselves in Swedish, how well they understood spoken Swedish and how good their pronunciation of Swedish was. Their perception of dialect was tested by having them attempt to discriminate between dialects presented in pairs and identify dialects as a forced choice between Malmö, Gothenburg, Uppsala, Umeå, Falun or Gotland. The collection of the material used as stimuli in these experiments is described in [1]. Comprehension (as measured here) can be seen from figure 1 to be a function of the length of time spent in Sweden, and reaches native standard after 12-17 years, or for some individuals, even earlier. Syntax and pronunciation appear to be much more difficult to learn, with only a few ever reaching native or near native standard. Those who do reach this standard may do so as early as after 6 or 9 years. Tables 1 and 2 show correlation matrices between the measured abilities and properties of the non-native and native listeners respectively. All the - values shown are significant at the 5% level.

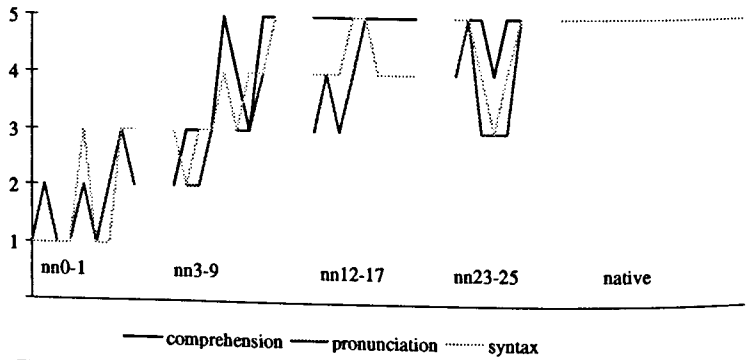


Figure 1.Individual language test results

From table 1 it can be seen that all three tested language skills, syntax, pronunciation and comprehension are strongly correlated with each other and also (somewhat less strongly) with the length of time the speakers had lived in Sweden and with the age of the speakers.

Table 1. Correlations between measured variables and results for all 33 non-native listeners

	age	years in Sweden	compre-hension	pro-nunciation	syntax	true *same*	false *same*
age	1						
years in Sweden	.65	1					
comprehension	.50	.78	1				
pronunciation	.40	.68	.85	1			
syntax	.46	.74	.92	.91	1		
true same	ns	ns	ns	ns	ns	1	
false same	-.36	-.55	-.68	-.48	-.61	ns	1
right dialect	.50	.84	.80	.74	.77	ns	-.60

DIALECT DISCRIMINATION

For the first listening task, dialect discrimination, the informants were to say whether pairs of speech samples were spoken by speakers from the same or different geographical regions. The listeners were told in advance that all the speakers came from one of the six places used in this study: Malmö, Gothenburg, Uppsala, Umeå, Falun or Gotland. The speech material used for this test was semi-spontaneous speech, where all speakers describe the same picture, so no non-phonetic information which may have helped the speakers identify dialects was likely to be present. Each of the six dialects was represented by two speakers. These were those judged as the most authentic by a panel of dialectologists.

Two speakers from each dialect gave twelve stimuli to be presented in 36 pairs. Previous work has shown that this kind of judgement can be made using short speech samples — listeners seem to be able to make up their minds very quickly about speakers (cf. e.g. [2]. Each stimulus was about 15 seconds long, and the two stimuli in a pair were separated by a tone. For each pair the listeners were to circle the words *same* or *different*. Instructions were given both on the answer sheet and orally on the stimulus tape in both Swedish and English.

Table 2. Correlations between measured variables and results for 15 native listeners

	age	true *same*	false *same*	right dialect
age	1			
true same	ns	1		
false same	ns	-.52	1	
right dialect	-.59	ns	-.45	1

There was considerable variation in the accuracy with which individuals in all groups were able to pick out the six pairs of speakers who spoke with the same dialect. The only listener to spot all six pairs had lived in Sweden no more than 4 years. There was overlap between the listener groups, and although the average score was highest for the native listeners, all the non-native listeners performed as well as or better than the worst native listeners. T-tests showed no significant difference between native and non-native listeners.

In many cases the listeners failed to detect dialectal differences between stimuli. Here there was a significant difference between native and non-native listeners ($p(t)>0.0001$): the non-native listeners more often failed to distinguish between dialects, although some of the non-native listeners who had been in Sweden 12 years or more were as skilful as the least skilful native listeners. Table 1 shows that length of residence in Sweden or competence on any of the three linguistic dimensions are not correlated to the ability to detect pairs of speakers of the same dialect. They are, however, significantly ($p(r)>0.01$) negatively correlated to the number of false *same* judgements, with comprehension and syntax skills having the strongest correlation to the ability to hear a difference between dialects when there is one.

For both native and non-native listeners, the Malmö dialect was not often confused with any other; the Umeå dialect was confused with all but the Malmö dialect; the Uppsala dialect was confused with the Gothenburg and Umeå dialects; the Gotland dialect was confused with Falun and Umeå; the Falun dialect was confused with Gothenburg, Umeå and Gotland and the Gotland dialect was confused with Umeå and Falun, although the native listeners had considerably fewer false *same* judgements than the non-native listeners as mentioned above.

DIALECT IDENTIFICATION

The final section of the listening test concerns listeners' ability to identify dialects. Semi-spontaneous speech from the four most authentic speakers of each dialect (as judged by the expert panel of dialectologists) was used for this test. This gave 24 speakers. The stimuli from each speaker were about 60 seconds long this time. The listeners were given a map of Sweden with the six places where the speakers originate marked on it for reference. They were also given an answer sheet for each speaker with the six place names on. They were instructed to circle the name of the place the speaker came from on the answer sheet. If they did not know, they were to guess.

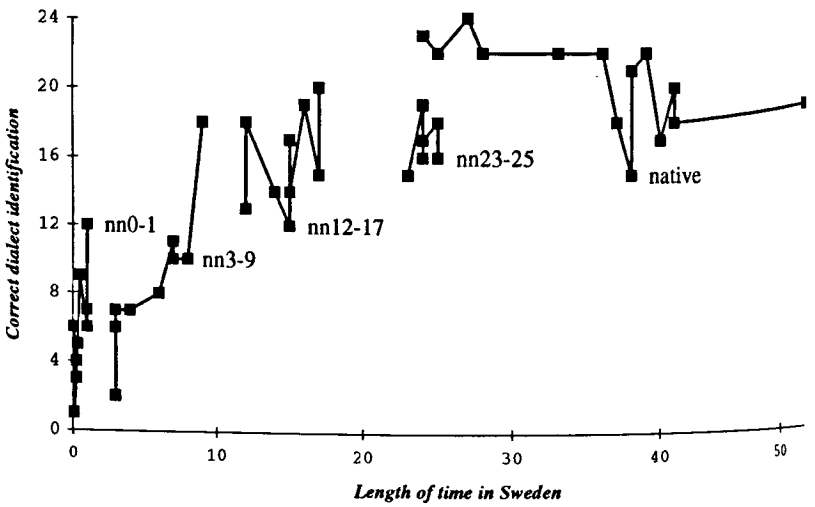


Figure 2. Number of correct dialect identifications for all listeners, plotted against the length of residence in Sweden

Figure 2 shows this to be a task that native listeners generally perform better than non-native listeners. A t-test shows a significant difference between native and non-native listeners here $p(t)<0.0001$, although figure 2 shows that the most accurate non-native listeners performed as well as the least accurate native listeners, one after as little as nine years residence in Sweden. An interesting feature which can be seen in tables 1 and 2 is that the length of time the non-native listeners have lived in Sweden is significantly correlated to their proficiency in the identification of dialects, even more strongly than their linguistic ability in Swedish, while the age of the native listeners is significantly *negatively* correlated to their degree of proficiency in this task.

Figure 3 shows how the different dialects were identified by the different kinds of listeners. Here too, the Malmö dialect seems to be the easiest to place correctly, closely followed by the Gotland dialect. Umeå and Falun seem to be much more difficult to identify accurately .

DISCUSSION

The questions posed at the beginning of this paper can now be answered. The results of the dialect discrimination test showed that many non-native listeners could distinguish between dialects as well as native listeners, even after a very short period of residence in Sweden. Non-native listeners were, however more likely to miss dialectal differences between speakers, and table 1 shows this to have more to do with their syntax and comprehension abilities than their pronunciation or the length of time they had been in Sweden. For both native and non-native

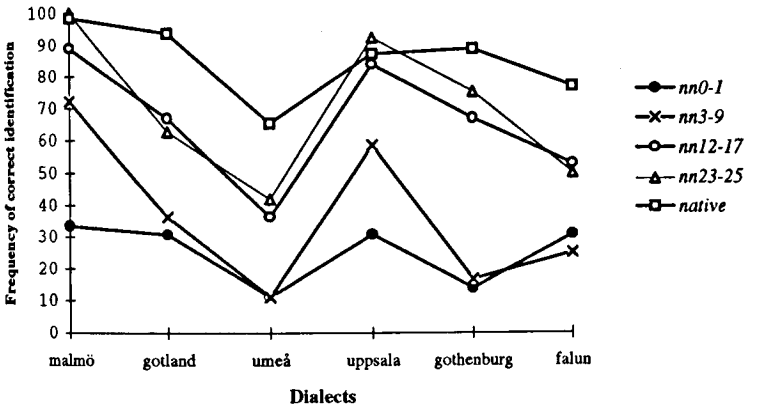


Figure 3. Frequency of correct identification of each dialect by each listener group

listeners there is, of course, a correlation between how many dialect distinctions the listener missed (false *sames*) and the number of speakers for whom they were able to correctly name the dialect.

The native listeners were significantly better than the non-native listeners at naming the six dialects they heard in a forced choice setting, although there was great variation between listeners in all groups. This ability may not have much to do with experience, since the older native listeners did not perform as well as the younger ones, and there were high scores in all but the least experienced listener group, NN0-1.

The conclusion here must be that few absolute differences exist between native and non-native listeners. The individual variation in achieved competence is very large regardless of the length of time spent in Sweden, although the listener's own competence in the Swedish language is related to sociodialectal awareness.

ACKNOWLEDGEMENTS

This work was financed by the Swedish Council for Research in the Humanities and Social Sciences.

REFERENCES

[1] Cunningham-Andersson, U. & Engstrand, O. 1994. "Data collection for a sociodialectal study". In G. Melchers & N.-L. Johannesson (eds.) *Non-standard varieties of language, Papers from the Stockholm symposium 11-13 April 1991*, 11-23. Almqvist & Wiksell International. Stockholm Sweden.

[2] Cunningham-Andersson, U. & Engstrand, O. (1989). "Perceived strength and identity of foreign accent in Swedish". *Phonetica* 46 (4), 138-154.

APPLYING PHONETICS TO THE TEACHING OF ENGLISH

Rebecca M. Dauer
University of Massachusetts, Amherst, MA, USA

ABSTRACT

Second language learners need both accurate phonetic details and a way of accessing them directly in order to improve their pronunciation of English. Examples of methods to teach tense and lax vowels, voiced and voiceless consonants, and stress and rhythm will be described.

INTRODUCTION

Although much research has been done on the phonetics of spoken English, only a little of this information has been incorporated into the teaching of English pronunciation to non-native speakers. Textbooks tend to use phonological labels, such as "tense-lax," "voiced-voiceless," and "stress," along with two-dimensional diagrams of "tongue positions" as explanations. Not only are these labels generally incomprehensible to the learner, but they may even mislead him by ignoring the major phonetic difference involved in a contrast. Learners need accurate phonetic details reflecting how English is actually spoken and processed in continuous speech, as well as a way of accessing this information directly, through their own experience, instead of merely understanding on an intellectual level. The same articulatory techniques used in teaching general phonetics to native speakers [1, 2] can be applied to teaching pronunciation to adult second language learners. These include becoming aware of one's own articulatory mechanism and being able to isolate, control, and recombine features in new contexts.

TENSE AND LAX VOWELS

One of the most difficult contrasts for non-native speakers to learn is the difference between the so-called tense and lax English vowels, /i - ɪ, u - ʊ/ (as in *seat-sit, suit-soot*). Writers of textbooks quite literally interpret this to mean that the muscles are either tense or relaxed, and advise students: "Say /i/ and then RELAX your tongue without moving it. This is the /ɪ/ sound" [3]. This kind of advice is accompanied by the typical line drawings of a sagittal section with tongue positions. As Ladefoged [2] points out, the difference has to do with the types of syllables these sounds can occur in; diagrams of tongue positions merely reflect the relationship of vowels in auditory space. In addition, students are told that /i/ and /u/ are long and /ɪ/ and /ʊ/ are short (often transcribed as /iː, uː/ vs /i, u/. In fact, all vowels can be long or short; length is determined primarily by the following consonant and stress; /ɪ/ in *bid* is longer than /i/ in *beat*; /ʊ/ in *should* (stressed) is longer than /u/ in *shoot*.

Teaching Suggestions

Students need to first find /ɪ/ and /ʊ/ and feel how they differ from /i/ and /u/. They can do this by gliding very slowly from /i/ to /ɑ/ (or /u/ to /ɛ/) a few times, both silently and aloud. Then they do it again, but stop as soon as they move away from /i/ (/u/); they hold on to this vowel and try some words that use it: *it, hid,... (hood, look)*. Or they can try getting at these the other direction, by slowly saying /aɪ/ (/aʊ/) and stopping just before the end. This method can also be used to find /ə/, another difficult vowel for non-native speakers: glide from /i/ to /ɑ/ but stop about half way, before your mouth opens completely; or glide from /ɛ/ to /ɔ/ and stop half way, before the lips start to round.

Lip position is also important for /ɪ, ʊ, ə/. Students can develop both kinesthetic and visual awareness of their lips by gliding continuously back and forth from /i/ to /u/, from /ɛ/ to /ɔ/, aloud while observing their lips in a mirror and silently (with eyes closed). They should then alternate /i - ɪ/, /u - ʊ/, making sure that the lips are close to neutral during /ɪ/ and /ʊ/.

Finally, they need to practice making all these vowels both long and short *without changing their quality*. The way to lengthen /ɪ/ or /ʊ/ is to move the tongue down and towards central /ə/--[ɪə], [ʊə] (*hid* [hɪəd], *hood*[hʊəd]), while they must keep pushing up and forward/back for /i/ and /u/. Once they have developed the kinesthetic awareness and ability to say these vowels in isolation, they are ready to do intensive practice in words, sentences, and longer discourse [4]. Minimal pair practice should include not only the traditional *seat-sit, heed-hid* type, but also *reach-ridge, heat-hid,* where short /i/ contrasts with long /ɪ/. This is difficult for most learners: they tend to pronounce /ɪ/ either too long or change it to /i/ in a voiceless environment, and are unable to stretch out /ɪ/ without changing it to /i/ in a voiced environment.

For some learners, even a preceding consonant can influence vowel quality; /l/ and /r/ can cause the tongue to retract so that *leaving-living*, or *reason-risen* sound alike. To overcome this, they should try saying the words without the initial consonant (*eaving-iving*), and then try to maintain the distinction while adding the consonant.

VOICED AND VOICELESS CONSONANTS

The teaching of voiced and voiceless consonants is another area where textbooks oversimplify and may even obfuscate the truth. Labeling is not enough: students need to be taught how to make sounds voiced and voiceless and given a method for self verification. As phoneticians have known for years, the major difference between initial and final stops in English is not voicing of the consonant, but rather aspiration and vowel length [2, 5]. For a learner trying to speak the language, the allophonic variations or phonetic realization rules are most important, yet are often overlooked.

Teaching Suggestions

Just like students in a phonetics class, learners can become aware of voicing by alternating back and forth between voiced and voiceless fricatives /szsz; fvfv; ʃʒʃʒ; θðθð/ [1]. They should try this at different speeds until they gain enough control to turn on and off voicing at will. By putting a hand over their larynx, they will be able to feel the voicing and experience it directly. For most learners, the voiceless fricatives are easier; by observing themselves in a mirror, they can check to make sure they are making no other changes to the tongue or lips when they add voicing. They need to separate what is going on in the larynx from what is going on in their mouths. It's also a good idea to have them try devoicing /r, l, m, n/ to prove they really have separate control over the larynx. This procedure enables many students to produce consonants (such as /ð/) which they had been unable to make before. Of course, further practice is needed in different environments in words and sentences. Textbooks seem to favor initial position, but it is not always the easiest. For example, the voiced fricatives are often easier to produce between vowels so that *mother* is easier to pronounce correctly than *this*; some find that /θ/ is easier to pronounce in final position (*math*) than initial position (*thing*). Students should experiment to find the words in which they can pronounce the sounds the best to use as starting points.

Final Consonants

Teaching the final consonants, particularly the stops and affricates, is entirely a different matter. The most common mistake for learners is to make them voiceless, which besides causing confusion in monosyllabic words, results in an unpleasant choppy rhythm. If students are told to lengthen the preceding *vowel* and leave the final consonant entirely voiceless, letting it naturally die out before a pause, most will be able to pronounce words like *bed, bag, badge, please* perfectly. On the other hand, if they strive to make the final consonant fully voiced, they often make it too long, too strong, or add an epenthetic [ə]. They need put their effort into the vowel, and weaken and shorten the consonant. In addition, they must link final consonants smoothly to following words, resyllabifying if the following word begins with a vowel (*find it = fin dit*), remembering to drop unstressed /h/. This has the added advantage of breaking up final consonant clusters and thus

making them easier to pronounce (*change his = changes* /tʃeɪn dʒɪz/). Respelling words can counteract the strong belief in the existence of spaces between words.

Other techniques for teaching final consonant clusters are to isolate and practice them as nonsense sounds /sks, tʃt, tʃtʃ, dʒdʒ/ and then try them in words, *desks, watched, which channel, orange juice*. Articulating words silently is another excellent method to help learners get in touch with and gain control over their speech organs. It can help them to coarticulate (link consonants together) as in *keep quiet, back door, right time* because they can verify more easily whether they are truly beginning the second consonant before releasing the first one.

Aspiration

Many non-native speakers do distinguish initial voiced and voiceless stop consonants, but they do it entirely by voicing--voiceless unaspirated versus fully voiced. In fact, aspiration is more important than voicing in distinguishing initial stops in English. Textbooks often teach aspiration by having students try to blow on a piece of paper to make it move. Some students will take a breath, tense up (i.e. make a glottal stop), then release a very crisp, yet totally unaspirated stop. The important thing is that aspirated sounds are actually more *relaxed* than unaspirated stops. In order for there to be a delay in voice onset time, the glottis must be open. This is more easily achieved by having students sigh and say words like *a key, a tie, a pie* on a long exhalation, as if they were very tired.

A fact that is usually ignored by textbooks is that /p, t, k/ are also aspirated when followed by /l/ and /r/. Students need to get the same relaxed feeling in *cold, crowd, cloud; pay, pray, play*. If they have been taught how to make any sound voiced or voiceless, they can devoice /r/ or /l/ (or think of saying /h/ and /r/ or /l/ at the same time), add the following vowel, and then the initial stop: [hr], [hreɪ], [phreɪ]. Aspirating, along with lengthening /r/ and /l/ in these clusters, also makes them more distinguishable from each other as well as from their voiced counterparts. Substituting completely voiceless, but unaspirated, stops for initial /b, d, g/ in these clusters can help those who tend to weaken them.

STRESS AND RHYTHM

In teaching stress and rhythm, modern ESL textbooks tend to emphasize listening. Stress is usually treated as unpredictable. Definitions of stress and reduction are often not sufficiently detailed: stress tends to be identified with high pitch, and reduction is simply a matter of using the vowel /ə/. The contribution of length, linking, and pausing to rhythm is often slighted.

Teaching Suggestions

The single most important thing that a non-native speaker can do to improve his comprehensibility is to slow down and pause more often. This is in fact very easy to teach since most students already know where they should pause. The necessity for slowing down becomes obvious even to naive students when they record and listen to themselves. Not only do pauses give the listener extra time to process what he hears, but they also give the speaker more time for speech planning. When learners focus on pausing more frequently and for a longer time, they often discover that other problems, such as omitting final consonants, disappear. Slowing down means stretching out stressed syllables, particularly monosyllabic content words, and slowing down before an actual or potential pause. Unstressed syllables may be pronounced as quickly as possible without dropping them and should be linked smoothly to other words in the phrase.

Vowel Reduction

Although reducing unstressed syllables and function words is essential to English, many learners have difficulty pronouncing /ə/, tending to pronounce it more like /ɑ/ or /ɔ/ (if spelled <o>). Students usually have much more success if they are told to pronounce unstressed vowels as /ɪ/ (e.g. *Washington, today, can, was, that*): it is inherently short and has no lip rounding. They can try pronouncing function words such as *can, was, some, them*, with no vowel: [kn, wz, sm, ðm]. This keeps them from opening their mouths too much and making the syllable too long. In practice, they need to concentrate on pronouncing the function words quickly and weakly, yet without shortening stressed syllables. Their goal should be to maximize the difference between stressed and unstressed syllables.

Since stress is primarily auditory, it doesn't lend itself as well to kinesthetic and visual feedback. However, students should be able to see a difference in the degree their mouths open in minimal pairs such as *a contract, to contract, an addict, to addict*, with a mirror or on videotape. The mouth should be almost closed (a hand can be put under the chin to keep it from opening if necessary) during reduced syllables. Noun-verb minimal pairs such as these are very useful in clearly pointing out to a listener whether a student is reducing enough.

Stress

Stress is predictable in the majority of words, and the rules are not difficult to teach [4]. When students in a class disagree or when they are unsure about which syllable is stressed, the teacher should say the word all possible ways (e.g. *'develop, de'velop, deve'lop*) to see if they can recognize which sounds right. They should then try to develop this capacity to stress any syllable of a word themselves. They also need practice moving stress in different forms of a word and alternating reduced and full vowels (*academy, academic*).

Words with various stress patterns also need to be practiced when they are non-tonic (without sentence stress). A common error is the automatic association of word stress with high pitch. This leads some students to jump up on every stressed syllable; as a result, their speech may sound choppy, words are not heard as grouped into larger syntactic units, and there is no focus or continuity in discourse. These students need to start each phrase on a low to mid pitch and delay jumping up to their highest pitch until they reach the tonic, which is typically towards the end of a pause group. Other students tend to destress (shorten excessively) all non-tonic stressed syllables; thus, words are all run together, it's hard for the listener to establish word boundaries, and words at the beginning of a sentence may be lost. These students need to concentrate on lengthening and clearly articulating all stressed syllables, even when they have no pitch prominence.

CONCLUSION

Just like students in phonetics classes, ESL students need to learn how to separate, independently vary, and then recombine phonetic features through experimenting with their own vocal tracts and moving from known to unknown. If they are to improve their production, they need to be given specific articulatory strategies as well as accurate descriptions of how fluent English is spoken. For most students, improvements in production also lead to better perception and comprehension of oral English.

REFERENCES

[1] Catford, J. C. (1988), *A practical introduction to phonetics*, New York: Oxford University Press.

[2] Ladefoged, P. (1975), *A course in phonetics*, New York: Harcourt Brace Jovanovich.

[3] Hagen, S. A. & Grogan, P. E. (1992), *Sound advantage: A pronunciation book*, Englewood Cliffs, NJ: Prentice Hall Regents.

[4] Dauer, R. M. (1993), *Accurate English: A complete course in pronunciation*, Englewood Cliffs, NJ: Prentice Hall Regents.

[5] Peterson, G. E. & Lehiste, I. (1960), "Duration of syllable nuclei in English," *JASA* vol. 32, pp. 693-703.

THE ACQUISITION OF MULTILINGUAL PHONOLOGY

Eduardo D. Faingold
State University of New York at Stony Brook

ABSTRACT

I am concerned with processes and strategies of early phonological and lexical development in multilingual children--Spanish, Portuguese, and Hebrew vs. English and Hebrew. The simultaneous acquisition of closely related languages such as Spanish and Portuguese vs. that of non-related languages such as English and Hebrew yields different results: The former 'prefer' maintenance, while the latter 'prefer' reduction. The Spanish and Portuguese-speaking children's high accuracy stems from a wider choice of target words, where the diachronic development of two closely related languages provides a choice of simplified words.

INTRODUCTION

Berman's study of the simultaneous acquisition of Hebrew and English phonology and lexicon discusses her daughter Shelli's strategy of reducing the number of syllables [1]. She also discusses the universally observed deletion of a final and initial consonant and the deletion of one member of a vocalic or consonant cluster; and she further presents a small number of 'full' words and a limited use of reduplication and transposition. Table 1 shows the breakdown of Shelli's first 175 words.

Table 1. Shelli's vocabulary (1;6;0 - 1;11;15) (Berman 1977)

	number	%
'Full' words	50	29
Reduction	100	57
Reduplication	10	5
Transposition	15	9
Total	175	100

Shelli's phonological development was contrasted to that of Noam's simultaneous acquisition of Spanish, Portuguese, and Hebrew [2]. Unlike Shelli, Noam's lexicon shows maintenance--a large number of perfect replicas of adult words, as well as other 'full' structures. Both children show: (a) the universally observed deletion of final and initial consonants; (b) the deletion of one member of the consonantal or vocalic clusters; and (c) a small number of transpositions in words which, presumably, present difficulties. Table 2 shows the breakdown of Noam's first 175 words.

Table 2. Noam's vocabulary (1;1;2 - 1;9;0) (Faingold 1990)

	number	%
'Full' words	79	45
Reduction	43	25
Reduplication	41	23
Transposition	12	7
Total	175	100

In sum, Noam and Shelli show two opposite strategies in the construction of their early lexicon--maintenance vs. reduction. The process of syllable reduction is focal in Shelli and minimal in Noam, while the use of syllable maintenance and reduplication is focal in Noam and marginal in Shelli. Final and initial consonant deletion and vocalic and consonant cluster reduction are systematic, universal, and language-independent in child language in general [4], and are also manifested in both Noam and Shelli, despite their different strategies and linguistic input. There are, however, quantitatively speaking, many more 'full' words in Noam's than in Shelli's lexicon.

DATA COLLECTION AND ANALYSIS

Nurit was visited two or three times a week from age 0;11 to 1;11 in her home in Jerusalem (Israel). Stimulus materials included picture books, drawings, and any object in her home. I kept a diary of Nurit, using the same transcription level as for Noam [2], [3]. Unlike Noam, Nurit was not tape-recorded, since this was felt as an intrusion into the home of Nurit's parents. The principle of "one person, one language" [5] was observed consistently by both Noam's and Nurit's parents and sitters. Both children were thus exposed to the same varieties of (La Plata) Argentine Spanish and (Saõ Paulo) Brazilian Portuguese.

As with [1] and [2], the data in this paper refers to one-word utterances produced by Nurit which offer a clear and consistent semantic interpretation; only words produced spontaneously were considered as part of her active vocabulary, while all imitations were excluded. The quantification and analysis of Nurit's data, i.e. the use of percentages, follows [1] and [2]. The results in this paper are thus fully comparable to those discussed by these authors.

NURIT'S PROCESSES AND STRATEGIES

Table 3 shows the breakdown of Nurit's first 73 words.

Table 3. Nurit's vocabulary (1;2;15 - 1;11;0)

	number	%
'Full' words	39	54
Reduction	22	30
Reduplication	9	12
Transposition	3	4
Total	73	100

'Full' words

As with Noam, Nurit produced a large number of 'full' words. Table 3 presents 39 'full' words (54% of the total). This set is divided into (i) perfect replicas (e.g. Pt. [ke] 'want', Sp. [papá] 'father'), and (ii) replicas with substitution (e.g. [mei] < Pt. [meu] 'mine', [tau] < Sp. [tŝau] 'bye'). Perfect replicas are copies of adult words in one of the input languages to which the child is exposed.

Reduction

Table 3 presents 22 words that suffer reduction (30% of the total). This set is divided into (i) reduction of segments (e.g. [ki] < Sp., Pt. [aki] 'here', [ma] < Sp. [mas] 'more') and (ii) reduction of syllables (e.g. [bo] < Pt. [bola] 'ball', [nana] <

Sp., Pt. [banana] 'banana'). Segments are consistently deleted in all positions by universally observed constraints on the production of initial and final consonants, as well as vocalic and consonant clusters. As with Noam, in (ii) syllables are deleted only occasionally, and they are generally maintained by reduplication as well as by the production of 'full' words.

As with Noam, Nurit's use of syllable reduction is marginal. However, both children present a significant number of deleted segments in initial and final and cluster positions, since this is a universal process of child language acquisition.

Reduplication

Table 3 presents 9 cases of reduplication (12% of the total). A reduplicated structure is a segment or a syllable that is not in reduplicated form in the input language. This set is divided into reduplication of (i) segments (e.g. [eme] < Sp., Pt. [kome] 'eat', [tota] < Hbw [toda] 'thanks'), and (ii) syllables (e.g. [papa] < Sp., Pt. [paula] 'Paula', [bobo] < Pt. [akabo] 'all gone'). Like Noam, she makes a creative use of harmonization rules to match the syllabic patterns of the adult model [2], [4]. However, unlike Noam, she presents a smaller number of reduplicated segments and syllables. The reason appears to be that while Noam capitalizes equally on both 'full' words and reduplication to produce a higher number of words that match the syllabic patterns of the adult input, Nurit uses almost exclusively 'full' words--which are mostly perfect replicas of the adult model. Nurit seems to be a slower and cautious learner, yet more accurate at hitting at adult targets, since her vocabulary is smaller yet more adult-like. Both children, however, obtain similar results. Despite the slightly different approaches to maintenance and the different number of items produced by each child (Noam's 175 vs. Nurit's 73) at comparable stages in language acquisition, their selection of phonological processes remains, in quantitative terms, almost identical for both children (compare Table 2 vs. Table 3).

Transpositions

Table 3 notes three cases of transposition ([lei] < Pt. [le] 'read', [toi] < Pt. [istoriña] 'tale', [abi] < Hbw. [omRi] 'Omri'). Nurit's use of transpositions is much lower than Noam's (or Shelli's). As noted, Nurit seems to be a much more cautious and accurate learner, and this fact probably accounts for her small number of transpositions as well as for her high number of 'full' words.

MAINTENANCE: A LANGUAGE-DEPENDENT STRATEGY

The majority of the words produced by Noam and Nurit are perfect replicas of adult words in one of the input languages to which the child is exposed. From the point of view of another cognate word in another input language, they are producing reduced versions of the adult word. For example, Nurit's [ke] 'I want' and [sai] 'go away' are perfect replicas of the Portuguese adult models, but they are also reduced versions of the Spanish cognates [kiero] and [salí]; similarly, Nurit's [papá] 'father' is a perfect replica of the Spanish adult model but it is also a reduced version of the Portuguese cognate [papai], all of which Nurit understands. In parallel fashion, as I have shown in [2], Noam produced adult Spanish (e.g. [si], 'yes' [asi] 'in this manner', [papá] 'father', [mamá] 'mother') and Portuguese words (e.g. [sai] 'go away') which might be in fact reduced versions of the adult Portuguese ([sim], [asim], [papai], [mamai]) and Spanish cognates ([salí]), all of which Noam understands. Even though [ke], [sai], [papá], [mamá], [si], [asi], etc. might be reductions of [kiero], [salí], [papai], [mamai], [sim], [asim], etc., they are all still 'full' words in one of the input languages. In this sense, all the perfect replicas produced by Noam and Nurit are 'legitimate' lexical items, since no deviations of the adult patterns occur.

Variation in the application of maintenance and reduction rules by Noam and Nurit vs. Shelli is thus systematic. The Spanish/Portuguese-speaking children's high rate of phonological and lexical accuracy results from a wider choice of target words, where the diachronic development of two closely related languages provides a simplified but legitimate model lexicon to the child. Thus, Nurit and Noam's high number of perfect replicas might be the result of an exploitation strategy or phonological preference [6]. In contrast, the simultaneous acquisition of unrelated languages such as English and Hebrew yields different results--a low number of adult replicas, as well as little reduplication and a high number of reduced syllables.

REFERENCES

[1] Berman, R. 1977. Natural phonological processes at the one word stage. *Lingua* 43:1-21.

[2] Faingold, E. D. 1990. The acquisition of syllabic and word structure: Individual differences and universal constraints. *Language Sciences* 12:101-113.

[3] Faingold, E. D. in press. *Child Language, Creolization, and Historical Change. Spanish in Contact with Portuguese*. Tubingen: Narr.

[4] Ingram, D. 1979. Phonological patterns in the speech of young children. *Language Acquisition*, pp.133-48. Paul Fletcher & Michael Garman (eds.). Cambridge: University Press.

[5] Leopold, W. 1939. *Speech Development of a Bilingual Child: A Linguist's Record. Vol 1*. Evanston: Northwestern University Press.

[6] Menyuk, P. & Menn, L. 1979. Early strategies for the perception and production of words. *Language Acquisition*, pp.49-70. Paul Fletcher & Michael Garman (eds.). Cambridge: University Press.

NATURALISTIC ELICITATION OF ACOUSTICAL PARAMETER SHIFT OF /r/ AND /l/ IN THE SPEECH OF JAPANESE ESL STUDENTS

S. N. Gynan and R. Weiss
Western Washington University, Bellingham, Washington, USA

ABSTRACT

This study documents the utility of naturalistic elicitation for identifying shift in parameters of interlanguage phonetic variables. Pre- to post-test production of /r/ and /l/ increased. /l/ was produced incorrectly far more often than /r/. The difference in accuracy between /r/ and /l/ was greater pre-vocalically than post-vocalically. /r/ is heard as [l] and /l/ as [ɹ] only pre-vocalically. For /r/, rhoticization increases gradually over time, and for /l/, lateralization increases gradually.

INTRODUCTION

The difficulty that speakers of Japanese have in distinguishing between the English liquids /r/ and /l/ is due to a number of factors. Japanese has one liquid phoneme, represented as /r/ and found intervocalically and word-initially, whereas the English liquids are found also in consonant clusters and post-vocalically. The Japanese /r/ has several allophones, including [l] and [ɹ], which resemble the principle allophones of the English liquids. The distribution of allophones of /r/ is, according to Vance, not at all inflexible, with substantial variation in pronunciation (see Table 1). In word-initial position, the onset of the articulation of Japanese /r/ is typically characterized by the location of the blade of the tongue on the alveolar ridge, producing a sound acoustically more similar to English [ɬ] than [ɹ]. Intervocalically, Japanese /r/ is often pronounced as a flap, [ɾ], a allophone of /r/ in some registers and dialects of English. Finally, Japanese /r/ assimilates to the following vowel, especially the high front vowel and semiconsonant.

Dickerson and Dickerson, in a study of Japanese accented-English, document the influence of phonetic environment on the articulation of /r/, noting accuracy of articulation depends on phonetic context, and higher accuracy before low vowels than high vowels, a phenomenon which reflects the assimilatory qualities of high and especially high front vowels in Japanese. Variability of pronunciation of /r/ is therefore conditioned by the phonetic context, as it is in Japanese.

Vance reports Japanese /r/ being labeled a variphone, a reference to the fluid nature of the sound. In English /r/ is also subject to variation, as are all vowel sounds. Labov notes that the range of variation in vowels can be represented on two dimensional formant displays as elliptical "envelopes" that partially overlap. The shape and size of the envelopes depend in part on the phonetic environment, as with Japanese /r/, and also on the sociolinguistic environment. The students in this study report being highly motivated to achieve native-like proficiency. They are likely to experience shifts in the shape and size of envelopes. It is this parameter shift which is ultimately of interest to the authors of the present study.

Table 1. Allophonic variants of /r/ and /l/ identified in the interlanguage of Japanese ESL students.

[r] voiced apico-alveolar trill

[ɾ] voiced alveolar flap

[l] voiced alveolar lateral approximant

[ɭ] voiced post-alveolar (retroflex) lateral approximant

[ø] omission of /r/ or /l/

[ɬ] voiced velarized alveolar lateral approximant

[ɹ] voiced alveolar frictionless continuant

[ɻ] excrescent vowel with /l/ or /r/

[ɹʷ] voiced labialized alveolar frictionless continuant

[ɚ] vocalized /r/

[uʷ] vocalized /l/

[ɖ] voiced retroflex stop

OBJECTIVES

One goal of this study was to identify phonetic variables that might be especially sensitive indicators of second language acquisition. The differentiation by native speakers of Japanese of the English phonemes /r/ and /l/ is a good focus for such a study because some of the allophones belonging to these phonemes are continuants or approximants, and it was felt that approximation of target-language norms could be incremental instead of segmental, and thus would be an appropriate target for acoustical analysis. A second goal was to determine the utility of naturalistic elicitation for carrying out second language phonetic analysis of an experimental variety. Specifically it was felt that naturalistic data, usually considered relatively uncontrolled, could be controlled sufficiently to allow for identification of the environments where pronunciation of these phonemes is most frequent, a pedagogical concern, and would allow for identification of the allophones which would most likely display "sensitivity" to the effects of target-language input to the phonological system.

SUBJECTS AND PROCEDURES

Subjects

This study involved 40 native Japanese ESL students, who were part of a larger study (originally 80 students) designed to measure changes in reading and speaking skills as well as attitudes as a result of their four-month study at Western Washington University. The subjects (ages 17 to 22) had had an average of seven years of English study in Japan.

Elicitation Procedures

The elicitation device required students to develop an oral composition about a series of pictures which depicted the past, present, and future of transportation. Students were requested to talk about the pictures for about one minute. They were assisted by the experimenter and assistants in carrying out the task of recording themselves in a laboratory setting equipped with high-quality Tandberg cassette recorders. The initial pre-test recordings were made in November, 1992 and repeated three months later as a post-test in February, 1993. The recordings varied in length between one and four minutes, with the average just over a minute. The complexity and sophistication of the morphological syntactic structures and accuracy varied considerably; however, the picture presentations elicited a number of repeated words, for example, virtually all subjects used words such as: *balloon, car, airplane, railroad, train, horse, shuttle,* etc.

The purpose for using such a technique is multifold. Speaking ability, commonly held to be the most critical language skill and the principle reason for students' study abroad, is not often tested. Interviewing techniques have been carefully refined over the years, but are still cumbersome, requiring highly trained personnel and a great deal of time. For experimental purposes they are also problematic because the interviewer actually intervenes in the procedure and contaminates the data. The current technique, however, affords several advantages over traditional oral proficiency measures: 1) it is non-interventionist, allowing the student to produce the narration; 2) many students can be tested at once; 3) vocabulary and verb tenses can be controlled by use of the pictures; and 4) no printed word interferes with the student's performance. The language thus elicited is more systematic.

For the object of study, i.e., parametric shifts of /r/ and /l/, specific phonetic environments ideally should be controlled to facilitate the precise study of the articulatory shifts. For the experimental phonetician the technique used may not afford sufficient control. This objection notwithstanding, the oral composition technique proved to be a reliable means of eliciting phonetic data in a variety of significant phonetic environments and that faithfully reflects the transitional linguistic competence of the speakers who produced it.

ANALYSIS PROCEDURE

A very broad transcription of the pre- and post test recordings was made by the authors. Each occurrence of /r/ and /l/ was noted with its environment. This stage of the analysis is presented in Table 2. Once environments were identified, phonemic distributions were tallied. These are presented in Table 3. Finally,

percentages of correct and incorrect responses were calculated, and major interlanguage allophonic variants of the phonemes along with their frequency of occurrence were identified. These data are presented in Table 4 and in Table 5.

RESULTS

Evidence for the limitations of the naturalistic elicitation method are documented in Table 2, where the seven of the original eleven environments are listed. Originally it was planned to examine the effect of the following vowel on the articulation of the liquids, but not enough tokens of liquids followed by different vowels were produced to enable a meaningful comparison. Consequently, post-vocalic environments were identified irrespective of phonetic features of vowels.

Table 2. Phonetic environments identified in the interlanguage of Japanese ESL students.

Formula	Description
# V	Word-initial, post-pausal
C + V	Syllable-initial, post-consonantal
V V	Intervocalic
C___V	Consonant cluster, pre-vocalic
V___C	Consonant cluster, post-vocalic
V #	Post-vocalic
C #	Syllabic

In Table 3, two more environments prove to have an insufficient number of tokens to allow for pre- and post-comparison of pronunciations of /r/ and /l/. Very few instances of either phoneme were found in syllable-initial position post-consonantally. Syllable and word-final consonant clusters consisting of obstruent plus liquid were underrepresented for the phoneme /l/. A successful analysis of /r/ and /l/ in these environments, as well as the pre-vocalic ones, would require a larger database of sounds elicited naturalistically, as many as 5,000 tokens of /r/ and /l/. Based on the current study, a total of 80 complete sets of data from subjects would be sufficient.

Across-the-board, but differential quantitative increases in production of /r/ and /l/ are documented in Table 3. The post test yielded 28% more tokens than the pretest ((1474-1148)/1148).

Table 3. Tokens of /r/ and /l/ identified in pre- and post-tests.

	/r/		/l/	
	pre-test	post-test	pre-test	post-test
# V	25	33	44	55
C+ V	2	4	2	5
V V	113	157	98	114
C V	175	190	84	120
V C	96	86	5	23
V #	255	309	51	105
C #	70	106	128	167
Total	736	885	412	589

There were 610, or 61% more /r/ tokens produced than /l/ tokens ((1621-1001)/1001). The difference between /r/ and /l/ expressed as percentage of /l/ was greater in the pre-tests than in post-tests (pre: (736-412)/:412=79%; post: (885-589)/589=50%)). Finally, whereas the pre- to post-test increase in production of the phoneme /r/ was 20%, the increase in the phoneme /l/ was 43%. Qualitative increases in and differences between pronunciation of /r/ and /l/ are also in evidence in Table 4. American /l/ was produced incorrectly far more often than /r/. The increase in correctness of /l/ went from 8% in the pre-test to 12% in the post-test. For /r/ these figures were 35% and 40%, respectively. Production of correct /r/ and /l/ variants improved in the post-test, /l/ by 53% and /r/ by 13%. Finally, the difference in accuracy

Table 4. Percentages of target-like /r/ and /l/ in pre- and post-tests.

	/r/		/l/	
	pre-test	post-test	pre-test	post-test
# V	24	33	2	9
V V	31	45	2	10
C V	68	70	5	2
V C	30	29	10	14
C #	39	43	14	16
Total	35	40	8	12

Table 5. Principal non-target-like allophones of /r/ and /l/.

	/r/				/l/			
	Pre-test	%	Post-test	%	Pre-test	%	Post-test	%
# V	[l]	36	[l]	12	[ɾ]	59	[ɾ]	65
	[ɾ]	16	[ɾ]	42	[l]	27	[l]	20
V V	[ø]	41	[ø]	28	[ɾ]	54	[ɾ]	46
	[ɾ]	11	[ɾ]	14	[ɹ]	22	[ɹ]	27
C V	[ɹʷ]	13	[ɹʷ]	15	[ɹ]	61	[ɹ]	64
	[ɾ]	12	[ɾ]	10	[ɾ]	29	[ɾ]	28
V C	[ø]	61	[ø]	59	[l]	35	[l]	37
	[ɚ]	4	[ɚ]	6	[l]	33	[l]	29
C #	[ø]	57	[ø]	53	[l]	42	[l]	47
	[ɹʷ]	1	[ɹʷ]	1	[l]	27	[l]	22

between /r/ and /l/ was much greater in pre-vocalic than in post-vocalic or syllabic environments.

Differential errors in pronunciation of /r/ and /l/ are illustrated Table 5. The pronunciation of /r/ as [l] and of /l/ as [ɹ] is only heard pre-vocalically. In post-vocalic and syllabic positions, /r/ and /l/, even when pronounced incorrectly, are realized with articulations roughly approximating the target-language norm. Evidence for parameter shift is also revealed in Table 5. Approximation of /r/ appears to occur in stages: [l] > [ɾ] > [ø] > [ɚ] > [ɹʷ] > [ɹ]. Approximation of /l/ appears to follow this sequence: [ɹ] > [ɾ] > [l] > [l] > [ɬ].

CONCLUSIONS

Naturalistic elicitation of /r/ and /l/ provided evidence of parameter shifting in the English interlanguage of native speakers of Japanese. Sufficient numbers of non-target-like allophones were produced to indicate that in approximating /r/, these subjects increase rhoticization gradually, and in approximating /l/, lateralization is increased. Syllable and word-final consonant clusters consisting of obstruent plus liquid were underrepresented for the phoneme /l/. A successful analysis of /r/ and /l/ in these environments, and pre-vocalic ones, would require a larger database of sounds elicited naturalistically, as many as 5,000. Based on the current study, a total of 80 sets of data would be sufficient.

The elicitation demonstrates that /r/ is produced with greater frequency than /l/. Use of /l/ increased dramatically from pre- to post-test, providing evidence for avoidance of /l/. Another sign that /l/ is the more difficult of the two phonemes is the fact that /l/ is pronounced with far less accuracy than /r/, both pre- and post-test.

The analysis of phonetic environments of /r/ and /l/ errors revealed that the difference in accuracy between the two phonemes was greater pre-vocalically than post-vocalically. Switching of [ɹ] and [l] was observed only pre-vocalically. Post-vocalically, inaccurate pronunciation of both phonemes bore more of a resemblance to the target allophones.

Further study should focus on quantifying acoustic parameters of [ɹ] and less rhotacized variants. Similarly, variation between [l] > [l] > [ɬ] should be examined acoustically in order to document incremental shifts away from the rhotacized lateral to the velarized variety of American English.

REFERENCES

[1] Vance, Timothy J. (1987), *An Introduction to Japanese Phonology*, Albany NY: State University of New York Press.

[2] Dickerson, L. and Dickerson, W. (1977), "Interlanguage phonology: Current research and future directions", in S. Corder and E. Roulet (eds.), *The Notions of simplification, interlanguages and pidgins and their relation to second language learning* (Actes du 5ème Colloque de Linguistique Appliquée de Neufchatel, pp. 18-29), Paris: AIMAV/Didier.

[3] Labov, William (1972), *Socio-linguistic Patterns*, Philadelphia PA: University of Pennsylvania Press.

[4] Pullum, Geoffrey K. and William A. Ladusaw (1986), *Phonetic Symbol Guide*, Chicago IL: The University of Chicago Press.

TEMPORAL RELATIONSHIPS IN NORWEGIAN AS A SECOND LANGUAGE SPOKEN BY ADULT VIETNAMESE

Olaf Husby,

Department of Linguistics, University of Trondheim, Norway

ABSTRACT

This study examines the ability of Vietnamese speakers to reproduce stress patterns in Norwegian words by using reiterated speech. Norwegians produce stress patterns where a syllable's duration is governed by a hierarchy related to stress degree and sequential position. Both long- and short-term Vietnamese residents in Norway fail to produce this pattern. Longer stay implies mastery of higher levels in the hierarchy.

INTRODUCTION

Research has clearly demonstrated the influence of native language phonetic patterns upon second language production. However, there has been relatively little research on the influence of first language rhythmic patterns upon second language rhythmic patterns.

Rhythmic patterns in both first and second language speech may be heavily influenced by differences in phonotactic structure related both to phoneme distribution and syllable structure. This problem is more or less avoided by using reiterant speech as one can use syllable structures and speech sounds which the two languages have in common.

In studies based on this method, subjects are asked to substitute a single syllable for each of the original syllables in a word or in a sentence. At the same time the relationship between stressed and unstressed syllables in the original word should be maintained. The syllables used in reiterant speech usually have a CV-structure, and syllables like /ba/ and /ma/ are quite common. Acoustic and perceptual analysis of reiterant speech has shown that the prosodic characteristics of the original utterance is maintained [1], [2]. Most languages contain a bilabial consonant and an open unrounded vowel, consequently syllables like /ba/ or /ma/ will be suitable for crosslinguistic investigations based upon reiteration.

As Vietnamese has a maximum syllable structure of C(w)V(C) and Norwegian has (CCC)V(CCCC), the method of reiterated speech seems applicable. The investigation was performed using the syllable /ba/.

EXPERIMENTAL PROCEDURE

Recordings

Recordings were made in the studio of Linguistic Department of Trondheim University and at Rosenhof school, Oslo. Three groups of speakers participated in this investigation, all of them male full-time students aged 20-27: 5 Norwegians studying Nordic languages at University of Trondheim (NOR), 5 Vietnamese with more than ten years stay in Norway and who had completed seconday school and high school i Norway, now students at the technical university (VIETLONG), and 5 Vietnamese students who had completed about 75% of an 500 hours introductory program in Norwegian as a Second Language at Rosenhoff (VIETSHORT). In order to maintain a consistent dialectal background, it was required that all informants had been living the major part of their stay in Norway in the south eastern part (i.e. vicinity of Oslo).

A list of 24 words were used to elicitate data of the informants reiterated speech. The words were parts of small dialogs of questions and answers:

"What is that?"	"That's (an) X"
"What did you say?	"I said X"

There were four sets of common Norwegians words with 1-4 syllables to be reiterated. In polysyllabic words all stress positions were exploited, and for all words the stressed syllable contained long and short vowel. For words with final stressed syllable, both words with open and closed final syllable were reiterated. For open syllables only long stressed were used as Norwegian does not allow short stressed vowels in this posistion. Letting <x> represent unstressed syllables and <X> stressed syllables, the following stress patterns were exploited:

Stress patterns	Words	Reitreation example
X	'li	/ 'ba:/
Xx	'loven	/ 'ba:ba/
xX	ba'nan	/ ba'ba:/
Xxx	'korene	/ 'ba:baba/
xXx	fo'rening	/ ba'ba:ba/
xxX	para'ply	/ baba'ba:/
Xxxx	'bygningene	/ 'ba:bababa/
xXxx	ro'manene	/ ba'ba:baba/
xxXx	foto'grafen	/ baba'ba:ba/
xxxX	revolu'sjon	/ bababa'ba:/

Figure 1. Reiteration patterns

Analysis

The reiterated speech patterns were analysed using *Signalyze*™ ver. 2.25 on a Macintosh Quadra 700. Durations of phrase, key word, stressed syllable, syllable onset and nucleus were registered, as well intensity level and F1, F2

ANALYSIS OF ACOUSTIC CUES

Intensity level differences

An analysis of differences in intensity between the word consisting syllables, shows that of 105 cases of comparable intensity level differences (35 per informant group), only 11 were greater than 2.5dB, none were greater than 3.7dB. 8 of the cases were related to differences between stressed and unstressed syllables for VIETSHORT.

The Just Noticable Difference in intensity level under favorable listening conditions is 1 dB [3]. Here the differences is related to natural speech conditions, and based on the findings one concludes that intensity level differences is objectively not a cue to stress placement for any group though VIETSHORT is making an attempt to exploit this feature.

Spectrum

An analysis of spectrum shows that each group reiterate key words with significantly different vowel qualitites. While NOR is using a back half open to open vowel [ɑ], VIETSHORT is using a more front open [a]. VIETLONG is intermediate [ɑ̟]. This is regarded as an approach to Norwegian vowel qualitites.

NOR demonstrates different formant values for initial stressed and unstressed syllables, but no differences is found for non-initial syllables. This pattern is not found for the Vietnamese informants where formant values are stable irrespective of stress position and degree.

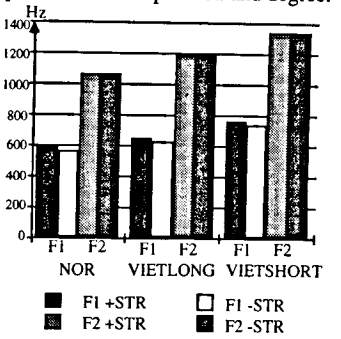


Figure 2. Formant frequencies stressed and unstressed syllables

F0 Pattern

There are significant different F0-patterns in stressed syllables. In accordance with the south eastern dialect, NOR and VIETLONG produces a *low* tone on the stressed syllable for all reiterated words. VIETSHORT, however, is using a *high* tone (fig. 3).

The reason for this is not clear. In the author's northern dialect a high tone is used in stressed syllables, so VIETSHORT's pattern might have been picked up as instructions were given. It is not likely that the short exposure to the northern dialect should overrule patterns acquired during the language course.

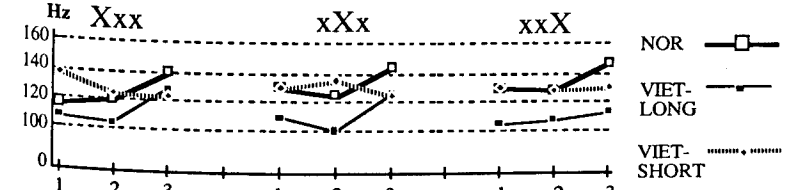


Figure 3. F0 curves trisyllabic words.

Duration

There are significant differences in durational patterns in the reiteration performed by the three groups. All groups produce identical duration of nucleus in monosyllabic word (215 ms). The stressed syllable of VIET-SHORT are however significantly shorter (p<.001) due to the short implosive bilabial [ɓ].

There are significant differences between NOR and both Vietnamese groups with respect to both duration of *phrase* (p=.002) and *anacrusis* (p=.001). There are no differences between the Vietnamese groups. The anacrusis constitutes 49,2% of NOR's phrase and 66,1% and 73,1% of VIETLONG resp. VIETSHORT. This pattern is found for all test words.

For polysyllabic words the investigation shows that both Vietnamese groups to some extent is able to reproduce Norwegian reiteration patterns. The reproduction ability is significantly related to duration of stay in Norway. The pattern is consistenly expressed for all polysyllabic words, but will demonstrated here with reference to four-syllabic words only (fig. 4-7).

For NOR the following pattern is found: stressed syllable is lenghtened, final syllable is lenghtened, if this two features co-occur, a cumulative effect is found. Unstressed syllables following the stressed syllable is lengthened if their position is odd-numbered (stressed syllable given number 1). This effect is labelled *rhythmical lengthening* and reflects the claims of metrical phonology where alternations of strong and weak syllables is described [4]. This effect is also cumulative in relation to final lengthening.

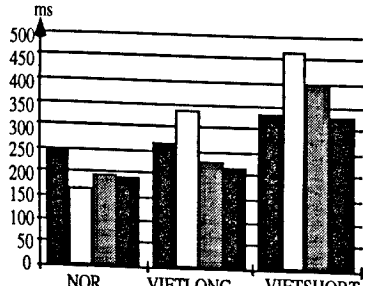


Figure 4 . Reiteration pattern Xxxx-words

(In figures 4-7 duration of syllables 1-4 is shown as columns from left to right).

For 4-syllabic words stressed on the 1. syllable NOR is using stress lengthening (1. syllable), rhythmical lengthening (3. syllable) and final lengthening (4. syllable). The unstressed 2. syllable is short. Both Vietnamese groups fail to produce this pattern.

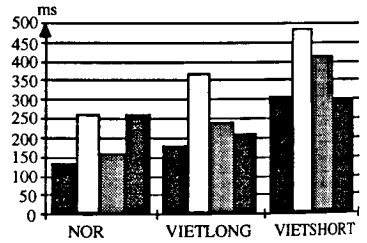


Figure 5 . Reiteration pattern xXxx-words

For four-syllabic words stressed on the 2. syllable NOR produces stress lengthening for the 2. syllable, and the 4. syllable undergoes both rhythmical and final lenghtening. The cumulative effect is shown as the duration of the final syllable is equal to the stressed syllable. The final syllable is also longer than in Xxxx-words. Both Vietnamese groups lengthen the stressed syllable only.

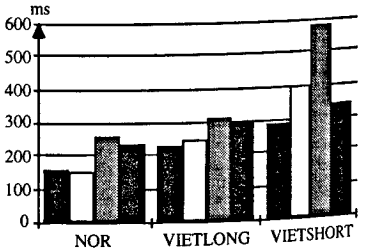


Figure 6. Reiteration pattern xxXx-words

For words stressed on the 3. syllable, NOR lengthens the stressed syllable, the 4. is lenghtened due to final position. VIETLONG lengthens the stressed and final syllable, VIETSHORT only the stressed.

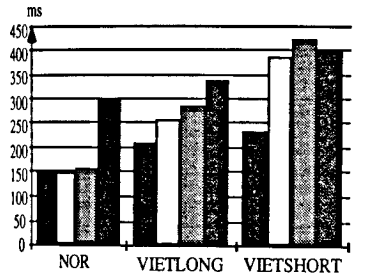


Figure 7. Reiteration pattern xxxX-words

For words stressed on the last syllable NOR lengthens this syllable which also is final. The same is found for the Vietnamese groups, who however demonstrates the lacking ability to shorten anacrustic syllables (se also fig. 5, 6).

An overall perspective (1-4 syllabic words) demonstrates the Vietnamese informants varying competence in using the lenghtening principles. In monosyllabic words, all three groups produce identical durations nuclei in stressed syllables. In polysyllabic words, stress lengthening is only found for NOR and VIETLONG. Lengthening of final syllable is used by the same groups, rhythmical lengthening by NOR exclusively as shown in fig. 5.

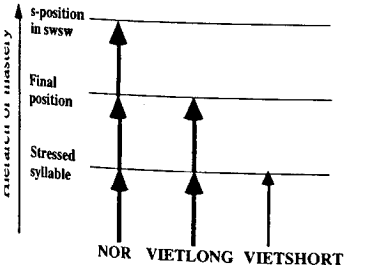


Figure 8. Lenghtening principles as exploited by the three informant groups. The thin line for VIETSHORT illustrates a partial ability to lenghtne stressed syllables.

CONCLUSION

This investigation has shown that in reiterating Norwegian words, Norwegian speakers apply three principles of lengthening both stressed and unstressed syllables to signal stress. The lengthening effects is working cumulative. As a secondary effect vowel quality is used to express differences between stressed initial on non-initial syllables. In accordance with the Norwegian dialect used, the stressed syllable is pronounced with a low tone. Intensity is not used as a cue to signal stress differences.

Vietnamese learners of Norwegian as a Second Language apply different strategies according to length of stay in Norway. Long term residents has partial access to the hierarchical lengthening system as stress and final lengthening is used, also in a cumulative way. Vowel quality, and intensity is not used used as cues of stress. The low tone is used on stressed syllables.

Short term Vietnamese residents has limited access to the hierachical lengthening system. Spectrum is not used as a cue to stress. Where the other informants use low tone on the stressed syllable, this group is signalling stress by the use of a high tone. There is also an attempt of using intensity differences as a signal of stressed syllables.

REFERENCES

[1] Larkey, L. (1983). Reiterant speech: An acoustic and perceptual validation. *Journal of the Acoustical Society of America, 73,* 1337-45

[2] Liberman, M.Y. & Streeter, L.A. (1978). Use of nonsense-syllable mimicry in the study of prosodic phenomena. *JASA 63,* 231-33

[3] Flanagan, J.L. (1957). Estimates of the maximum precision necessary in quantizing certain "dimensions" of vowel sounds. *JASA, 24,* 533-4

[4] Hogg, R. & McCully, B.E. (1987). *Metrical Phonology: a Coursebook.* Cambridge University Press

PERCEPTION OF CONTINUOUS SPEECH IN SECOND LANGUAGE ACQUISITION

S. Komar
Department of English and German Studies, Ljubljana, Slovenia

ABSTRACT

The purpose of the study is to show the influence of phonological and phonetic features on perception and comprehension of continuous speech in listeners who are not native speakers of English, but who have been taught the standard British English for at least eight years. Results of the testing show one common feature: several successive exposures to the same acoustic material do not improve the perception and comprehension but they confirm the initial impression.

SPEECH PERCEPTION IN THE MOTHER TONGUE

The study of speech perception is faced with two major problems: the first is to determine the phases involved in recognizing the spoken word; the second is to identify the nature of different types of context that influence these phases.

Phases of spoken word recognition

The process of recognizing a spoken word begins with the 'initial contact phase' when a listener takes the speech wave as input from which he abstracts the representation which contacts the internally stored form-based representations associated with each lexical entry [1]. These are not yet linguistic units, i.e. phonemes and syllables, but they function as mediators between the acoustic wave and abstract linguistic units. They are the so-called 'acoustic cues'.

The lexical entries that match the contact representation to some degree are then activated - the 'activation phase'. Theories differ in determining the status of activated words: some believe that all lexical items are equally activated, while others claim that their status depends on their frequency of occurrence in the language.

After initial contact and activation of the lexicon, the sensory input continues to be mapped onto the chosen lexicon unit until the intended lexical entry is selected - the 'selection phase'.

The end-point of the search phase is 'word-recognition', i.e. the moment when a listener has already determined which lexical entry he has heard. The precise moment when this takes place is said to be before a listener has completely heard the word.

Context effects

Context is said to play in important role in spoken word recognition. Psycholinguists generally agree that lexical processing depends on two broad classes of information: on the one hand, there are representations computed from the sensory input, on the other, there are representations constructed from the previous context using higher sources of knowledge (e.g. lexical, syntactic, semantic, pragmatic).

A very broad distinction between two types of context is the distinction between structural and non-structural context.

Structural context are the constrainsts according to which elements can be combined into higher-level linguistic units and can be applied to the phoneme, morpheme, phrase, utterance and discourse levels.

Non-structural context does not influence the building up of higher-level representations, but rather it influences the recognition of one word in relation to another.

There are different views regarding the exact moment when different contexts influence the recognition of words and phrases. In the so-called interactive view contexts may intervene throughout lexical processing, thus altering the choices that have already been made.

Perception of continuous speech

The basic difference between the study of isolated speech sounds and the study of continuous speech is that the former investigation is carried out in more or less artificial circumstances, whereas the study of continuous speech normally observes conversations with meaning and substance, where people listen for the message, not isolated sounds.

There are two views regarding the perception of continuous speech in mother tongue: the passive one, according to which listeners try to identify each word as if it were an isolated word, and whenever they fail, they guess. According to the active view, listeners use linguistic constraints in the perception process. It looks as if they listened for some words or phrases and ignored the others. A reason why this view is a more appealing one is that normal speech is in itself unintelligible and it is only an illusion that we think it is clear [2]. Another evidence in favour of the view that the perception of continuous speech is not merely putting together of isolated sounds is what Clark & Clark wittily call 'cocktail party phenomenon'. The idea is that although we are surrounded by several speakers, speaking about different things and to different people, our ears are able to pick up from the excess of linguistic and above acoustic information only those items which belong to the speech of the person we are talking to.

SPEECH PERCEPTION IN NON-NATIVE SPEAKERS OF ENGLISH

The question is whether non-native speakers of a certain language perceive the spoken word in the same way as the native speakers. At this point we can only imagine some points in speech perception where problems are likely to arise.

The four phases in the recognition proces - taking in of the acoustic material, attempting to organize phonological representations into constituents, building constituents into higher representations, and identifying all propositions - are not applied one after another, but are all in action at the same time and they work well with the native speakers.

With non-native speakers the first problem arises when they cannot succeed in identifying the speech sounds correctly. There are several reasons for that among which the most important one is the lack of expectation about what words are likely to sound in the foreign language. This may particularly be the problem of those non-native speakers who are used to a strong relationship between sounds and spellings in their mother tongue, and are exposed to a language where this is not the case (e.g. Slovenes learning English). Even if the initial phase is successfully fulfilled a non-native speaker may not recognize the speech sequence correctly because of the grammatical nature of the sequence or poor vocabulary.

What can be expected from students of English with advanced knowledge of English, when they are exposed to spoken English, and particularly when they are exposed to spoken non-standard variants of English? For that purpose an experiment with native speakers of Slovene, all first year students of English at the Department of English and German Studes at the Ljubljana University, was carried out.

THE EXPERIMENT

The purpose of the experiment was to examine the perception and understanding of spoken RP and two non-standard British accents: popular London (Cockney) and Scouse with regard to phonological and phonetic diversities of the three variants. Another purpose was to find out how and to what extent different contexts help informants to better understanding.

The informants were tested in listening comprehension and in dictation. The dictation test was to show whether different context effects are in play throughout the perception process, thus correcting wrong selections simultaneously.

The informants were divided into two groups of fifty students of which one group was tested in listening comprehension, while the other did the dictation test of the same recording. They were tested in all three varieties of British English. Results of both tests were first analysed individually and then compared.

ANALYSIS AND EVALUATION OF RESULTS

Listening comprehension

In listening comprehension tests students listened to the same recording twice: after the first round they were given a few very general questions, while after the second round the questions were more detailed. The results, regardless of the accent, showed one common feature: after the second listening the comprehension of the recording did not improve; with some students the number of correct answers after the second listening even fell. This brought us to the conclusion that several successive exposures to the same acoustic material do not influence the perception and understanding in the sense of improving it. Instead, they confirm the initial

impression, rather than correct or improve it.

The influence of the phonemic and phonetic differences of Cockney and Scouse as against RP were not essential for success in listening comprehension. They must have been overcome by the help of the semantic context.

Dictation

Types of mistakes that appeared in dictation can be roughly divided into two groups: (a) misperceptions on the basis of phonemic resemblance, i.e. correct words or phrases replaced by words or phrases that acoustically resembled the correct words, but were completely inappropriate in the given context; (b) omissions of words, phrases or whole sentences. Most interesting were mistakes under (a) which were either single word errors or multiple word errors. The latter we believe resulted from the correct perception of the stressed syllable on the one hand, and incorrect perception of word boundaries on the other. It looks as if the phonemic and phonetic variations affecting the consonant clusters at word boundaries were responsible for these misperceptions.
Examples:
RP: trouble getting *their breaths*, became, trouble getting *their best*, or, the *Upper Circle*, became, *up in circle*.
Cockney: they've all been *sentenced to the guillotine*, became, they've all been *sent guilty*.
Scouse: contamination *source*, became, contamination *sauce*.

Very many multiple word errors resulted in homphones which did not fit the grammar or the context. This was particularly frequent in Cockney and Scouse, where a sequence containing one or two "unfamiliar" phonemes was not identified correctly.
Examples.
Cockney: the second one *today*, became, the second one *to die*, or, *they* became *I*.
Scouse: *My old fellow used to fish in that river* became *My offer was officially that river*, or, *I've no fancy letters after me name* became *I've no funny lessons*, or, *intermittent dumping* became *intimate jumping*.

Results like these led us to the conclusion that in the case of dictation, the perception process in non-native speakers of English resembles more the perception of isolated words than the perception of continuous speech where the role of semantic context is decisive. We believe that the non-native speakers when they are asked to note everything they hear, listen for sounds and words rather than meaningful sequences. Their efforts to catch every single word very often result in meaningless (sometimes completely ungrammatical or nonsense) word sequences. This proves that the influence of any context is absent.

Listening comprehension vs. dictation

The comparison of the results in listening comprehension and dictation showed that the relationship between the errors in dictation and incorrect answers in listening comprehension, except in one or two cases, did not match. In other words, misperceptions of certain phrases or even sentences in dictation did not have their counterparts in false answers in listening comprehension. Moreover, the students answered some questions which were closely related to the misperceived sequences in dictation correctly. This confirmed our above-mentioned presumption that in listening comprehension, where the students were primarily interested in getting the message, the role of context is important. In dictation, where the task is to put down every single word, the students were mainly in search for words and did not pay any attention to any kind of context.

CONCLUSION

With respect to different types of context that influence the perception of continuous speech, as well as how and when they influence it, we may conclude that:
(i) in listening comprehension lexical and semantic contexts can help listeners to catch the meaning of a speech sequence they are exposed to, so that obstacles which "unfamiliar" sounds and words present can often be easily overcome;
(ii) in dictation the interest of the listeners is mainly to catch individual words, so they focus their attention to individual sounds and words. The influence of lexical and semantic contexts is minimal;

With regard to other factors which influence the perception of speech in non-native speakers of English, we conclode that:
(i) the phonological differences between the Slovene phonological system and the phonological systems of the three British accents do not interfere. The results neither indicate any cases where the Slovene phonological system would help in perception of RP, Cockney or Scouse. Although we cannot deny the influence of the mother tongue upon the perception of a foreign language, we, nevertheless, believe that at an advanced level of knowledge the influence of the phonological system of the mother tongue upon the perception of the foreign language is minimal;
(ii) phonemic and phonetic differences among the three British accents can be responsible for some errors in dictation, mainly in the two non-stabdard accents;
(iii) the general knowledge that people have about the culture, civilization, politics, geography, etc. of the country the language of which they learn, is also important and is often neglected in the evaluation of foreign learners' perception and production skills;
(iv) the experiment proved our initial assumption that speech perception in native speakers differs from that in non-native speakers. In the former case it is a more or less subconscious process, whereas in the latter a considerable amount of effort and work is required from listeners.

REFERENCES

[1] Frauenfelder, U. H. & L. Komisarjevski Tyler (1987), *Spoken Word Recognition, A Cognition - Special Issue*, The MIT Press.
[2] Clark, H. H. & E. V. Clark (1977), *Psychology and Language*, Harcourt Brace Jovanovich.

PRODUCTION OF SCHWA BY JAPANESE SPEAKERS OF ENGLISH: A CROSS-LINGUISTIC STUDY OF COARTICULATORY STRATEGIES

Yuko Kondo
University of Edinburgh

ABSTRACT

The present study addresses the question of how L2 learners acquire the coarticulatory strategies of L2. English and Japanese manifest interesting contrasts in their coarticulatory patterns. The present study looks into the shift from the L1 to L2 coarticulatory pattern by observing the production of schwa by Japanese speakers of English.

INTRODUCTION

There are a number of interesting contrasts in the coarticulatory strategies of English and Japanese. Firstly, the reduced vowel, schwa, of English seems to be phonetically unspecified in F2 [2][3]. That is, the rhythm of English traditionally characterized as the alternation of full and reduced vowels may be described as the contrast of targeted and targetless vowels. In other words, different degrees of contextual assimilation are observed between full and reduced vowels of English. In Japanese, on the other hand, there is no unspecified vowel such as schwa and presumably all its vowels are targeted. Secondly, acoustic studies of V-to-V coarticulation have shown that English has stronger carryover than anticipatory effects in F2 [1] [5]. On the other hand, stronger anticipatory effects have been observed in Japanese [3] [4].

If Japanese speakers of English successfully shift their coarticulatory pattern from the L1 (Japanese) to the L2 (English) system, they would manifest a contrast in context dependent variability between schwa and full vowels. They would also show a shift in the relative strength from the R-to-L to L-to-R coarticulatory effects. As coarticulatory strategies are closely related to prosodic or organizational aspects of languages, a crosslinguistic study may yield an interesting insight into how speech is organized in these languages.

METHODS

Vb_bV sequences with the English schwa, the full vowel /æ/ and the Japanese vowel /a/ as the middle vowel were embedded in natural sentences. The contextual vowels were /ɪ/ and /æ/ or /ə/ for the English sequences and /i/ and /a/ for the Japanese sequences resulting in 12 different sequences in all as follows.

- *The campaign for Women's Lib abysmally failed.*
- *The inspector considered the lab abysmal.*
- *We found the crib abandoned in the car park.*
- *The crab abandoned its prey as it sensed something approaching.*
- *The fib Abbey National 's TV advert was said to contain turned out to be quite legal.*
- *When today's students were in the crib Abba were superstars.*
- *The robbers planned to grab Abbey National's armoured van.*
- *Nostalgia fans like to grab Abba records when they see them.*
- *Mukashi Babironia-to-iu kuni-ga arima'shita.*
- *So'to-dewa o'tsiba-bakari kasakasa-to oto'-o ta'te-te-imasu.*

Table 1. The results of ANOVA's for the English vowels /ə/ and /æ/ produced by British English and Japanese speakers and the Japanese vowel /a/ produced by Japanese speakers. The symbol + shows that the main effect of the preceding or the following vowel was significant by $p < 0.01$. The symbol – means that no significant effect was obtained.

		preceding vowel			following vowel		
speaker	vowel	onset	midpoint	offset	onset	midpoint	offset
English	ə	+	+	+	+	+	+
	æ	+	+	–	–	–	+
Japanese	ə	–	–	–	+	+	+
	æ	–	–	–	+	+	+
	a	+	+	+	+	+	+

- *Sei'sho-niwa Babironia-no-koto'-ga iroiro ka'ite-arimasu.*
- *Kawa'-niwa ka'ba-bakari-de-na'ku kiken-na wa'ni-mo imasu.*

Eight male British English speakers and five Japanese male speakers participated in the experiment. The native English speakers produced the 8 sequences with the English vowels while the Japanese speakers produced all the 12 sequences. Each sentence was repeated 5 times in a randomized order. In order to make the number of observations between native and non-native speakers more or less equal, the middle three repetitions were taken from each native speaker's production while all the 5 repetitions were used for the non-native speakers' production resulting in 24 observations per sentence type for native and 25 observations for non-native speakers respectively. The sentences were sampled at 16 kHz into a UNIX SUN workstation with WAVES speech analysis facilities. Formant values were obtained by running the FORMANT program for LPC analysis with a 25 ms cos**4 window moving in 5 ms steps. The measurements were taken at the onset, midpoint and offset of the vowel. Only the second formant values were studied as schwa was observed to be targetless in F2 [2][3].

RESULTS

Three-way ANOVAs were performed with the onset, midpoint and offset of the English vowels /ə/, /æ/ and the Japanese vowel /a/ produced by English and Japanese speakers as dependent variables. The independent variables are preceding vowel, following vowel and speaker. Table 1 shows the results for the main effects. Significant interaction was observed between the preceding vowel and speaker at the onset of /ə/ for the native speakers' production. For the Japanese speakers' production, significant interactions were observed between the following vowel and speaker at the three points of the Japanese /a/ and the English /ə/ and at the midpoint and offset of the English /æ/. Where significant interactions were observed, speakers showed different degrees of V-to-V effects. One Japanese subject had higher mean F2 for schwa in the context of /æ/ than in that of /ɪ/ at all the three points of the segment.

A number of interesting observations may be made from Table 1. First of all, the reduced vowel /ə/ is more transparent than the full vowel /æ/ when they are produced by native speakers. That is, the effects of both the preceding and following vowel are observed right through the schwa whereas the effect of the preceding vowel is stopped at the midpoint of /æ/ and the effect of the following vowel is stopped at the offset. Secondly, the Japanese vowel /a/ is as transparent as the native speakers'/ə/. Thirdly, the Japanese subjects did not show any difference in sensitivity to context between the production of /ə/ and /æ/. Lastly, when the Japanese subjects produced the non-native vowels /ə/ and /æ/, their preference for anticipatory effects seems to be much more pronounced. Somehow their coarticulatory pattern is distorted and

different from either the L1 or the L2 pattern. There seems to be a strong over-projection of the preferred coarticulatory pattern of the L1 onto the interlanguage system.

Figures 1 and 2 show the differences in the mean F2 values as a function of the preceding and following vowel at the onset, midpoint and offset of the three vowels /ə/, /æ/ and /a/. When Figures 1 and 2 are compared, the native speakers' schwa shows stronger carryover than anticipatory V-to-V effects. Significant differences are observed right through the schwa for both carryover and anticipatory effects. For the full vowel /æ/, the differences are smaller for the carryover effects and the effects diminish around the midpoint of the segment. For anticipatory effects, the difference is observed only at the offset of the vowel.

For the Japanese vowel /a/, anticipatory effects are greater in magnitude compared to carryover effects. Significant differences are observed right through the segment for both carryover and anticipatory effects. The extent of V-to-V carryover effects observed on the Japanese /a/ is intermediate in degree between that observed for the English speakers' /ə/ and /æ/. However, for anticipatory effects, the Japanese vowel /a/ shows the greatest effects among the three. The Japanese speakers showed hardly any carryover effects for the English

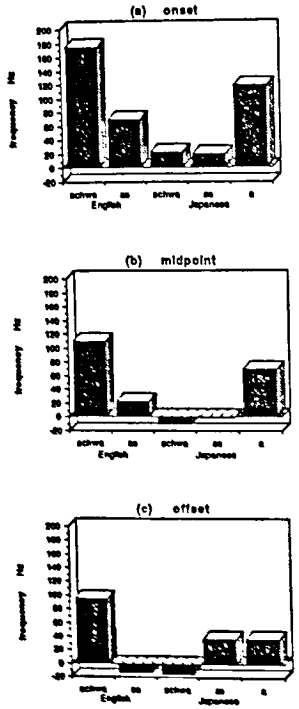


Figure 1. The differences in F2 values as a function of the preceding vowels /l/ or /i/ and /æ/ or /a/ for the English vowels /ə/ and /æ/ produced by English and Japanese speakers and the Japanese vowel /a/ produced by Japanese speakers.

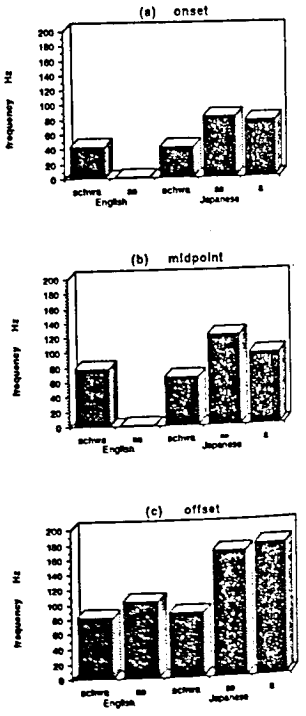


Figure 2. The differences in F2 values as a function of the following vowels /l/ or /i/ and /æ/, /ə/ or /a/ for the English vowels /ə/ and /æ/ produced by English and Japanese speakers and the Japanese vowel /a/ produced by Japanese speakers.

/ə/ and /æ/. Also, unlike the native speakers' pattern, they showed stronger effects for the full vowel /æ/ than for schwa in anticipatory V-to-V coarticulation. These results seem to suggest that the Japanese speakers in the present study are not successfully shifting from the L1 to L2 coarticulatory strategies in speaking English. The general pattern seems to be that of a transfer, or more precisely an over-projection of their native coarticulatory pattern onto the interlanguage.

DISCUSSION

The context dependent variability of schwa in the indefinite article *a* is reported in [3]. In this study, VCəCV sequences with the consonantal contexts of /p, t, k/ and the vocalic contexts of /ɪ, æ, u/ are embedded in natural sentences such as *You may pick a kitten from the basket.*. Two groups of non-native (Japanese) speakers of English produced these sequences. These groups may be described as fluent and non-fluent groups. In producing schwa, non-fluent speakers showed a coarticulatory pattern which may be described as a transfer from the Japanese vowel /a/ in F2 and non-systematic variability in F1. On the other hand, fluent speakers of English exhibited large and systematic variability in F2 which is very similar to the native speakers' pattern. These speakers seem to have acquired the phonetic underspecification of schwa in F2. Furthermore, two of the three fluent speakers of English showed stronger V-to-V carryover effects in the labial context where most native speakers of Japanese would show stronger anticipatory effects in Japanese. The results of the above study seem to contradict with the results of the present study. Different subjects used in the two studies may explain this contradiction. However, the subjects participated in the present study were relatively fluent speakers of English as well. Another possible explanation is the relative difficulty of the sentences used in the two experiments. The content words used in the present study, such as *abysmal*, *fib* and *crib* were not familiar to the Japanese subjects. This may have affected their performance. The L2-like coarticulatory pattern reported in [3] may not immediately affect all the lexical items in the interlanguage. The L2-like coarticulatory pattern may gradually spread from more familiar words to less familiar words.

CONCLUSION

Contrary to the prediction, Japanese speakers of English showed less systematic contextual variability on /ə/ than on the full vowel /æ/. The pattern of V-to-V coarticulation across /b/ observed for the Japanese speakers' /ə/ and /æ/ may be characterized as a transfer or more adequately as an over-projection of the coarticulatory pattern of L1 onto the interlanguage. Their coarticulatory pattern of /ə/ was more deviant from the L1 pattern than that of /æ/. However, instead of approaching the L2 pattern they seem to have shifted away from the L2 as well. On the other hand, the coarticulatory pattern of /æ/ was more L1-like. This seems to suggest that the Japanese subjects in the present study had some awareness of schwa being a unique and different vowel, but somehow failed to produce the correct coarticulatory pattern.

REFERENCES

[1] Bell-Berti, F. & Harris, K.S. (1976) "Some aspects of coarticulation." *Haskins Lab. Status Report on Speech Research*, SR45/46, pp. 197-204.

[2] Kondo, Y. (1994) "Phonetic underspecification of schwa." *Proceedings of the International Conference on Spoken Language Processing*, Yokohama. Vol. 1, pp. 311-314..

[3] Kondo, Y. (1995) "Production of schwa by Japanese speakers of English: A crosslinguistic study of coarticulatory strategies." *Ph.D. dissertation*, University of Edinburgh.

[4] Magen, H. (1984) "Vowel-to-vowel coarticulation in English and Japanese." *JASA* 75, Suppl. 1:S11.

[5] Öhman, W.E.G. (1966) "Coarticulation in VCV utterances: Spectrographic measurements." *JASA* 39, pp. 151-168.

THE EFFECTS OF TALKER VARIABILITY ON THE ACQUISITION OF NON-NATIVE SPEECH CONTRASTS

James S. Magnuson and Reiko A. Yamada
ATR Human Information Processing Laboratories, 2-2 Hikaridai, Seika, Soraku, Kyoto, 619-02, Japan; email: magnuson@hip.atr.co.jp, yamada@hip.atr.co.jp

ABSTRACT

In previous /r/-/l/ training studies that stressed talker variability, stimuli were blocked by talker. We compared mixed-talker training (in which the talker changed from trial to trial) with blocked-talker training, to see if blocked training gives subjects adequate experience to adapt to talker differences. We found that mixed training led to better talker-adaptation, although the effect was mitigated when the number of talkers used in training was increased.

INTRODUCTION

Previous /r/-/l/ training paradigms for Japanese listeners have shown that talker variability in training is a crucial element for post-test generalization to new talkers and stimuli. Subjects trained with stimuli produced by only one talker sometimes fail to acquire generalization ability [1], whereas subjects trained with stimuli produced by five talkers consistently show good post-test generalization (e.g., [2]). When high talker variability has been stressed, stimuli have been blocked by talker. Results from perceptual studies suggest that how such training is structured may be an important consideration.

Yamada and Tohkura [3] reported that when untrained Japanese adults attempt to identify English /r/ and /l/, they appear to set criteria based on the range of cues they hear in a series of trials. Given the entire series of stimuli from a synthesized /r/-/l/ continuum, Japanese response functions were quite similar to English speakers'. However, given only the most /r/-like or /l/-like half, Japanese boundaries shifted such that they continued to respond "R" approximately 50% of the time (as opposed to native speakers, whose nearly-categorical functions were not affected by such changes). It seems that Japanese subjects expected to hear equal numbers of /r/ and /l/ tokens, and set response criteria accordingly when the range of stimuli changed.

We have found a similar *range-bias* effect due to talker variability [4]. When Japanese adults identified /r/ and /l/ stimuli produced by only one talker, they responded "R" approximately 50% of the time. When stimuli from five talkers were mixed, the overall rate of "R"-response (*R-rate*) was still about 50%. However, the *R-rate* to particular talkers differed significantly between blocked-(single) and mixed-talker conditions. It seems that subjects set a single criterion based on the range of cues they heard within a block, rather than evaluating each stimulus independently.

This strategy was successful when the stimuli were produced by one talker. However, when stimuli from different talkers were mixed, some talkers' /r/s and /l/s sounded "R-like" or "L-like" relative to other talkers' productions, as was reflected in significant, talker-specific changes in bias (*R-rate*).

Given that previous studies which stressed the importance of talker variability in /r/-/l/ training have only presented stimuli blocked by talker [2], it follows that trained subjects might have difficulty adapting to between-talker differences, as did subjects in our perceptual tests [4]. The experiments we report here were designed to examine this possibility.

EXPERIMENT 1

We trained two groups of subjects to identify English /r/ and /l/. One group of subjects was trained in a mixed-talker condition (i.e., the talker could change on any trial), and the other was trained in a blocked-talker condition (i.e., the talker remained constant within a block). This comparison was made to determine whether either condition better promotes the ability to adapt to talker differences in non-native speech contrast perception.

Method

Subjects. 12 native speakers of Japanese with limited English training were paid to participate in Experiment 1.

Stimuli. The training stimuli were 79 minimal pairs of real English words contrasting /r/ and /l/ in 5 phonetic contexts: initial singleton, initial cluster, inter-vocalic, final singleton and final cluster. The stimuli were produced by 2 male native speakers of American English. One of the talkers was found to be "R-like" relative to several others (i.e., *R-rate* to his productions was ~.50 in a blocked-talker condition, but in a mixed-talker condition, increased significantly), and the other was found to be "L-like" [4]. The stimuli were selected from the set used in [2].

The test stimuli were 25 minimal pairs of real English words contrasting /r/ and /l/ in initial position (selected from the set used in [2]). The stimuli were produced by 4 native speakers of American English: the 2 training talkers and 2 female talkers (one R-like and one L-like [4]).

A set of generalization stimuli produced by a male talker not used in training or testing consisted of 32 minimal pairs of real English words contrasting /r/ and /l/ in the five training contexts, and 8 pairs of filler items contrasting other phonemes. These items were used in previous training studies [1,2] and were included to facilitate comparisons with those studies.

Procedure. A 2-alternative forced-choice paradigm was used for all sessions. On each trial, orthographic forms of a minimal pair of words were displayed on a CRT. Then, one of the pair was presented over headphones. The subject responded by pressing a key indicating the side of the screen on which he or she thought the word played over the headphones was displayed. The orthographic forms were randomly assigned to the right or left side of the CRT. Subjects received feedback about their responses only in training trials. If the subject answered correctly, a chime sounded and the next trial began. If the subject answered incorrectly, a buzzer sounded, the orthographic forms were randomly assigned to the left or right, and the word was played again. This continued until the subject answered correctly. For every three words identified correctly on the first attempt, subjects received an additional 1 yen as a monetary incentive.

There were 7 parts to the experiment: pretest, training sessions 1 and 2, mid-test, training sessions 3 and 4, and post-test. Subjects participated in the pretest and training session 1 on day 1, training session 2 on day 2, mid-test and training session 3 on day 3, training session 4 on day 4, and the post-test on day 5.

Test sessions consisted of nine blocks. Four were 50-trial blocks of stimuli produced by only one of each of the four testing talkers (blocked-talker). The same 200 stimuli were presented in four mixed-talker blocks. The order of blocked and mixed blocks was determined randomly. The ninth block consisted of the generalization items.

Subjects were randomly assigned to two training groups of six subjects. The blocked-training group heard training stimuli produced only by the "R-like" talker in the first 2 training sessions, and only stimuli produced by the "L-like" talker in the last 2 training sessions. Both groups were trained with four 948-trial sessions (total training trials: 3792). The mixed-training group heard equal numbers of stimuli produced by each talker presented in random order in each training session.

Results and Discussion

For this paper, we will focus on pretest--post-test comparisons; space constraints prevent us from including mid-test results. Both groups showed significant accuracy improvement on the /r/-/l/ items in the generalization set between pre- and post-tests ($F(1,10)>7$, $p<.05$; blocked group: .64 to .74; mixed: .62 to .72).

Both groups also improved on test items between pre- and post-tests, as can be seen in Table 1. An analysis of simple effects showed that the post-test difference between groups was significant for at the blocked level of talker condition ($F(1,10)=30$, $p<.001$). Note that subjects in the blocked group were significantly less accurate in the mixed-talker portion of the post-test than in the blocked-talker portion ($F(1,5)=19$, $p<.01$). That there was no such difference in the mixed group suggests that subjects in the mixed group applied

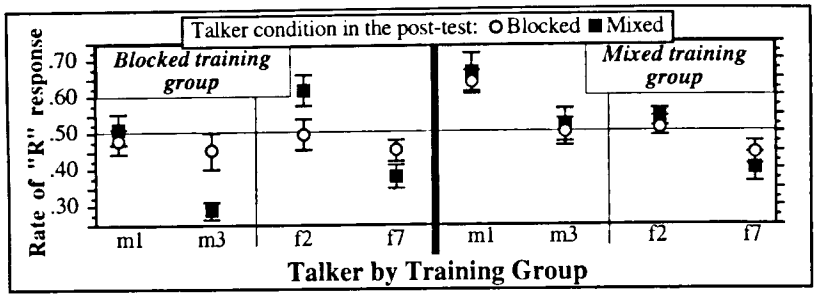


Figure 1: The talker by talker condition interaction by group in the Experiment 1 post-test. m1 and m3 were training talkers; f2 and f7 were heard only in tests. m1 and f2 appeared R-like to untrained subjects and m3 and f7 appeared L-like [4]. Bars show standard error.

similar response criteria in both talker conditions in the post-test.

ANOVAs on the *R-rate* data confirm that the mixed group used similar criteria in both talker conditions. Both groups showed interactions of talker and talker condition in the pretest ($p<.05$ for the mixed group, $p=.05$ for blocked). The blocked-talker training group still showed a strong post-test talker by talker condition interaction ($F(3,15)=16$, $p<.001$). In contrast, the interaction was not significant for the mixed training group ($F(3,15)<.5$, $p>.7$; see Figure 1). These subjects' responses to particular talkers were not affected by talker condition. Both groups of subjects responded "R" approximately 50% of the time in the blocked- and mixed-talker conditions. However, in the post-test, the blocked group showed large talker-specific differences in *R-rate*, similar to those observed previously with untrained subjects; *R-rate* increased for R-like talkers and decreased for L-like talkers.

Table 1: Accuracy in Experiment 1. All pre-post differences were significant (Tukey HSD post-hoc test, .05 level).

Group	*Talker condition*	*Pretest*	*Post-test*
Blocked training	blocked	.58	.73
	mixed	.59	.67
Mixed training	blocked	.58	.63
	mixed	.56	.63

It appears that the mixed training condition promoted greater ability to adapt to talker-specific differences in /r/ and /l/ productions, since subjects trained in that group responded "R" equally often to particular talkers -- even unfamiliar ones -- in both talker conditions in the post-test. However, the accuracy differences between groups suggest that there may be trade-offs between talker variability and stability in training: stability promoted higher accuracy for the blocked-training group.

EXPERIMENT 2

The results of Experiment 1 indicate that mixed-talker training may promote better adaptation to talker differences for even unfamiliar talkers. However, they do not provide an adequate basis for comparison with previous /r/-/l/ training paradigms that have stressed talker variability, as those studies have used 5 talkers in training [2]. In Experiment 2, we increased the number of training talkers to 5 and the number of testing talkers to 7, and added a third training condition.

Method

Subjects. 30 native speakers of Japanese with limited English training were paid to participate in Experiment 2.

Stimuli. We used the same sets of testing, training and generalization stimuli we used in Experiment 1. However, the test stimuli were produced by 7 talkers (the same 4 used in Experiment 1, with the addition of 2 males and 1 female). The training stimuli were produced by 5 talkers (the 2 males used in Experiment 1 with the addition of 1 male and 2 females).

Procedure. The same 2-alternative forced-choice paradigm used in Experiment 1 was used, although the

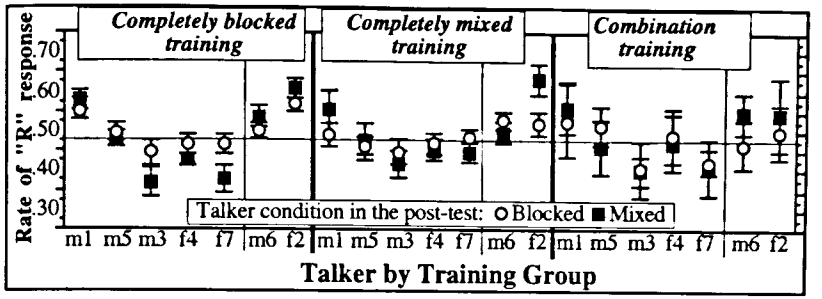


Figure 2. The talker by talker condition interaction by group in the post-test in Experiment 2. Talkers m1, m5, m3, f4 and f7 were training talkers; m6 and f2 were heard only in tests. m1, m5 and f2 have appeared R-like to untrained subjects, and m3, f4 and f7 were L-like [4] (m6 has not been tested). Bars represent standard error.

design was changed slightly. On the first day, subjects participated in a test with the generalization materials and a multiple talker pretest similar to the one used in Experiment 1, with the addition of 50 trials in blocked and mixed conditions for each of the 3 new talkers. On each of the next 5 days, subjects were trained in 158-trial sessions with feedback (total training trials: 3950). Subjects were assigned to one of three training groups (blocked, mixed, or combination, described below). At the time of this writing, 5 subjects had been assigned to the combination group, 12 to mixed, and 18 to blocked.

The blocked group heard 1 talker on each training day (a different talker each day). The mixed group heard equal numbers of stimuli produced by each of the training talkers mixed together in random order each day. The combination group also heard equal numbers of stimuli produced by each talker, but the stimuli were blocked by talker. On day 7, a post-test on the generalization and test materials used in the pretest was given.

Results and Discussion

Blocked, mixed and combination groups all improved significantly from pretest (.64, .65, and .63, respectively) to post-test (.76, .76, and .73) on the /r/-/l/ generalization items (Tukey's HSD post-hoc test, $p<.05$). All groups also showed significant improvement on the test materials between pretest and post-test in both talker conditions ($F>8$, $p<.05$). There were no significant differences between groups.

The interaction of talker and talker condition in the post-test *R-rate* results for each group are shown in Figure 2. The interaction was significant for subjects in the mixed ($F(6,66)=3.1$, $p<.01$) and blocked groups ($F(6,102)=3.7$, $p<.01$). The interaction was not significant for subjects in the combination group ($F(6,24)<1$, $p>.6$). It appears that the addition of more talkers increases the variability in the training set to such a degree that the mixed vs. blocked training advantage observed in the first experiment is diminished. This suggests that a combination of stability and variability may promote the ability to adapt to differences between talkers.

CONCLUSION

Our results indicate that talker variability, and also how it is organized, are important considerations in non-native speech contrast training. Further work is required to determine the nature of the observed trade-offs between stability and variability in learning non-native contrasts.

REFERENCES

[1] Magnuson, J.S., Yamada, R.A., Tohkura, Y., Pisoni, D.B., Lively, S.E., & Bradlow, A.R. (1995). *Proc. Acoust. Soc. Japan, Spring, 1995*, 393-394.

[2] Lively, S.E., Pisoni, D.B., Yamada, R.A., Tohkura, Y., & Yamada, T. (1994). *J. Acoust. Soc. Am.*, 96, 2076-2087.

[3] Yamada, R.A., & Tohkura, Y. (1992). *Perception & Psychophysics*, 52, 376-392.

[4] Magnuson, J.S., & Yamada, R.A. (1994). *J. Acoust. Soc. Am.*, 95, 2872.

ON FORMAL IDENTIFICATION OF SOME PROBLEMS IN NON-NATIVE FRENCH PRONUNCIATION

*J. F. Malet * and N. Vigouroux***
**California State University, Sacramento, USA*
***Laboratoire IRIT, URA-CNRS n° 1399, Toulouse, France*

ABSTRACT

Description of a method to identify non-native learners' phrase-level pronunciation problems. Devised to inform efforts at automatic detection through HMM modelizing, the method is based on a threefold concern with: 1) preparation of relevant phonetic identifiers, through 2) robust acoustic evidence, applied to 3) training data that is also suggested by traditional didactic practices Examples discussed in light of evidence seen in preparing units for modelization.

INTRODUCTION

In a previous publication [1] a set of informally identified pronunciation problems —commonly encountered by English-speaking learners of French— were discussed from the point of view of potential automatic detection and correction.

Formal identification and reliable detection of such problems —as they occur within speech data supplied by learning subjects— require that they be defined and classified according to a method that interacts with the traditional practices used in non-native language acquisition; as opposed to a mere off-shoot application of phonetic science with a distinct, not easily related didactic module (e.g., SPELL Project [2]).

As it turns out, whereas some mispronunciation items are easily identifiable from the dual point of view of their respective theoretical inception levels (e.g., articulatory, phonotactic, morpho-syntactic, etc.) and of their typical acoustic manifestations, some other items are much more difficult to pinpoint and, even more so, to characterize either theoretically or datawise. A threefold investigative method, catering to the above considerations, is therefore attempted.

From a corpus of 56 French phrases (offering wide phonemic and intonative variety) read (under uniform technical recording conditions) by 42 American English-speaking learners of French, a number of apparent pronunciation problems are selected for a multi-level confirmation of their likely manifestation within the acoustic signal. For each problem taken up, data-analytic strategies are sought, aiming at 1)HMM modelization of phonetic units and 2)at relevant phonotypical mapping into an adapted organic version of the evaluation grids, traditionally used in testing for oral proficiency —e.g., accuracy, delivery rate, stress, phonemic quality, etc.

Phrase-level examples are discussed in light of initial observations, made in preparing for modelization of units that are formal phonetic identifier candidates.

METHODOLOGY

Ultimately, identification of the specifics that constitute a language learner's problem-realization, requires a reliable decisional process. And such a process has better chances of succeeding, of course, if it is highly predictive of variability within the test-data submitted to it.

To achieve this with our method, training-data is prepared with two main considerations in mind: 1)selecting this training data (from a set of learners's realizations) for its efficient problem-oriented value and 2)ascertaining physical evidence of apparent pronunciation problems.

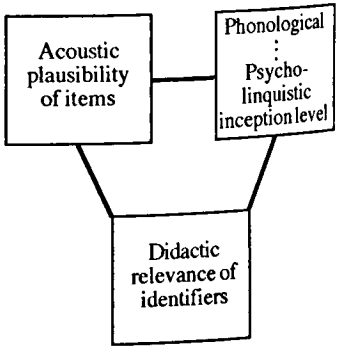


Figure 1. Three-pronged approach to selection of training-data items and formalization of phonetic identifiers.

Training-data Selection

This is a process that can begin with informal observation of non-native accent. However, although common stereotypification of "foreign accents" can initially be somewhat useful, it is preferable to turn to more informed ways, such as long language-lab or classroom experience.

The procedure for choosing training-data is strictly governed by a triad of major concerns that lies at the heart of our method (Figure 1).

In choosing the items, which are to be acoustically modelized in order to acknowledge and/or create useful, formal phonetic identifiers, all observed problem units of speech —and relevant supra-segmental phenomena associated to them— have to be eventually subjected to a decisional process embodying:

- a pertinence strategy guided by didactic needs,
- an acoustic robustness clause excluding speculative observation, and
- a proposal for formal classification of candidate units within a larger, ontological system .

1.The didactic area of concern involves:

a) the preparation of a relevant corpus directly dictated by the content and purpose of a given lesson,

b)once the speech specimens are collected, the preparation of competence groups among the learners —recognized by human expertise,

c) the selection of common learning problems (e.g., in reading aloud, deciphering efforts vs. easy and intelligent reading).

2.The concern with acoustic robustness aims at establishing the physical basis of problems perceived in speech. This involves:

a. the preparation of acoustico-phonetic (AP) data: choice of subjects, collection of speech specimens, manual segmentation and labelling of relevant signal data (a step based on observation of wave-form, spectral display, frequency analysis and listening[3]),

b. a classification of AP data files in competence groups, in general with respect to perceived quality of pronunciation, but also in particular with respect to mastery of chosen difficulties (See examples),

c. a further classification of AP data-files with a view to retaining data that is reliably observable, and suggesting the use of well mastered parameters to bring out potential candidate units or clusters to formal phonetic status.

3.The third area of concern involves the detection of the (most likely) inception level of problem units of realization; whether articulatory, phonotactic, ..., lexical, syntactic, supra-segmental (prosodic), or of higher order (i.e., linguistic and psycho-linguistic levels sharing their competence with didactic concerns —e.g., deciphering versus intelligent reading, as in First Example, below). Determination of the inception level of problem-realizations is necessary to clinch the formalization process of identification. It calls on the joint expertise of teachers, phoneticians, linguists, psycho-linguists, etc..

Summary Of Purpose

Basically, we are looking for types of corpus realization that can be HMM-modelized and confirmed as formal identifiers of a non-native kind of speech.

However, given the high degree of variability, this search cannot be a random one and is better served by a didactic line of progress.

An identifier does not necessarily have to be a special interlanguage unit so long as it can serve an interim learning strategy [4]; it can be a known phoneme (or an infra-phonemic unit, or a phonemic cluster, or again a supra-segmental phenomenon) that is realized in a specific location where it should not occur (e.g., /t/ in First Example, below) or again that is not realized, or hardly so, when it should be (/p/ or [m] in Second Example).

EXAMPLES

Two sample sets of AP datafiles are now briefly examined and commented.

First Example

Our database contains the phrase (actually a sentence): Et c'est sain. (Eng.: "And it's healthy too.") with barebone phonetic transcription /eses$\tilde{\epsilon}$/. This phrase has been selected on the following grounds:

1) it involves twice the same consonant, /s/, which is easy to realize,

2) it uses the same vowel, /e/ twice in a row; inducing assimilation of the third vowel /ɛ̃/ to [ẽ], especially with American Anglophones as such speakers are not used to an [ ]-producing buccal aperture concomitantly with essential nasalization,
3) its metric structure is somewhat anapestic and it can possibly be stressed for rhythm on the last syllable.
As a result, it should not be difficult to choose a didactically useful phonotypical transcription leading to comparison of non-native realizations. Such a theoretical templet can entail:
• no pause (minimum phrasal continuity, reasonably demanded from learners),
• a strictly unvoiced fricative (/s/),
• an easily perceptible and measurable time structure with variants quite directly reflecting nuances in the meaning of the phrase,
• a strict exclusion of any diphthong,
• some fading on the median vowel is preferable to avoid a staccato effect.

A set of 41 learners supplied the voice data. These were divided into 3 groups according to the perceived resemblance of their speech to either French or (American) English pronunciation: Grp.F (15 speakers) was perceived as fairly francophone, Grp.A (17 speakers) as dominantly anglophone. Speakers in the third group were difficult to assess.

Not all speakers realized a sixth phoneme /n/ but, in Table 1, figures are adjusted on the virtual presence of this sixth phoneme.

Mean duration:	Fr.-type 780 ms	Ang.-type 1 963 ms	
% phrase time		Ratio MeanEn./time	
Franco.	Anglo.	Franco	Anglo
e 16.7	e.. 14.9	5.77	4.92
s 19.4	s 16.8	5.71	4.95
e 13.2	ei 17.0	7.15	4.60
s 20.3	s 15.6	5.60	4.95
ɛ̃ 28.4	e 26.8	7.60	2.64
ŋ 10.5	ŋ 14.8	6.10	4.10

Table 1.

As can be observed, both Grp.F and Grp.A realized a somewhat anapestic metric structure, devoting nearly half of phrase-time to the last syllable. However, a wide difference is to be noted between the two groups in that a slight fading of the median vowel occurs in Grp.F, relieving the staccato effect perceived in Grp.A. The monotone ratio figures of mean energy to duration of units might also account for the perception of such an effect. The phenomenon might in fact be owed to strenuous efforts at deciphering each syllable as it comes up in the reading, as opposed to a more competent, flowing rendering of the written corpus.

Aside from these global observations, a number of unwanted pauses were detected. Two cases of extreme fading of the median vowel with a resulting centralizing to /ə/ and an English type of accentuation on the last syllable. On the other hand, ten speakers realized this median vowel taking anywhere between 150 and 320 ms, with five of these realizing a perceptible dipthong [ei]. Six speakers realized a [t] with five of them supplying a phonetic profile [etses] and one [etsetsə].

All such unwanted units are potential candidates for potential modelization and identifier status.

Second Example

The French phrase *un pseudo voeux* (Eng.: "a pseudo wish") was chosen for three didactic *a priori* 's:
1) the gender distinction of the singular indefinite adjective,
2) the requirement to delete an English phonological rule (initial occl.+ /s/ —> /s/),
3) the mastery of vocalic timbres for /ø/ and final /o/.
Out of 37 learners, who read the phrase, two different types of realizations turned out for the first three syllables: respectively, along a fairly francophone phonetic scheme /œ̃psødo.../ (12 speakers) and along a quite definitely anglophone profile [œ̃nsyIdʰəo...] (15 speakers) where [yI] symbolizes a vowel that can be perceived either as some French /y/ or as an English /I/ but is neither. This [yI] can be considered a good candidate to become an interlanguage vocalic identifier.
In Table 2, it can be seen that the anglophone type of subjects took, on average, some 45 % more time to realize the first three syllables of the phrase. While a more detailed examination of the data points to, at least, three areas of discrepancy in the duration of the phenomena looked at:
1) francophone-type [p] and (in this specific context only) its *de facto* substitute [n], with a possibility of treating [œp+ s] and [øn+s] as clusters, for /œ̃[m]sødo/ is

Mean duration:	Fr.-type 1 034 ms	Ang.-type 1 499 ms	
% phrase time		Ratio MeanEn./time	
Franco.	Anglo.	Franco	Anglo
œ̃ 12.2	œ̃ 6.1	5.93	6.41
p 9.2	n 4.5	4.86	7.62
s 16.6	s 12.6	5.26	3.42
ø 8.2	yI 4.2	8.52	8.59
d 6.4	dʰ 4.2	7.60	7.36
o 17.0	əo 49.9	3.89	

Table 2.

also a native francophone type of realization
2) francophone-type [ø] and interlanguage candidate identifier [yI],
3) francophone-type [o] and American English-type [əo] or [əɔ].
Whereas [p] and [n],in this context, reveal a very different ratio of mean energy to duration —4.8 vs 7.62— vowel realization is not as clearly differented by this ratio and calls for frequential parameter definition (through formant tracking or, more recently, noise in certain frequency bands [5]).

CONCLUSION

The method, presented in this article, offers seemingly endless investigative possibilities, while it remains *de facto* contained by the practical and realistic demands of assisted non-native language acquisition.

At the same time, whereas a certain tendency to sprawling investigation is inherent to a thorough observation of bulk AP data, such a tendency is checked by the necessity to justify the creation of ever more formal identifiers —for these must eventually be fitted in the larger context of a language science.

However, as an empirical tool serving didactic purposes whence it is in great part derived, this method appears promising.

Obviously,the training-database required to achieve automatic detection of mispronunciation in test-data supplied by learners, has to be considerably enlarged so has to enable HMM modelizing of potential formal phonetic identifiers.

REFERENCES

[1][Malet 1992] J.F. Malet, G. Pérennou & N.Vigouroux, "Repérage automatique d'unités acoustico-phonétiques pour l'enseignement de la prononciation française," *Journal de Physique*, Suppl. IV, 1992.

[2][Lefèvre 1992] J.P. Lefèvre, M. Jack, C. Maggio, M. Recife, M. Savino & L. Santagelo, "An Interactive System For Automated Pronunciation Improvement" in *Proceedings Of ICLSP-92*, Banff, Canada.

[3] *La parole et son traitement automatique*, CALLIOPE, Eds. J.P. Tubach, L.J. Boe, P. MArtin, J. Caelen, J.M. Pierrel R. Descout, C. Sorin, J.J. Mari-ani (Massons: Paris, 1989).

[4] N. Yamada, "Japanese Accentuation Of Foreign Learners And Its Interlanguage," ICSLP'94, Yokohama, pp. 1227-1230. IV, 1992.

INVESTIGATING IMITATIVE ABILITY

Duncan Markham
Department of Linguistics and Phonetics, Lund University, Sweden

ABSTRACT

This paper introduces a newly initiated research project dealing with imitation, entitled Human Imitation: Perceptual and Productional Processes. The project is motivated by the poor understanding of phonetic behaviour and ability in second language acquisition, and voluntary adoption of non-native dialect, accent and speaker characteristics.

INTRODUCTION

At present there exists little conclusive research to explain why phonetic acquisitional capacity (or any other acquisitional capacity in linguistics) appears to deteriorate with age. It is generally assumed, and has been empirically demonstrated, that both perceptual discrimination of unfamiliar phonetic contrasts (eg [1]) and productional ability (accentedness in L2) (eg [2]) do deteriorate with age, but there have also been investigations indicating no age effect (cf [3, 4]). Other investigations have demonstrated that general multilingual experience does not facilitate imitative/acquisitional ability [5], contrary to another common belief. Much of the published research, including some mentioned above, draws powerful conclusions from comparisons with other research in phonological acquisition, despite often grossly different research methods and controls. This research project attempts to take into account as many of the potential influencing factors as possible, and specifically addresses the question of individual ability — an issue ignored almost entirely in phonological performance studies.

PILOT INVESTIGATION

The results of a pilot investigation of imitative ability will be presented here. The investigation served as a test of the elicitation methodology and experimental design to be applied in the main investigation. Our primary interest in the pilot investigation was the experimental structure and the reactions of the informants, rather than a detailed investigation of their performance. This was necessary, as the design places considerable demand on the speakers' ability and concentration, and involves a large amount of preparation and training. In the following section, the investigation and its results will be described. The structure of the main investigation will then be described.

A text, approximately semantically equivalent, was prepared for a number of languages. Each text was read by at least one native speaker at a slow speed and at a faster casual/colloquial speed. A word list was then read once by each speaker. The word list contained all words from the text, each word appearing twice in the list. Two minutes of spontaneous speech were also recorded for each speaker (in most cases, the speaker described his house). Native speakers were selected on the basis of age (20-45) and modal voice, as far as practicable, as the imitators were to be of similar age (20-35), and in order to maximise the number of phrases which could be extracted as stimuli (see below).

The recordings were made on DAT at a sampling rate of 44.1kHz. They were then resampled using the SoundScope hardware/software package on a Macintosh computer at 22.2kHz. The texts were examined for short phrases or subphrases suitable for use in the experiment. The phrases had to be syntactically whole, in that constituents couldn't be divided. For example, in the phrase *the linguist parses the sentence*, the subphrases *the linguist, the linguist parses, parses the sentence, the sentence* would be acceptible, whilst **linguist parses, *linguist parses the, *parses the, etc.* would not. Furthermore, the phrases had to exhibit fairly modal voicing characteristics. Syntactically conditioned creak was acceptible, but speakers who exhibited strong creak as part of their reading register in general, or in their lower range for F0, were not used. The phrases were excerpted from the texts at both speeds, as were the relevant words

*short tone (450Hz, 125ms) // three instances of a fast phrase eg, *the linguist parsed the sentence*;
*short tone // three instances of the slow phrase;
*two-tone signal (450Hz, 125ms; 600Hz, 125ms) // two instances of the first grammatical word in the phrase (from the word list) eg, *the ... the*;
*two-tone signal // two instances of the second grammatical word in the phrase;
...
*two-tone signal // two slow instances of the first subconstituent eg, *the linguist ... the linguist* ;
*two-tone signal // two fast instances of the first subconstituent;
...
*two-tone signal // two slow instances of the phrase;
*two-tone signal // two fast instances of the phrase.

Figure 1. Stimulus presentation.

from the word lists.

A battery of stimuli were prepared for each language. For each phrase (two per language), the informants would hear complete phrases, and then the individual words or subconstituents, as shown in Figure 1. The stimuli were then recorded to DAT again, preceded by the fast text, the slow text, and (for some languages) the spontaneous speech.

Two informants (one male, one female) were chosen for the pilot investigation, on the basis of their above average performance in the pronunciation of some foreign languages. It was, however, necessary to choose informants of lesser apparent imitative ability than those participating in the main investigation, in order to maximise the number of good imitators remaining for the latter investigation (informants could not participate in both the pilot and main experiments). The informants were native speakers of Swedish, 26 and 29 years old respectively. The male spoke Southern Swedish (*skånska*), whilst the female spoke with a mixed regional accent of Halland, on Sweden's west coast (*halländska*) and *skånska*.

Three weeks before the experiment, the informants received an audio cassette containing the read texts and spontaneous speech for a selection of the languages. They were required to listen to the cassette twice on ten separate occasions (2 x 10=20 listenings) during the three weeks. The informants then came to the Department for the experiment. Recordings were made in the Department's recording studio.

Prior to the imitation task, the informants were fitted with a lightweight (30 gm) headset microphone, and asked to read a small part of the text used above (and with which they were, by now, familiar) in a dialect other than their own. They were also asked to read text excerpts, in as native-like a manner as possible, for languages in which they had adequate competence. For all dialects or languages they were also asked to introduce themselves as fictional persons, according to details they were given on paper (eg, Johnny, 29, London, student, photography, New York). These tasks were used to gain an impression of each informant's a) delayed imitative ability (strength of existing dialect models), b) base pronunciation performance in L2s, and c) performance in a non-reading (more freely structured) task.

For Swedish, the texts and introductions were done with regional pronunciation for Stockholm, and Finland-Swedish, with which all Swedes can be said to have some familiarity. For English, 'British English' and 'American English' were specified. For German and French no specific dialects were requested.

The informant was then told of the structure of the imitation task, and how the stimuli would be presented. This was done using visual information — pictures of the approximate sequence of signal-tones, texts, phrases, etc. The stimuli were then presented over high-quality headphones. The informant could not see the author, but communication was possible via a loudspeaker in the recording studio. After each two-tone signal, the informant heard two instances of a stimulus and then attempted to imitate it. Adequate time was available for each imitation.

After completing this task (total duration: 32 mins), informants were permitted to leave the studio, and were provided with some refreshments. After approximately 10 mins, the imitation task was repeated.

Canadian English (female, 24, Toronto) — TRAINING LG
have to work nights: have, to, work, nights
Japanese (female, 43, Osaka) — TRAINING LG
Sutefan wa pailotto de (Steven is a pilot): Sutefan wa, pailotto de
futari wa Nihon Kookuu de hataraite imasu (they both work for Japan Airlines): futari wa, Nihon Kookuu de, hataraite imasu, futari was Nihon Kookuu de, Nihon Kookuu de hataraite imasu (whole phrase not imitated, due to length)
Finland-Swedish (female, ca 40, Helsinki)
men omväxlande arbete (but varied job): men, omväxlande, arbete
till många storstäder runt om i världen (to many cities all over the world): till, många, storstäder, till många storstäder, runt, världen, runt om i världen
French (male, 45, Grenoble, no strong regional accent) — TRAINING LG
au deuxième étage (on the second floor): au, deuxième, étage
c'est pareil pour Maria (the same is the case for Maria): c'est, pareil, pour, Maria, c'est pareil, pour Maria
Mandarin (male, ca 50, Beijing)
dōu yǒu qìchē (each has a car): dōu, yǒu, qìchē
sānge fángjiān (3 rooms): sān, ge, fángjiān

Figure 2. Language and stimuli used.

RESULTS

Text Reading and Self-Introduction

The male informant, JE, showed interesting behaviour for the two Swedish dialects and two English dialects elicited. The text and introduction had been designed to highlight allophonic differences between speakers' native dialect and the imitated dialect. For JE, the clearest allophonic changes required for Stockholm Swedish were [ɧf] → [ʂ/ɧx], [ʀ] → [r/ʐ], [ᵊi:] → [i:]. and a change in pitch accent realisations. The first allophonic change, and the pitch accent change, were not attempted, which was surprising as they are considered to be highly salient dialect cues. For Finland-Swedish, on the other hand, JE made attempts to produce dialect specific [r ʃ y: ʉ:], typical breathiness and pitch accent, but not [tɕ] for [ɕ]. His productions of British English (attempted RP) sounded non-native, primarily because of the realisation of /ai/ and /a:/ with [ɒ] instead of a more front a-vowel. American English was more acceptable, though problems with the diphthong just mentioned were also observed. JE's normal English production tends towards RP, but residence in an English speaking environment was limited to the USA.

The female speaker, PO, showed much less resistance to pitch accent change required for three Swedish dialects (*skånska* was added), but she didn't modify *intonational* behaviour. Her segmental production showed clear and fairly good attempts at dialect-allophony for all dialects. PO's productions of both varieties of English were at times characterised by typically Swedish dental stops, but most other segmental and prosodic requirements were well met. Both informants had problems with reduction in both American and British English. The text included the sequence *so that they can*, and proved exceedingly difficult, as they tried to produce each syllable and fricative, rather than reducing the second and fourth syllables, which would normally result in something resembling [sᵊʊˈðɛɪkʰn̩]. JE's productions were always less native-like and sounded more exaggerated than PO's for all dialects.

Imitation

The two informants behaved at times quite differently in this task. JE behaved very poorly with regard to intonation reproduction, even for single word stimuli. There was slight improvement on the second trial. PO was considerably better, though still showed many errors. Her weakest area, however, was vowel quality, contrasting with fairly good performance for consonants. JE was far more variable, though generally not as good as PO, as he had difficulties with segmental production, utterance length (tempo), and rhythmical characteristics in French and Japanese (especially gemination in the latter — unexpected, as Swedish makes use of gemination contrasts).

Of considerable interest was the way in which intonation in phrases was usually reproduced. Generally, the onset of imitation was good, but the highly salient final tonal movements were not. This is curious, considering the perceptual salience of final tonal movements for expressing both syntactic and pragmatic functions, such as questions, scepticism, sadness, etc. Similar behaviour was observed for imitation of prosody only (rather than the segmental characteristics of an utterance) in a previous study [6].

MAIN INVESTIGATION

The main investigation takes essentially the same form as the pilot investigation. There are approximately eight informants (native speakers of Swedish), though it is hoped that further informants may be located. All informants show near-native ability in the pronunciation of at least one second language (L2). It is relatively difficult to find and recruit speakers who are this good and are prepared to be made aware of their imitative deficits, as language, and especially phonetic, ability is a sensitive part of the human ego.

Speakers will be screened prior to the training phase, so as to gain an impression of their present abilites. The screening involves both the initial reading and introduction task described above, and some imitative tasks. Shortly thereafter, half of the informants will receive an audio tape containing training material. This material will be different for each informant, as it will not include L1 or L2 dialects in which a given informant is already competent. The remaining informants receive no training.

The imitation task will take a similar form to the pilot investigation. The number of languages/dialects will, however, be increased to six. Additional imitations of the same stimuli will be made after the first attempt, rather than presenting the entire battery again. This means that rapid improvement for imitations of the same stimulus might be observable, as informants can monitor and correct their productions with little delay. This should also reveal the stimulus characteristics which remain unidentified or unproduced by a given informant, more readily than the delayed 'second chance' given in the pilot experiment. After the phrase tasks for a given language, informants will also be presented with a number of extra-contextual word stimuli, and phonological contrast-pairs (eg chair~share), for allophones or phonetic contrasts which are known to be problematic for speakers of Swedish (eg, alveolars, voicing in fricatives, syllabic consonants).

The elicited material will be edited and test tapes will be prepared for the next phase of the investigation. Trained phoneticians will assess the closeness of the imitations, and degree of non-native accent, auditorily (including discrimination tests — accented~unaccented) and attempt to describe deviations. The material will then be analysed instrumentally to assess acoustic deviation and its relationship to perceived deviation.

REFERENCES

[1] J. F. Werker and R. C. Tees (1984), Cross-language speech perception: Evidence for perceptual reorganization during the first year of life, *Infant Behavior and Development*, vol. 7, pp. 49-63.

[2] S. Tahta, M. Wood and K. Loewenthal (1981), Foreign accents: Factors relating to transfer of accent from the first language to a second language, *Language & Speech*, vol. 24, pp. 265-272.

[3] J. E. Flege (1987), The production of 'new' and 'similar' phones in a foreign language: Evidence from the effect of equivalence classification, *Journal of Phonetics*, vol. 15, pp. 47-65.

[4] C. E. Snow and M. Hoefnagel-Höhle (1977), Age differences in the pronunciation of foreign sounds, *Language & Speech*, vol. 20, pp. 357-365.

[5] J. F. Werker (1986), The effect of multilingualism on phonetic perceptual flexibility, *Applied Psycholinguistics*, vol. 7, pp. 141-156.

[6] D. J. Markham (1994), Prosodic imitation: Productional results. In ICSLP'94, vol. 3, pp. 1187-1190, Yokohama: Acoustic Society of Japan.

AERODYNAMIC EFFECTS ON SECOND LANGUAGE ACQUISITION

Geoffrey S. Nathan, Southern Illinois University at Carbondale

Alice Faber, Haskins Laboratories

ABSTRACT

While languages normally select VOT values for their stop series from three possible categories—long lead, short lag and long lag, most languages prefer adjacent pairs. English, with primarily short and long lag stops contrasts with Spanish (and many other languages) which exhibit lead and short lag pairs. Israeli Hebrew is one of a few languages which may exhibit the extreme values. Speakers of voicing lead-type languages do not need to adjust their voiced stops to produce comprehensible English, but two different studies reported here show that speakers of such languages do indeed begin to produce a significant number of short lag stops for velar targets. We suggest that the increased physiological effort required to sustain transglottal airflow while the supraglottal tract is obstructed relatively close to the glottis (as with velar stops) explains the spontaneous devoicing of velar stops by bilinguals acquiring an English-type system.

It is generally understood that languages have available three categories of stop 'voicing'. Using the terminology adopted by Lisker and Abramson [1], we find overall three possible categories: voicing lead (a relatively long period of voicing before release of the articulators), short lag (a relatively short period of voicelessness after release of the articulators--the lag can vary from none at all to roughly 30 ms) and long lag, a relatively long period of voicing following the release of the articulators. It is well-known that the VOT systems of languages differ systematically within these quite constrained parameters. As [1] and numerous others have shown, most languages tend to belong to one of two possible types—voiceless aspirated vs. voiceless unaspirated as one choice (Mandarin, German, probably English) and voiceless unaspirated vs. voiced (French, Italian, Russian) as the other. This is documented extensively in [2]. We also find, of course, languages such as Thai and Hindi with all three categories, but this paper will be limited to consideration of only the two-category languages, which Maddieson, [3], shows to be the majority of languages in his survey

As pointed out in various places (e.g.[4], [5]) languages seem to prefer the use of adjacent VOT contrasts. Maddieson, [3], finds that of the 162 languages in his survey which have a two-way contrast, 72% have a lead/short lag contrast. The vast majority of remaining systems (exact percentages are unavailable) consist of short lag/long lag contrasts. Research reported in [5] indicates that very few languages seem to opt for a contrast between voiced and aspirated stops without the presence of the intermediate category.

One of the languages which has been reported as belonging to the set of lead/long lag languages is Israeli Hebrew, although the exact degree of aspiration within the so-called aspirated series is somewhat in doubt. [6] argues that Israeli voiceless stops are not aspirated but [7], [8] says they are, at least to some extent. Recent instrumental studies also show the presence of aspiration in the voiceless stops, although there is some variation in the results reported.

On the other hand, Spanish is a classic lead/short lag language, with all voiced stops showing long voicing lead, and all voiceless stops showing short lag (data can be found in [1] and numerous other locations).

Bilinguals present an interesting question. What happens when speakers of a language with one system confront a second, conflicting system? We know that speakers of Spanish normally transfer their Spanish system to their English, at least at the early stages. [9], [10] has shown that eventually Spanish speakers acquire the aspirated category.

although he reports that they do not modify their voiced stops towards the English system of short lag for 'voiced' stops. In a recent study ([5]) it was found that bilingual Israeli Hebrew speakers maintain their voiced stops at traditional voicing lead levels, but that their voiceless stops seem to take an intermediate value between short and long voicing lag.

In several studies carried out by the authors we have investigated voicing lead in the acquisition of English, and we find that the results are somewhat more interesting. In the first study, [4], a longitudinal examination of the initial stops of Spanish first-language speakers was reported. In that study speakers of various dialects of South American Spanish produced English words with initial stops twice during their period of study at Southern Illinois University, with an interval of eighteen months separating the taping. The first set of data contained long voicing lead stops for all three points of articulation, but the second set of data showed considerable movement at the velar point of articulation. Here we find a significant number of 'g' tokens produced with short lag values. Almost fifty percent of the /g/'s were produced with a positive value for VOT.

The second study we will report deals with an apparently balanced bilingual speaker of Israel Hebrew and American English. The subject speaks both languages with no apparent accent. The Hebrew values are exactly what one would expect from a speaker of Israeli Hebrew (see Table 1)

Table 1. VOT values for Hebrew stops of English-Hebrew bilingual.

VOT	Bilabial	Dental	Velar
Vcd	-112.8	-137.9	-73.6
Vcl	50.3	54.7	99.6

Here we can see that the voiced values fall well within the normal voicing lead range, while the voiceless values are a little short for voiceless aspirated values, but much too long for 'short' lag.

Much more interesting, however, are the values for English stops (Table 2):

Table 2: VOT values for English stops of English-Hebrew Bilingual

VOT	Bilabial	Dental	Velar
Vcd	-109	-116.3	-22.9
Vcl.	60.8	95.5	119.8

While the voiceless values are typical aspirated stop types, and the voiced bilabial and dental values are almost identical to the Hebrew set, the velar is radically different. And in fact we have a case where the mean obscures a bimodal distribution. We found the following (Table 3):

Table 3: Individual VOT values for English /g/ target.

	Short Lag	Long Lead
	33.5	-72.7
	45.8	-127.7
	36.8	-52.1
	45.1	
Mean	40.3	-84.1

Clearly our subject has a significant number of short lag 'voiced' stops, with a mean VOT well within the short lag value for velars (which tend in general to have longer VOT lags than labials and dentals).

Again we have found a pattern that velars do not seem to fit into the VOT category of the other two stops in cases where there is a conflict of systems. Interestingly, this pattern also occurs in looking at overall phoneme inventories. Maddieson says [3] that among those languages which have voiced stops, /g/ is much more likely to be missing than /b/ or /d/.

Why should the velar stops be somehow resistant to voicing? And why only in conditions of instability such as in second language acquisition? One possible explanation can be given in terms of aerodynamics. In order for vocal-cord vibration to take place, it is necessary for air to flow past the vocal cords. In order for airflow to occur, there must be a sufficiently large supraglottal cavity to permit the air to go there. In addition, some change in the size of some air cavity is required to produce the airflow in the first place. Of course, the act of expiration would cause

some airflow, but since we are dealing with stops, unless the supraglottal cavity expands there is no region of reduced pressure towards which the air can flow. In the case of labial and dental/alveolar stops there is the general oral cavity, and a number of researchers have found that there are slight gestures of oral cavity expansion accompanying long lead stops. Westbury, [11], Bell-Berti, [12] and Bell-Berti and Hirose, [13] have presented evidence of slight openings of the jaw, expansion of the cheeks and elevation of the velum accompanying voicing of stops. Westbury and Keating, [14], present a computer model of vocal tract aerodynamics confirming the claim that extra effort is required to maintain voicing during closure.

For velar stops, however, we have only limited use of the above supraglottal gymnastics, since most of the oral cavity is blocked off by the velar contact. Thus cheek expansion, and jaw opening will not increase the size of the pharyngeal cavity (although velar raising might). Consequently, it requires considerably more effort to produce truly voiced velars than voiced versions of the other points of articulation.

If we assume that speakers are always attempting to reduce the amount of effort they are required to produce, we can suggest a possible reason for the fact that voiced stops in acquired English would be vulnerable to pressure for devoicing. In English there is no contrast between long lead and short lag stops, and most dialects of English show somewhat free variation between them, with a preference for short lag versions. Interestingly, no one has ever investigated the conditions that govern the variation in voicing of 'voiced' stops in English in any detail.

In any case, English, in a sense, 'doesn't care' what kind of voiced stops it uses. Spanish does, and Hebrew does, but if we assume that speakers are under constant pressure to do as little as possible, these bilinguals have apparently succumbed to the articulatory desire to reduce effort by modifying their English velar stops to the easier, short lag VOT values.

REFERENCES

[1] Lisker, L., and A. S. Abramson. (1964), "A cross-language study of voicing in initial stops: Acoustical measurements." *Word* vol. 20, pp. 384-422.

[2] Keating, P., W. Linker, and M. Huffman. (1983), "Patterns of Allophone Distribution for voiced and voiceless stops." *Journal of Phonetics*, vol. 11 pp. 277-90.

[3] Maddieson, I. (1984), *Patterns of Sounds*. Cambridge: Cambridge University Press

[4] Nathan, G. (1987), "On second-language acquisition of voiced stops." *Journal of Phonetics*, vol. 15, pp. 313-322.

[5] Raphael, R., Y. Tobin, A. Faber, T. Most, H.B. Kollia and D. Milstein. (1995), "Intermediate values of voice onset time." in *Producing speech: contemporary issues for Katherine Safford Harris*. F. Bell-Berti and L.J. Raphael, eds. Woodbury N.Y.: American Institute of Physics Press.

[6] Rosén, H.B. (1966) *A textbook of Israeli Hebrew. second, corrected edition*, Chicago: University of Chicago Press.

[7] Blanc, H. (1956), "A selection of Israeli Hebrew speech." *Leshonenu*, vol. 21, pp. 33-39

[8] Blanc, H. (1964), "Israeli Hebrew texts." in *Studies in Egyptology and Linguistics in Honor of H.J. Polotsky*, Jerusalem: Israel Exploratory Society.

[9] Flege, J.E. (1987), "The production of 'new' and 'similar' phones in a foreign language: evidence for the effect of equivalence classification. *Journal of Phonetics*, vol. 15, pp. 47-65.

[10] Flege, J.E. (1991), "Age of learning affects the authenticity of voice onset time (VOT) in stop consonants produced in a second language." *Journal of the Acoustic Society of America, vol.* 89, pp. 395-411.

[11] Westbury, J.R, (1983) "Enlargement of the supraglottal cavity and its relation to stop consonant voicing," *Journal of the Acoustical Society of America*, vol. 73 pp. 1322-36.

[12] Bell-Berti, F. (1975) "Control of pharyngeal cavity size for English voiced and voiceless stops," *Journal of the Acoustic Society of America*, vol. 57, pp. 456-61

[13] Bell-Berti, F. & H. Hirose. (1975), "Palatal activity in voicing distinctions: a simultaneous fiber optic and electromyographic study," *Journal of Phonetics*, vol. 3, pp. 69-74.

[14] Westbury, J. R. and P. A. Keating. (1986), "On the naturalness of stop consonant voicing." *Journal of Linguistics*, vol. 22 pp. 145-166.

PERCEPTUAL LEARNING OF JAPANESE MORA SYLLABLES BY NATIVE SPEAKERS OF AMERICAN ENGLISH: EFFECTS OF TRAINING STIMULUS SETS AND INITIAL STATES

Tsuneo Yamada
Osaka University
Suita, Osaka 565, Japan

Reiko A. Yamada
ATR Human Information Processing Research Laboratories
Seika, Kyoto 619-02, Japan

Winifred Strange
University of South Florida
Tampa, Florida 33620, USA

ABSTRACT

Native speakers of American English were trained to identify Japanese short vowels, long vowels and special phoneme /Q/. This laboratory training using natural tokens was effective for generalization into unfamiliar phonemic contexts and unfamiliar talker. Different sets of training stimuli had a significant effect on the early stage of learning processes, but not on the final stage.

1. INTRODUCTION

Native speakers of American English (AE) have difficulty in perceiving some of the sound contrasts in Japanese. For example: /i/ plus /j/ versus palatalization of the preceding consonant (e.g., /ki-ja/ vs. /kja/), long vowel versus short vowel (e.g., /sa-a/ vs. /sa/), presence versus absence of the phoneme /Q/, which makes a geminata with the following consonant by prolonging the following consonant (e.g., /sa-ta/ vs. /sa-Q-ta/ or /sa-t-ta/), and so on. Suprasegmental contour and moraic unit play essential roles in distinguishing these contrasts. In the above examples, hyphens show the moraic boundary. Note that members in each contrast differ in the number of morae.

Difficulties arise for adult English speakers learning Japanese due to the differences in the phonological systems of the English and Japanese. It had been believed that adult learners hardly learn novel phonetic contrasts that do not occur in their own language. Recently, however, it has been experimentally demonstrated that even adult learners can modify their phonetic categories, if adequate training methods are used (cf. [1,2,3]). Segmental phonetic contrasts, such as English /ð/ vs. /θ/ for native speakers of French, and English /r/ vs. /l/ for native speakers of Japanese, were used as training items in those studies.

In this paper, we focused on the suprasegmental contrasts rather than segmental contrasts, and examined AE speakers' perception of Japanese short vowel vs. long vowel, and short vowel with /Q/ vs. short vowel without /Q/. In the experiment, these two contrasts were examined together, using "short vowel", "long vowel" and "short vowel plus /Q/" triplets. We trained adult AE speakers to identify Japanese /CV-CV/, /CV-V-CV/ and /CV-Q-CV/ syllables in order to determine whether adult learners can improve their ability to distinguish these suprasegmental contrasts. Processes of learning and generalization, and the effects of variability in the training stimuli on those processes will be discussed.

2. METHODS

2.1. Subjects

Fifteen native speakers of American English served as subjects. They were monolinguals whose dialect was General American, and had no Japanese language experience.

2.2. Stimuli

Stimulus materials were Japanese syllables which constituted minimal triplets, /CV-CV/, /CV-V-CV/, /CV-Q-CV/, where the first CV and the final CV were identical among the members (Table 1). Triplets with various consonants and vowels were produced by three female and one male native speakers of Japanese (Table 2). There were two sets of training materials, the C2-variable and V1-variable sets. In the C2-variable set, C2 varied among 10 consonants whereas V1 was fixed as /o/. In contrast, in the V1-variable set, V1 varied among 5 vowels whereas C2 was fixed as /k/. In order to have the same number of triplets in these sets, two different combinations of C1 and V2 were used in the V1-variable set, and only a single combination was in the C2-variable set. Thus, each set consisted of 10 triplets. The training stimuli were produced by the three female talkers.

Table 1. Examples of the minimal triplets contrasting short vowel, long vowel and /Q/ used as stimulus materials.

Response Category	Structure	Examples
short V	C1V1-C2V2	/kato/
long V	C1V1-V1-C2V2	/kaato/
Q	C1V1-Q-C2V2	/kaQto/

In addition, there were two sets of test materials, the full-test set and mini-test set. In the full-test set, all possible combinations of five vowels and twelve consonants, i.e., 60 triplets, produced by all four talkers, were used. Three of these talkers were those used in the training, and the fourth was the male talker not used in the training. The mini-test set was a subset of the full-test set (see Table 2), and consisted of only the tokens produced by the three talkers used in the training. Syllables in the test sets were selected among the ones which were not used in the training sets.

2.3. Procedures

Subjects were assigned to the C2-variable training group, V1-variable training group or control group. In addition to the pretest and post-test, a mini-test on every training day and an interim test just in the middle of the entire training phase were administered. The experiment consisted of eleven daily sessions. Familiarization and pretest were administered on Day 1, and followed by four days of training (Day 2-5), an interim test (Day 6), four days of training (Day 7-10) and a post-test (Day 11). On every training day, while subjects in the C2- and V1-variable training groups took both one training session and one mini-test session, those in the control group took only the mini-test session.

In the pretest, interim test and post-test, the full-test set of 720 stimuli (60 triplets by 4 talkers) were tested without repetition (720 trials). In each training session, 10 triplets by 3 talkers (i.e., 90 stimuli) were presented randomly with three repetitions (270 trials). In the mini-test session, 14 triplets by 3 talkers (i.e., 126 stimuli) were presented without repetition (126 trials).

Table 2. Combinations of V1 and C2. Combinations for the V1-variable training were enclosed with a square, and those for the C2-variable training were enclosed with ellipses. Combinations marked with "f" were used in the full-test, and those marked with "m" were used in the mini-test.

C2 \ V1	a	i	u	e	o
k	f,m	f,m	f,m	f,m	(f,m)
s	f	f	f	f	(f,m)
t	f	f	f	f	(f,m)
h(F)	f	f	f	f	(f,m)
g	f	f	f	f	(f,m)
z	f	f	f	f	(f,m)
d	f	f	f	f	(f,m)
b	f	f	f	f	(f,m)
p	f	f	f	f	f
ʃ	f	f	f	f	(f,m)
tʃ	f	f	f	f	f
dʒ	f	f	f	f	(f,m)

A forced-choice task from three alternatives, "short-V," "long-V" and "short-V+Q", were used in all the tests and the training. Subjects were seated in a quiet room, and listened to the stimuli through a headphone (Stax, SR-Λ Signature). They responded by typing one of the keys, which corresponded to the alternatives. In the tests, any feedback was not given to the subjects. In the training, in contrast, feedbacks were given to them immediately after the responses; chime sound and graphical coin for correct response, and buzzer for incorrect response. Experimental events and data collection were controlled by workstation (NeXT Cube Turbo).

3. RESULTS & DISCUSSION

3.1. Effects of training: Comparisons in pretest, interim test and post-test

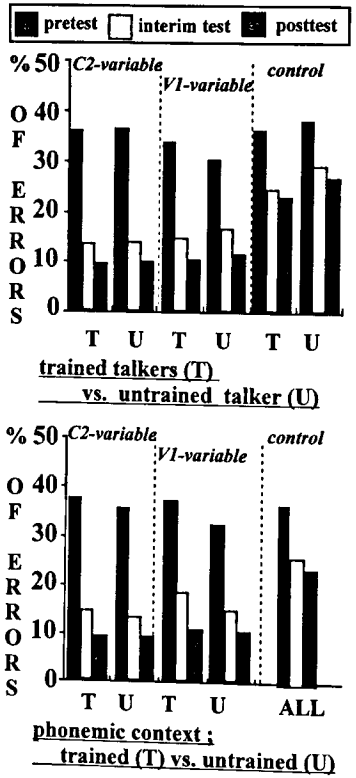


Figure 1. Generalization processes in interim test and posttest. The ordinate indicates mean percentage of errors in sub-categories, (1) talkers used in training vs. an unfamiliar talker (upper) , and (2) trained phonemic contexts vs. untrained contexts (lower).

Error rates in pretest, interim test and post-test are shown in Figure 1. Mean error rates in the pretest were around 30-40% in all three groups, which were lower than chance level of 67%. Such low error rates stemmed from the fact that AE listeners could distinguish short vowels from long vowels without training in some phonemic contexts, although they had difficulties when the contrasts occurred before /s/, /ʃ/, /tʃ/, or /k/.

Both the C2-variable and V1-variable training groups showed significant improvements in the interim test and post-test. These were true not only for the stimuli produced by trained female talkers but also for those produced by an untrained male talker (Figure 1, top panel), suggesting that training effects generalized into the unfamiliar talker. The equal amounts of improvements were also observed in trained phonemic contexts and untrained phonemic contexts in both training groups (Figure 1, bottom panel), suggesting that training effects generalized into unfamiliar phonemic contexts. These results showed that, through such training, adult AE speakers modified their phonetic categories, so that they can distinguish short vowel, long vowel and short vowel plus /Q/.

3.2. Acquisition process: Analysis of mini-tests.

Error rates in mini-tests are shown in Figure 2. While the C2-variable training group showed drastic improvements in the first three sessions, the V1-variable group needed larger number of training sessions to generalize into both trained and untrained phonemic contexts.

There observed a significant effect of consonant (C2) on the error rates: The voiceless fricatives (/s/, /h(F)/ and /ʃ/) made the distinction of the preceding target part, contrasting short vowel, long vowel and short vowel plus /Q/, difficult to distinguish. Since Japanese /Q/ is a geminata, in which the second consonant is preceded by the homorganic consonant, the /Q/ mora is filled with noise when the second consonant is a fricative, while it is filled with silence for the other consonants. The above result implies that the /Q/ filled with silence is easier for AE speakers to distinguish in the contrasts than the /Q/ filled with noise. In contrast to the effect of consonant, there was no effect of vowel (V1). These results pointed out that variation in the following consonant affects the perception more strongly than variation in the preceding vowel: Furthermore, the stimulus set with high variability in the following consonant is effective for generalizing in the early stage of learning on such contrasts.

4. CONCLUSIONS

The native speakers of AE learned the distinction of the Japanese moraic contrasts, among short vowel, long vowel and short vowel followed by /Q/. The training effects generalized not only into the untrained phonemic environments, but also into the untrained talker. The acquisition processes were clarified by analyzing performance in the mini-tests, which were administered after every training session. It was found that distinction in the contrasts is difficult when a fricative is the following consonant, and that training with high variability in the following consonants facilitates acquisition in the early stage of training. However, the enough amount of training trials promotes the equivalent improvements in the training with low variability in the following consonants.

Speech perception is originally to be *learned* in an individual linguistic environment. Therefore, the viewpoint that speech perception has dynamic aspects as an acquisition process is also necessary. In this paper, it was found that the acquisition process is affected by the structure of training stimuli. It was implicated that analysis of the learning process is essential for clarifying the mechanisms to acquire non-native contrasts.

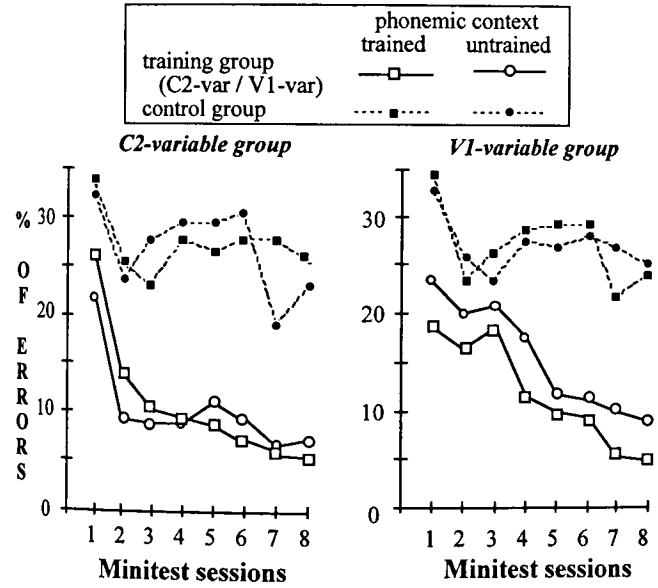


Figure 2. Generalization processes in the minitests. The ordinate indicates mean percentage of errors in each training condition, that is, (1) training group vs. control group and (2) trained phonemic contexts vs. untrained contexts. Constituents of each context condition in control group correspond to those in each training group.

ACKNOWLEDGMENTS

This research was partly supported by a fellowship from the Ministry of Education, Science and Culture, Japan to T. Yamada. We wish to thank Desiderio Saludes and Brett H. Fitzgerald for their assistance in conducting this experiment.

REFERENCES

[1] Jamieson, D.G., & Morosan, D.E. (1986). Training non-native speech contrasts in adults: Acquisition of the English /ð/-/θ/ contrast by francophones. *Perception & Psychophysics*, **40**, 205-215.

[2] Logan, J.S., Lively, S.E., & Pisoni, D.B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *Journal of the Acoustical Society of America*, **89**, 874-886.

[3] Strange, W., & Dittmann, S. (1984). Effects of discrimination training in the perception of /r-l/ by Japanese adults learning English. *Perception & Psychophysics*, **36**, 131-145.

SOME CHARACTERISTICS OF VOT IN PLOSIVES SPOKEN BY ARABIC LEARNERS OF ENGLISH

Mohamed ZAHID
Institut de Phonétique, Université de la Sorbonne Nouvelle. CNRS, Paris
19, Rue des Bernardins, 75005, Paris, France.

ABSTRACT

The aim of this study is to determine to what extent Arabic learners of English are able to correctly realize the Voice Onset Time (VOT) of both Arabic and English /t/ when they learn English as a second language (L2) in adulthood. Acoustic measurements of VOT revealed that although the Arabic learners of English were able to detect the acoustic differences between Arabic and English /t/ and to produce /t/ with more aspiration in English than Arabic, they were unable to reach the phonetic norms of English /t/ because of equivalence classification.

1. INTRODUCTION

It is generally assumed that voicing contrasts are marked differently along the Voice Onset Time (VOT) dimension in languages such as English as opposed to French or Spanish. In English, voiceless stops /p,t,k/ are produced with a long-lag VOT accompanied by aspiration, whilst in Spanish and French, they are realized with a short-lag VOT. Arabic was chosen as the counterpart to English in this study because the phonetic contrast between voiced and voiceless stops in Arabic appears to differ from that of English and because Arabic lacks /p-b/ contrast [1], but not that of /t-d/ and /k-g/. These cross-language differences offered the opportunity to asses how a difference in phonological inventory as well as more subtle differences in the phonetic implementation of a phonological contrast would affect production of foreign language speech sounds by adult language learners. Moreover, cross-language studies [2, 3, 4, 5] that were interested in VOT production of similar consonants by second language (henceforth, L2) learners confirmed that while early learners were able to match VOT values of native speakers of English, late learners manifested VOT values that were intermediate to those observed in native speakers of their first language (henceforth, L1) and in native speakers of English. In light of these findings, the present study is designed to test the Speech Learning Model (henceforth, SLM) hypothesis [6, 7] regarding the production of similar L2 consonants. More specifically, we want to learn whether Arabic learners of English can accurately produce the VOT of /t/ at the beginning of English words.

2. METHODS

The speech material consisted of 10 CVC words in Moroccan Arabic and American English (C1= /t/, V= /a/ and C2= /b, d, p, t, k, q, h,ʕ,ʒ/ inserted in a carrier sentence "He said_ two times." Twenty-four subjects: 8 native Americans, 8 native Moroccans and 8 Arabic learners of English produced five repetitions of each CVC word. Acoustic measurements of VOT were made in Arabic words spoken by Arabic monolinguals, in English words spoken by American monolinguals, and in Arabic and English words spoken by Arabic learners of English. The speech of Arabic and English speaking subjects was examined to estimate the phonetic norms of Arabic and English. Audio recordings (Sony TCD 5M) were made in a soundbooth with a microphone (Nakamichi CM 300) placed about 6 in. from the mouth. The tape-recorded stimuli were digitized at 10 KHz with a 16-bits amplitude resolution and analyzed with the Unice speech analysis program [8]. The VOT of /t/ was measured from the beginning of the burst release (seen as a wide-frequency vertical striation) to the onset of periodicity in the region of the second and higher formants (seen as quasi-periodic striations) [9].

3. RESULTS

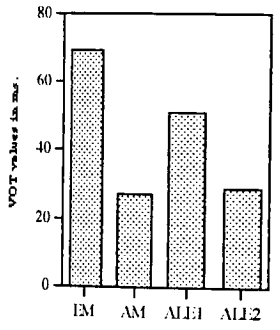


Figure 1. Mean VOT values for the three groups of subjects: EM(English Monolinguals),AM(ArabicMonolinguals) ALE1(Arabic Learners of English producing English stimuli), ALE2(Arabic Learners of English producing Arabic Stimuli).

The results presented in figure 1 indicate that:

- The mean VOT values for /t/ of the American monolinguals (69ms) is substantially longer (42ms) than the Arabic monolinguals (27ms). This implies that English /t/ is marked by the presence of an appreciable interval of aspiration after stop release. Arabic /t/, on the other hand, has less aspiration and thus short VOT [10] (see figure 2).
- The Arabic learners of English realized Arabic /t/ with a slightly longer VOT (29ms) than the Arabic monolinguals (27ms). However, a paired t-test analysis revealed no significant difference between the two groups ($t(39)= 0.69$; $p<0.494$, two-tailed). This result indicates that learning English as a second language has no significant effect on the production of the native language.

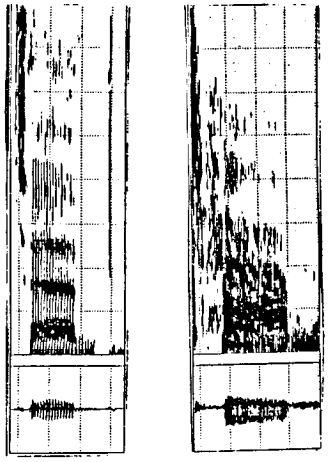


Figure 2. Spectrograms of the word /tab/ in Arabic (left) produced by Arabic monolinguals and in English (right) produced by American monolinguals.

- The mean VOT values of the Arabic learners of English vary as a function of the target language: a long-lag VOT (51ms) for English /t/ and a short-lag VOT (29ms) for Arabic /t/ (see figure 3).

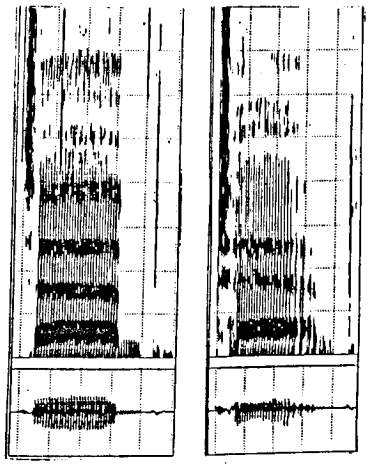


Figure 3. Spectrograms of the word /tab/ in Arabic (left) and in English (right) produced by Arabic learners of English.

- The mean VOT values for the English /t/ produced by the American monolinguals

are longer (69ms) than those produced by the Arabic learners of English (51ms). A paired t-test analysis revealed a significant difference between the two groups (t(39)= 17.07; p<0,001, two-tailed). This shows that the Arabic learners of English have not yet reached the phonetic norms of English. This finding is consistent with previous studies of L2 speech production [11, 12, 13].

The overall pattern of results obtained in the present experiment revealed that the Arabic learners of English produced /t/ with significantly longer VOT values in English than Arabic, but with significantly shorter VOT values in English words than the American monolinguals did. This suggests a pattern of partial phonetic approximation rather than of complete mastery of English by Arabic learners of English. A comparison of the three groups of subjects indicated that the Arabic learners of English realized the English /t/ with VOT values intermediate between those of the Arabic monolinguals and those of the Americans.

4. DISCUSSION

The discussion deals with the factors that led the Arabic learners of English to realize English /t/ with intermediate VOT values.

4.1. Age of learning English

The Arabic learners of English began learning English late (at 16 years of age) at a Moroccan University. Even though this late learning helped them to detect auditorily subtle acoustic differences between Arabic and English /t/ and to produce /t/ with more aspiration in English than Arabic, they failed to judge these two sounds as realisations of two different phonetic categories and to establish a new phonetic category for English [t^h] to produce it correctly. According to the SLM developed by Flege, this failure in phonetic category formation may be blocked by equivalence classification. This perceptual mechanism leads L2 learners into "equating" (identifying) an L2 sound with an auditorily distinct sound in the L1 inventory thereby rendering them unable to make effective use of sensory input in speech learning.

4.2. New phonetic category

Arabic learners of English in this study produced the /t/ in English words with VOT values that were intermediate between short-lag and long-lag values typically observed for Arabic and English, respectively. This means that they merged the phonetic properties of similar L1 and L2 phones within a new phonetic category [t^h] different from that of the American monolinguals. To realize /t/ accurately in English, Arabic learners of English must, in addition to the establishment of a new phonetic category, either modify the realisation rules used for outputting existing phonetic categories (i.e. [t]); or develop new realisation rules to be used when speaking English.

4.3. Phonetic Input

By phonetic input, we refer to the origin of the learning conveyed to our subjects. That is, who taught English to them?

It is likely that our Arabic subjects were exposed to English spoken by native speakers of Arabic (Moroccan) in which /t/ was realised with VOT values intermediate to the short-lag and long-lag values typifying Arabic and English, respectively. Thus, the subjects examined here may have produced the English /t/ with about the same intermediate VOT values they heard. Perhaps our non-native subjects would have produced better /t/ had they received better (i.e., more accurate) English-language phonetic input.

5. CONCLUSION

The main conclusions to be drawn from this study are as follows:

To improve English learning, it seems preferable for the Arabic learners of English to:

1- Begin learning English at an early age in order to acquire sufficient L2 experience and more phonetic input.

2- Spend more time in a country where English is the dominant language.

3- Receive sufficient native-speaker (American English) phonetic input.

4- Develop English phonetic realisation rules to lengthen the VOT of English /t/.

[1] A similar L2 phone is a phone that is related to a corresponding phone in the L1 yet differs acoustically from the L1 counterpart. E.g., /t/ is found in both French and English, but it is implemented as a short-lag stop with dental place of articulation in French and as a long-lag stop with alveolar place of articulation in English

ACKNOWLEDGEMENTS

We would like to thank N. Clements and J. Y. Dommergues for their comments and suggestions on a former manuscript, and all the subjects who participated in the recording sessions.

REFERENCES

[1] Al-Ani, S. (1970), *Arabic Phonology* (The Hague).

[2] Flege, J. E., and Port, R. (1981), "Cross-language phonetic interference: Arabic to English", *Language and Speech*, vol. 24, pp. 125-146.

[3] Port, R., and Mitleb, F. (1983), "Segmental features and implementation in acquisition of English by Arabic speakers", *Journal of Phonetics*, vol. 11, pp. 219-231.

[4] Flege, J. E., and Eefting, W. (1987a), "Cross-language switching in stop consonant production and perception by Dutch speakers of English", *Speech Communication*, vol. 6, pp. 185-202.

[5] Flege, J. E., and Eefting, W. (1986), "Linguistic and developmental effects on the production and perception of stop consonants", *Phonetica*, vol. 43, pp. 155-171.

[6] Flege, J. E. (1988), The production and perception of speech sounds in a foreign language, In *Human communication and its disorders: a review* 1988 (H. Winitz, editor). Norwood, NJ: Ablex.

[7] Flege, J. E. (1990), Perception and production: the relevance of phonetic input to L2 phonological learning, In *Crosscurrents in second language acquisition and linguistic theories* (C. Fergusson and T. Huebner, editors). Philadelphia: John Benjamins.

[8] LIMSI.

[9] Lisker, L. and Abramson, A. S. (1964), "A cross-language study of voicing in initial stops: Acoustical measurements", *Word*, vol. 20, pp. 384-422.

[10] Yeni-komshian, G. H., Caramazza, A. and Preston, M. S. (1977), " A study of voicing in Lebanese Arabic", *Journal of Phonetics*, vol. 5, pp. 35-48.

[11] Flege, J. E. and Hillenbrand, J. (1984), "Limits on pronunciation accuracy in adult foreign language speech production", *Journal of the Acoustical Society of America*, vol. 76, pp. 708-721.

[12] Nathan, G. (1987), " On second-language acquisition of voiced stops", *Journal of Phonetics*, vol. 15, pp. 313-322.

[13] Mack, M. (1989), " Consonant and vowel perception and production: Early English-French bilinguals and English monolinguals", *Perception and Psychophysics*, vol. 46(2), pp. 187-200.

INDEX

Speech materials used to elicit production in the Arabic and English stimuli.

Arabic stimuli

tah (wandering)
tab (repented)
taʕ(belong)
taʒ(crown)
taq (believed)

English stimuli

tab
tap
tack
tad
tat

HOW SUCCESSFULLY DOES VISUAL FEEDBACK TRAIN LISTENERS TO PRODUCE AND PERCEIVE NON-NATIVE PHONOLOGICAL CONTRASTS?

M. Ziolkowski and K. Landahl
University of Chicago, Chicago, IL, USA

ABSTRACT

This study evaluates the efficacy of three techniques for teaching students of a second language to phonologize natively allophonic variants. Each technique—"listen-repeat-compare" with a tape recorder, tutoring with a native speaker, and computer-driven visual feedback—was used to train ten English speakers in producing phonemic vowel and consonant length and pitch accent in Japanese. Though results were mixed, visual feedback proved its utility.

THE PROBLEM

A native speaker and instructor of the Japanese language issued a plea for help in her effort to assist her second-year students, native English speakers, overcome deficiencies in producing and perceiving properties of Japanese phonology with even rudimentary proficiency. Particularly salient were her students' difficulties with Japanese vowel and consonant quantity, and pitch accent.

All three of these phonological properties are phonemic in Japanese—*ojiisan* (long [i]) means 'grandfather', *ojisan* (short [i]) 'uncle'; *aka* (short consonant) denotes 'dirt', *akka* (long consonant) 'dirty money'; *kaME* (low-high pitch pattern) refers to a 'large-mouthed jar', *KAme* (high-low) to 'tortoise'. So, respecting these properties is crucial not just for achieving more native-like Japanese pronunciation and listening skills, but is a necessity for understanding Japanese and being understood correctly when speaking it.

What is interesting here from a linguistic point of view is that none of these Japanese phonological properties is wholly absent from the English learners' native system. Instead, each serves a more or less active role in English allophony. Vowel length has been observed to vary systematically with the voicing of a following consonant; consonants geminate when like segments meet at morpheme or word boundaries. Pitch may figure in the realization of English stress where it is one of the constellation of relevant features, but it bears the functional load solo in Japanese pitch accent.

The challenge for native English speakers learning Japanese appears to be one of changing the functionality of these properties. The challenge to the instructor is increasing the salience of these properties, hoping they may be put towards a phonemic end. We responded to the challenge with visual feedback.

THE SOLUTION?

We decided that a computer-based visual feedback system provided the appropriate means for increasing the salience of the acoustic cues to these properties, and selected the Visi-Pitch Model 6095/6097, manufactured by Kay Elemetrics of Lincoln Park, NJ. It affords real-time display of fundamental frequency and intensity contours for input utterances, and allows these data as well as their graphical representation to be stored on a personal computer for later analysis and display. Stored model utterances can then be redisplayed, and contours for student renditions overlaid on native models in real-time.

Figure 1 shows the fundamental frequency contours (f_0 traces) associated with a pair of words, uttered by a male native speaker, which contrasts all three Japanese properties. The upper half of the display traces the f_0 contour for *GAchcho* 'joint authorship', and the lower half, the trace for *gaCHOO* 'goose'.

Each of the phonological properties is iconically displayed in the figure, where time runs horizontally and f_0 vertically

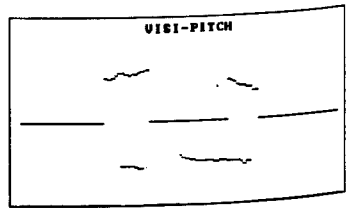


Figure 1. GAchcho *and* gaCHOO.

with greater displacements of the trace from the baseline associated with higher f_0. In both halves of the screen there is a break in the f_0 trace for the voiceless, medial consonant "ch", and the longer drop to the baseline in the upper half correlates with the longer "ch" in *GAchcho* compared to the short affricate in *gaCHOO*. Consonant quantity, then, is cued by the extent of baseline drops. The length contrast for [o] at the end of these two words is depicted by the lengthier last "chunk" of f_0 trace in the lower contour, and pitch accent patterns "high-low" versus "low-high" by the relatively higher vertical displacement of the first and second "chunks" of f_0 in the upper and lower halves of the screen, respectively.

In Figure 2, the Visi-Pitch's potential for training phonological properties one at a time becomes apparent. A female native speaker's f_0 traces for *itai* 'dead body' and *iitai* 'one body', a minimal pair distinguished by length of the word-initial vowels, appear in the two halves of the display. When the two contours are more similar in overall appearance, as they are here, the acoustic correlate of the feature differentiating the two words, vowel quantity, is more visibly transparent. The first "chunk" of the f_0 trace is clearly shorter in the upper half of the screen than in the lower—the first vowel in *itai* is shorter than the first vowel in *iitai*.

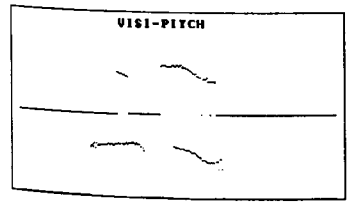


Figure 2. itai *and* iitai.

Thus, visual feedback exercises can be constructed offering learners straightforward, graphical cues to the phonological properties which have proven difficult for them to master. Visual feedback presents such properties much more saliently than do traditional strategies for teaching these distinctions. The only feedback offered, for example, by a Level III tape recorder may be compromised by students' probable deficiency in perceiving differences in the first place. Individual instruction with a native speaker of the language may be preferable, but the superior feedback associated with the greater cost typically is limited to a ready source of native examples and to the tutor's "affective" support for successful productions. Neither tape nor tutor affords salient presentation of the properties which cue non-native contrasts as well as visual feedback does.

But is visual feedback more effective than these other techniques? Does extracting relevant acoustic properties hone learners' abilities to perceive and produce non-native phonological properties? Common sense seems to dictate visual feedback's superiority to a tape recorder certainly, and to a tutor possibly, and pedagogues [1] have encouraged its use for fine-tuning phonetic distinctions in second language learning under the assumption that it works. Previous research (see [2]) has found increases in the perceived nativeness of utterances produced after visual feedback training on sentential intonation. But, to our knowledge, no systematic study of the extent of visual feedback's efficacy in fine-tuning phonetic capabilities or phonologizing the allophonic has ever been undertaken. Nor has visual feedback been rigorously compared to instruction with a tutor. The following study begins to fill these gaps.

THE STUDY

We employed the three teaching strategies discussed above—tape, tutor, visual feedback—in production-directed training on the three Japanese phonological properties mentioned there—vowel and consonant length, and pitch accent. Our goal was to quantify each strategy's effect on learners' abilities to produce these properties directly, through acoustic analysis of their productions. Also, we wondered whether production-directed training enhanced subjects' perception, and examined subjects' perceptual acuity with the distinctions before and after treatment.

Subjects

Thirty University of Chicago students participated in the study. Half were completing their first year of Japanese language study, and thus were familiar with the three phonological features to be trained; half were Japanese-naive. No subject had studied any other language that phonemically exploits segmental quantity, pitch accent or tone. The ex-

periment demanded about four hours over three weeks and subjects were compensated for their participation.

Stimuli

Minimal pair lists were constructed illustrative of each of the three phonological properties of interest. 9 pairs represented the vowel quantity contrast, 11 pairs constituted the list for consonant length, and 8 pairs appeared on the pitch accent list. Words were chosen to facilitate identifying segment boundaries on the Visi-Pitch in later acoustic analyses.

These 56 words were presented to 10 native speakers of Japanese for elicitation. The utterances of 2 speakers—one male, one female—were selected as models on the criterion that they were judged to be the best examples of the Tokyo dialect by the native speaker/second-year instructor.

Design

A third of the learners (5 Japanese-naive, 5 experienced) were randomly assigned to one of three treatment groups, each of which invoked a different training method (tape, tutor, visual feedback). Training lasted for one hour; twenty minutes spent on each of the three phonological properties of Japanese. After pretesting, subjects in the visual feedback condition received a twenty minute "crash course" in operation of the Visi-Pitch, which served as the visual feedback tool, before using it in training.

Subjects completed a battery of speaking and listening tests 1) immediately before training to measure baseline performance, 2) immediately after, 3) one week after, and 4) three weeks following instruction.

Within each testing session, subjects completed three different tasks, always administered in the following order. A "read-aloud" production test was followed by a "listen-and-repeat" mimicry task, and the session concluded with a forced-choice perception test. In the reading task, subjects were given a randomized list of stimulus words and asked to pronounce them "cold" to prompts by an experimenter at a rate determined by a light-flashing metronome. In the mimicry task, subjects were asked to repeat after a tape consisting of the male and female model speaker utterances presented in a randomized order. In the perception test subjects were played a tape with another randomization of the stimulus set and asked to circle the member of the minimal pair they thought each token represented. For example, *itai* appeared next to *iitai* on the response sheet, and subjects had to circle which word they thought the native speaker was saying.

A subset of the full stimulus set was used in the training and pre-treatment testing sessions. The stimulus sets expanded from 28 words (2 x 5 vowel length pairs, 2 x 5 consonant dyads, 2 x 4 pitch accent pairs) to 56 words from a subject's first session to last. We will determine, in later analysis, if training on a phonological property generalizes to items not explicitly trained upon.

Analysis

All subject data in production tests were acoustically analyzed for vowel and consonant length, and pitch accent as relevant. (Due to measurement difficulties we excluded data from three subjects: one Japanese tape, one naive tape, one naive tutor.) For quantity, the absolute segment durations were determined and used to compute the dependent measures for quantity described below. For pitch accent, analysis was restricted to that subset of stimuli for which the Visi-Pitch allowed word segmentation into syllables. In these cases the average f_0 for the two syllables "bearing" the pitch accent (high-low or low-high) was measured. Since each word must be self-normalizing (native speakers can determine the accent pattern of words spoken in isolation), the relevant value for the dependent measure for pitch accent was the difference in average f_0 between the two syllables, divided by whichever average was smaller. (A two-way ANOVA on the model speakers utterances showed a main effect of accent-type but not speaker identity for this metric.)

The dependent measure of subject performance in the reading and mimicry production tasks was different for the quantity and pitch accent tokens. For each quantity distinction, a regression equation was fit for each subject in each test session for reading and for mimicry. The equation consisted of μ, the average speaking rate, plus α, quantifying how much longer long segments are and how much shorter short segments are than μ, plus an error term, ε. α was the dependent measure for both vowel and consonant length performance. For pitch accent the dependent measure was the difference between the average values of the metric described above for the two accent types.

For the perception tests the percentage correct was the dependent measure.

Results

Repeated-measures analyses of variance were run for each of the production and perception measures with language experience (Japanese or naive) and training type (tape, tutor, visual feedback) as grouping factors.

For analysis of subjects' performance on the three phonological properties, exercise type (reading, mimicry) was an additional grouping factor in the ANOVA. Session (pre-, post-, week 1, week 3) was the repeated measure.

For vowel duration three significant values obtained: a main effect for session ($Pr>F=0.0001$), a main effect for exercise type, ($Pr>F=0.0376$), and an interaction between exercise type, language experience and training type ($Pr>F=0.0791$). For consonants there was a main effect for session ($Pr>F=0.0001$), an interaction of exercise and session ($Pr>F=0.0351$), and a marginal interaction of exercise, session and treatment type ($Pr>F=0.0228$). For pitch accent main effects of exercise type and session were observed ($Pr>F=0.0001$ for both).

For the perception data, baseline performance was a covariate and the last three sessions were the repeated measures for a two-way ANOVA with language experience and training type as grouping factors. The analysis revealed no statistically significant results, although a main effect of training type did approach significance ($Pr>F=0.088$).

Discussion

The results for production of the contrasts have been discussed in greater detail elsewhere [3], but those of relevance for evaluation of the efficacy of visual feedback in comparison to tape and tutor training derive from the two interactions in which training type is implicated. Although robust effects for training type were not obtained, it is not surprising they were not, given the small number of subjects in each of the groups and the extent of individual performance variation we have observed in the data.

Our interpretation of the interactions obtained for vowel and consonant quantity performance is the following. For all subjects, mimicry generally improves to native-like following training, but those trained using the tape recorder tend to exaggerate the length distinctions when reading. Visual feedback subjects, then, more closely resemble the tutored subjects by showing more or less native-like durations when reading without a native example.

The perception results are intriguing. Remember that the main effect for training type approached statistical significance for the perception data as a whole. A robust finding for perceptual acuity is that naive visual feedback subjects and Japanese-experienced tutored subjects perform best, followed by naive tape and Japanese visual feedback subjects with Japanese tape and naive tutored learners performing worst. This generalization holds for perceptual ability as a whole, as well as for perception of each phonological feature individually. We can think of no principled reason why this strict ordinal pair-wise ranking should have obtained. We suspect the best explanation to be "luck of the draw"—the naive visual feedback condition happened to be populated by subjects with exceptional perceptual skills.

ACKNOWLEDGEMENT

Thanks go to Mayumi Usami, Brenda Tunnock, Sema Yetim, the Consortium for Language Teaching and Learning, and the University of Chicago Language Laboratories and Archives.

REFERENCES

[1] Chun, D. (1989), "Teaching tone and intonation with microcomputers", *CALICO Journal*, vol. 7, pp. 21-46.

[2] de Bot, K. (1983), "Visual feedback of intonation I: Effectiveness and induced practice behavior", *Language and Speech*, vol. 26, pp. 331-350.

[3] Landahl, K., M. Ziolkowski, M. Usami, and B. Tunnock (1992), "Interactive articulation: Improving accent through visual feedback", *Proceedings of the Second International Conference on Foreign Language Education and Technology*, pp. 283-292.

ASSIMILATION OF IRISH VELARISED & PALATALISED STOPS

Ailbhe Ní Chasaide and Liam Fitzpatrick
Centre for Language and Communication Studies, Trinity College, Dublin, Ireland

ABSTRACT

Assimilatory patterns for [t̪ʲt̪ʲ c k] in Irish were examined for one speaker in $VC_1\#C_2V$, using EPG and limited EMA data. All assimilations were anticipatory. Palatalised and palatal stops were much more prone to assimilation than velarised and velar stops. To this extent, coronals do not all assimilate more readily than dorsals. Many effects may be explained in terms of mechanical and dynamic lingual constraints. Segments involving contiguous articulators assimilate more readily, and follow a different assimilation route than the non-contiguous. Fewer assimilations occur when the tongue gesture required for the cluster follows the "preferred" anti-clockwise trajectory.

INTRODUCTION

The four way opposition of lingual stops in Irish involves differences in secondary articulation (palatalisation and velarisation) as well as differences in the primary place of articulation. As such, they offer a rich testing ground for theories of assimilation. As part of a broader study, we report here on the assimilations that occur when these stops form clusters across a word boundary.

METHODS

The assimilation patterns for every combination of the four lingual stops of Irish /t t' k' k/ = [t̪ʲt̪ʲ c k] across a word boundary were examined for one female speaker of Connemara Irish in $'VC_1\#'C_2V$ utterances where V = /a/. This yielded four homorganic clusters as well as twelve non-homorganic clusters. Each stop was also elicited in #CV and VC# contexts. Five randomised repetitions of all utterances were recorded using the Reading EPG system and audio. The EMA illustrations below are drawn from a separate recording which included EMA, EPG and audio (same materials, same speaker, 10 repetitions). The four lingual EMA coils were positioned: (1) at 0.5 cm behind the tip; (2), (3) and (4) respectively at about EPG row 5, row 7 and 0.3 cm behind the EPG palate (during a swallow).

As all assimilations were anticipatory, the description of the unassimilated stops below is based on the VC# context. However, to decide on the extent to which a given non-homorganic cluster (for example, /t#k'/) had assimilated, the temporally more comparable homoranic clusters (/t#t/ and /k'#k'/) were used as a basis for comparison. We focused on the dynamic patterning in the interval from 40 ms before closure of C_1 to 25 ms after. A "unique" contact template was calculated for each EPG frame over this interval, describing that part of the contact pattern which was found in any repetition of /t#t/, but never in any repetition of /k'#k'/. Similarly, we established on a frame-by-frame basis the "unique" pattern for /k'#k'/ never present in /t#t/. This interval for the non-homorganic /t#k'/ was then compared to the "unique" templates. The degree of assimilation corresponded to the extent to which it approached the "unique" contact pattern of C_2 and differed from that of C_1. This process was carried out for every combination of C_1 and C_2.

RESULTS

The unassimilated stops

Figure 1 presents EPG contact patterns for the unassimilated stops in VC# as well as the approximate tongue contours, at the timepoints of consonant closure and maximum EPG contact. The tongue contour outlines show where midsagittal contact occurs, as well as approximate tongue height in the regions corresponding to the EPG rows 5-8.

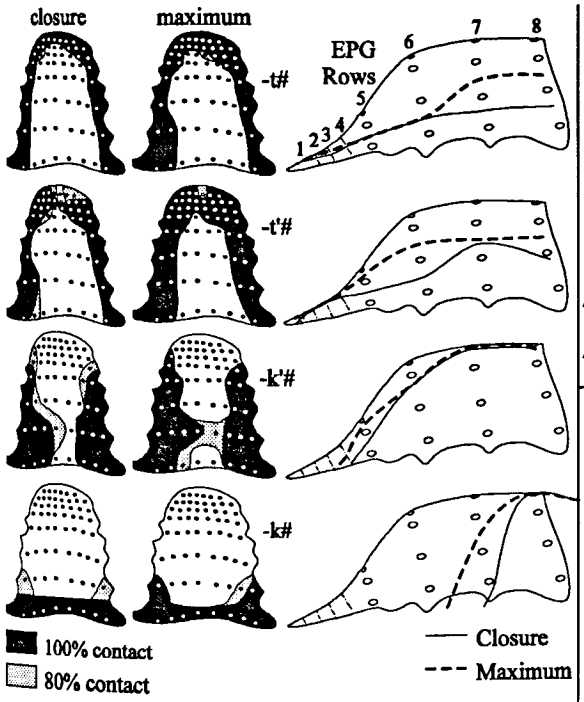


Figure 1. Averaged EPG contact patterns and approximate tongue contours for stops in VC#.

For coronal stops /t/ and /t'/, despite considerable overlap in the area of contact at the maximum time point, the dynamics at onset and offset suggest two rather different articulations. The initial and final point of contact for /t/ is in the dental region (row 1). For /t'/ the contact pattern at closure and release suggests a more posterior, laminar articulation with lowered tongue-tip (see discussion of similar patterns in [1]). The different relative positions of tip and blade for the two stops at closure can be visualised in Figure 2.

As concerns secondary articulation, the greater fronting of the tongue body for /t'/ is evident from the tongue contours of Figure 1.

For the dorsal stops, /k/ and /k'/, one cannot easily differentiate primary and secondary aspects of the articulation. The main stricture for /k'/ is formed with a raised tongue front and a more anterior point of contact than for /k/. There was some variability in the precise location of the occlusion for /k'/, and this gives the impression in Figure 1 of incomplete closure in row 7 of the EPG palate. For /k/, much of the contact is likely to be behind the artificial palate.

Assimilating Environments

On the whole, these data did not reveal large numbers of assimilations. Under what we term assimilation (we deal mainly here with the primary aspects

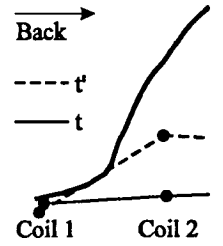


Figure 2. Average positions of EMA coils (1) and (2) for /t/ and /t'/ at closure.

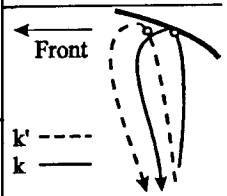


Figure 3. Trajectory of EMA coil (4) for /k'/ and /k/. The small circle marks the coil position at the time of closure.

of the articulation), we observed rather different kinds of phenomena. Results in Table 1, are presented in terms of different types of assimilation and some of these are illustrated in Figure 4, which shows four repetitions of the -t'#k- cluster. For each repetition, EPG contact on the two rows best representing the primary articulatory gestures for /t'/ and /k/ are plotted over the interval of both stops. This was always row 8 for /k/, and varied as between rows 2 and 3 for /t'/. Closure and release of both stops are shown.

None: no/little evidence of assimilation (repetition 1 in Figure 4)

Len: a partial assimilation where an appropriate gesture for the first consonant of the cluster is retained, but the stricture never becomes a stop. **Len1** = fricative-like strictures held for a relatively long interval (e.g. repetition 2, Figure 4); **Len2** = such strictures when of very short duration and/or of a lesser degree of narrowing (e.g. repetition 3, Figure 4).

Blend: partial assimilations where the first part of the (potential) cluster has features of C_1 and C_2, but doesn't necessarily include all the features of either.

Full: no/little EPG evidence of C_1 features (repetition 4 in Figure 4).

In terms of resistance to assimilation (based on the numbers in the None column), we get the following hierarchy: **k > t > k' > t'**. The palatalised stops emerge as the "weakest". Note that a rather analoguous phenomenon has been discussed for Russian, where palatalisation is tending to disappear in segments which precede velarised consonants in word-internal clusters [2].

For each of the stops, the likelihood of assimilation appears to be greatest when it abuts a stop whose primary articulation involves a contiguous, mechanically linked part of the tongue (e.g. tip and blade). Thus, abutting coronals or abutting dorsals are more prone to assimilation than clusters which involve a combination of coronal and dorsal.

When C_1 and C_2 involve contiguous articulators, the assimilation route seems to involve a blending of the C_1 and C_2 gestures in the first part of the cluster. When the consonants involve non-contiguous articulators, the route assimilation takes would seem to involve a gradual lenition (reduction in the amplitude of the gesture) of C_1. Generally, the C_1 gesture in the more assimilated cases has a shorter duration and tends to overlap more with the C_2 gesture. These findings are broadly compatible with predictions of articulatory phonology [3].

Table 1. Number and type of assimilations found for all clusters.

	None	Len1	Len2	Blend	Full
-t#t'-	2			3	
-t#k'-	5				
-t#k-	5				
-t'#t-		1		1	3
-t'#k'-	2		3		
-t'#k-	1	1	2		1
-k'#t-	4	1			
-k'#t'-	3	2			
-k'#k-	1			2	2
-k#t-	5				
-k#t'-	5				
-k#k'-	5				

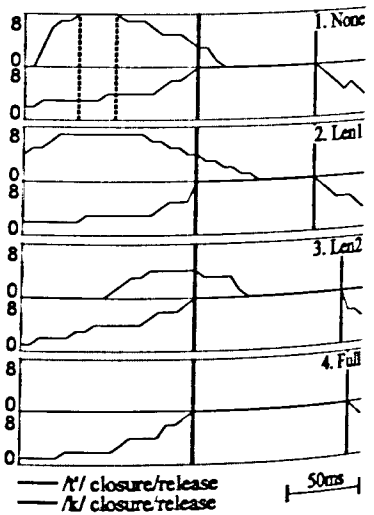


Figure 4. For 4 repetitions of -t'#k- are shown EPG activation (0-8 electrodes) in those rows representing the region of C_1 and C_2 primary articulation, aligned to C_2 closure.

It is also worth noting that the two coronal stops are extremely different in terms of their propensity to assimilation. Whereas /t'/ bears out the frequently observed high propensity to assimilation, /t/ emerges as very resistant, assimilating much less than the palatal /k'/.

It is striking that when either coronals or dorsals abut, the order of occurrence appears to matter greatly. Note that no assimilations were found for -k#k'-, whereas for -k'#k- there were 4 assimilations (in the 5 repetitions).

This asymmetrical behaviour that results from the juxtaposition of two dorsals might possibly be explained in terms of tongue dynamics. As can be seen in Figure 3, the preferred tongue body gesture appears to follow an anti-clockwise elliptical gesture (see discussion of this in [4]).

We suggest that this basic gesture may need little modification to produce -k#k'-: it might simply be a matter of allowing the tongue forward movement to be more extensive. However, the articulation of -k'#k- may require movement that runs counter to the preferred tongue body gesture. One should note that even in the (potentially) assimilating environments of -k'#k- we found no examples of a clockwise gesture.

This may explain why palatalised stops generally assimilate more readily before velarised stops than vice-versa. It would also explain the clearly similar phenomena discussed for Russian by Barry [2].

Finally, this same principle may also explain the somewhat greater propensity to assimilation in /t'#k'/ than in /k'#t'/.

CONCLUSIONS

The different propensities to assimilation of the consonants may well have their origin in mechanical and dynamic properties of the tongue. The tendency to assimilation is heightened where C_1 employs a contiguous articulator to that of C_2. When assimilated, C_1 here "blends" features of C_1 and C_2. Where C_1 and C_2 involve non-contiguous articulators, C_1 assimilated through a gradual weakening of the gesture.

Palatalised consonants are much more prone to assimilation than are the velarised. It is hypothesised that this asymmetry arises because tongue-body gestures in an anti-clockwise direction are preferred to clockwise gestures. So, for example, a sequence of velar + palatal may represent an easy target, less likely to assimilate, than the sequence palatal + velar.

ACKNOWLEDGEMENTS

This work was supported by Esprit-BRA, Working Group, ACCOR II.

REFERENCES

[1] Pandelli, H. (1993), "The articulation of lingual consonants", *Unpublished Ph.D. thesis*. Cambridge University

[2] Barry, M. (1992), "Palatalisation, assimilation and gestural weakening in connected speech", *Speech Communication*, 11, pp. 393-400

[3] Browman, C. and Goldstein, L. (1989), "Articulatory gestures as phonological units", *Phonology*, 6, pp. 201-231.

[4] Mooshammer, C., Hoole, P., and Künhert, B. (1995), "On loops", *Journal of Phonetics*, 23, in press.

THE EFFECTS OF PROSODY ON VOICELESS STOP-RELATED GESTURES IN ENGLISH

André M. Cooper
The University of Michigan, Ann Arbor, Michigan USA

ABSTRACT

This study examines the contextual variation in a number of indices of aspiration that appear in the literature in order to determine whether the variation in these indices can be accounted by phonological rules based on syllable structure. Results suggest that effects of both stress and word structure, rather than syllable-based category changing rules, best describe the contextual variation in the indices.

1. INTRODUCTION

In English voiceless stops vary between aspirated and unaspirated. Aspirated stops are said to occur word initially, and syllable initially in stressed syllables. Although phonologists have devised category changing rules to describe the distribution of these voiceless stop allophones (e.g., [1], [2], [3]), it unclear exactly what aspect of a voiceless stop's production corresponds to aspiration. If the category changing rules capture facts about speakers' productions of voiceless stops, there should be a corresponding physical index of aspiration (IOA) that varies as a function of syllable structure.

The goal of the present study is to examine a number of possible IOAs that appear in the literature, in order to test which, if any, is categorically distributed. An additional goal is to test whether the contextual variation in the indices is best accounted for by a categorial rule (aspiration-category proposal), or by continuous effects of prosody on oral and laryngeal gestures and on the way these gestures are coordinated in time and space (prosody proposal). The IOAs include: (a) voice onset time--the interval from the release of a stop and the onset of glottal pulsing (e.g., [4], [5]), (b) peak glottal magnitude--the maximum degree to which the glottis opens for devoicing, (c) glottal magnitude at oral release--the degree to which the glottis is open at the time of oral release [6], (d) the timing of the onset of glottal adduction relative to the oral release of a stop [7]. Longer VOTs, larger glottal openings and later onsets of glottal adduction relative to oral release are all associated with greater amounts of aspiration.

2. EXPERIMENT 1

2.1. Methods

Two male speakers of English, ES and KM, spoke the nonsense words /pipip/, tɪtɪt, kɪkɪk/ with primary stress on either the initial or the final syllable in the carrier phrase "say___again." Both acoustic (as a measure of oral closure and release) and transillumination (as a measure of glottal activity) signals were recorded synchronously.

2.2. Data Analysis

For each speaker separate ANOVAS were performed comparing the effects of stress and word position on each IOA for each stop category. When there were significant interactions, separate protected t-tests were performed for each word position to determine whether there was an effect of stress on the IOAs. Approximately twenty four repetitions of utterance are included in the analyses. The aspiration-category proposal predicts that there should be a significant interaction between the effects of stress and word position such that there is no difference between word-initial IOAs and a significant difference between medial IOAs. The prosody proposal predicts there should not be an interaction between stress and word position on the IOAs. Finally, if there were both a significant interaction and a significant stress effect on IOAs in both initial and medial positions, this result would be ambiguous between the aspiration-category and the prosody proposals. Probabilities less than .01 are considered significant.

2.3. Results

2.3.1. Acoustic Data

The VOT data for ES lend strong support for the aspiration-category proposal. There is a significant interaction between stress and word position. In addition, there is no significant difference between prestressed and preunstressed VOT in initial position and a significant difference in medial position.

The data for KM are not as uniform. The data for /t/ support the prosody proposal, while the data for /p, k/ are ambiguous between the two proposals. For /p, k/ there is a significant interaction of stress and word position on VOT, as the aspiration-category predicts, but there are also significant stress effects in word-initial and medial positions.

2.3.2. Transillumination Data

For both speakers the peak glottal magnitude data for the labial stops are ambiguous between the aspiration-category and the prosody proposals, while the data for the lingual stops support the prosody proposal.

The pattern of results for glottal magnitude at oral release are virtually identical to those for the peak glottal magnitude. Peak glottal magnitude, however, is greater than the glottal magnitude at oral release.

2.3.3. Interarticulator Timing

There are no consistent patterns of oral-laryngeal timing across speakers or across stop categories within speakers. Thus, it does not appear that the onset of glottal adduction always occurs later for stops which are supposed to aspirate than for stops which are supposed to be unaspirated.

2.4. Summary and Discussion

Both the aspiration-category and the prosody proposals find some support. Surprisingly, there was no single IOA, whose distribution (across speakers and stop categories) fit either proposal exactly. These finding suggests that aspiration, however defined, is neither solely a function of syllable structure, nor solely a function of prosody.

An alternative interpretation of these results is that both stress and word position (prosody + word position proposal), rather than syllable structure, per se, affect the contextual variation in the IOAs. This interpretation could subsume each of the previous proposals. Specifically, IOAs in word-initial position might simply hold a unique status.

Consider the fact that in many dialects of English, word-initial stops are distinguished physically by differences in voicing lags, rather than by voicing lead versus voicing lag. Thus, word-initial /b, d, g/ are realized physically as [p, t, k] and word-initial /p, t, k/ are always realized as [p^h, t^h, k^h] in order to be differentiated from their voiced counterparts. In word-initial position the articulators may work to minimize stress-related differences in IOAs so that voiced and voiceless stops will not be confused perceptually.

Finally, these data suggest a difference in the way that the glottal IOAs for labial and lingual stops behave. This difference may serve to enhance the perceptional distinction between voiced and voiceless labial stops. All else being equal, voicing will begin earlier for labial stops than for lingual stops because it takes a longer time to build a pressure differential across the glottis (a physical condition necessary for voicing to occur) for smaller cavities than for larger ones. Consequently, in order to insure an a voicing lag sufficient for bilabial stops to be perceived as voiceless, more extreme laryngeal and oral-laryngeal timing maneuvers may be required for labial stops than for lingual stops.

3. EXPERIMENT 2

Experiment 2 seeks to distinguish between the aspiration-category, the prosody, and the prosody + word-position proposals by examining voiceless stops in additional segmental contexts. Two types of stimuli are used. For both stimulus types voiceless stops appear in contexts where aspiration categories should not vary as a function of syllable structure. The effects of stress on IOAs for these stimuli are then compared with the data in Experiment 1.

According to the aspiration-category proposal stress is predicted to have a large effect on IOAs only when it produces a change in syllable structure. For the present stimuli the aspiration-category proposal would find further support if stress effects on the IOAs were small, and comparable to those for word-initial singletons in Experiment 1

that supported the aspiration-category proposal.

If, however, stress effects on these IOAs are large and comparable to those for medial singletons in Experiment 1, it can be argued that stress has similar effects on IOAs regardless of the intended aspiration categories (prosody proposal), with the caveat that stress can affect word-initial IOAs differently than it affects non-word-initial IOAs (prosody + word-position proposal).

3.1. Methods

The two males from Experiment 1 spoke the nonsense words /pɪspɪp, pɪpsɪp, pɪsɪp, pɪptɪp, pɪtpɪp/ in the carrier phrase "say___again." KM did not produce any of the final stops in the target words or the word /pɪtpɪ/. Primary stress occurred on either the initial or the final syllable. Again, both acoustic and transillumination data were collected synchronously.

3.2. Data Analysis

First two-way ANOVAs were performed to investigate how the combined effects of stress and utterance type affect each IOA for the present stimuli. Then ANOVAs were performed to examine the effects of stress and utterance type on the stop-stop clusters vs. the initial singleton stops from Experiment 1. Finally, ANOVAs IOAs were performed to examine the effects of stress and utterance type on the stop-stop clusters vs. the medial singleton stops. Where there were interactions, additional ANOVAs were performed to determine their source. Probabilities less than .01 are considered significant.

3.3. Results

3.3.1. Acoustic Data

The VOT data for stop-stop clusters support the prosody + word position proposal. For both speakers stress has significant and equivalent effects on VOT for /pt, tp/. Furthermore, the magnitude of stress effects on VOT for stop-stop clusters, for which stress should not produce shifts in aspiration categories, is comparable to that for singletons where stress is predicted to affect aspiration categories. Stress did not have a significant effect on VOT for /sp/ for either speaker. It was not appropriate to measure VOT for the other utterances with fricatives.

3.3.2. Transillumination Data

Like the VOT data, the present results provide support for the stress + word position proposal. For both speakers stress effects on peak glottal magnitude for stop-stop clusters are comparable to those for singletons in whatever position stress effects are greatest.

For both speakers stress effects on peak glottal magnitude differ for stop-stop clusters versus the utterances with fricatives. Stress affects all of the utterances with fricatives in a uniform fashion and there is a significant main effect of stress.

4. General discussion and conclusions

The results for the stop-stop clusters directly support the prosody proposal and implicitly support the prosody + word position proposal. In particular, stress effects on the IOAs for stop-stop clusters are most similar to those for singletons which vary between aspirated and unaspirated.

The stop-stop cluster data also have implications for the VOT results in Experiment 1. Recall that these data appeared to offer support for the aspiration-category proposal. The present results, however, indicate that the contextual variation in the IOAs cannot be explained by a need to produce aspirated allophones in certain phonological environment and unaspirated allophones in others. Rather, it appears that stress has large effects on nonword-initial IOAs--regardless of whether stress differences cause changes in syllable structure--and potentially smaller effects on word-initial IOAs.

The stops in /s/ stop clusters do not aspirate. Therefore, the IOAs are predicted to exhibit little or no stress effect according to the aspiration-category proposal. According to the prosody + word position proposal, however, these IOAs might be expected to show relatively large stress effects since the stops are not word initial. Although the peak glottal magnitude shows significant stress effects, the VOT results show no stress effect for /sp/, suggesting that the lack of stress-related variation in VOT might be most economically described by an aspiration rule that makes reference to syllable structure.

Browman and Goldstein [8] explain the defective distribution of stops in /s/-stop clusters as a constraint on the articulatory structure of English words. They propose that English words can begin with only one devoicing gesture. Thus, the constituents of /s/-stop clusters share a single devoicing gesture rather than each having its own. The lack of voicing and aspiration for the stop is then explained by a general principle governing gestural coordination in English. The devoicing gesture begins at the onset of the fricative and ends at the release of the following stop, thus, generating a voiceless unaspirated stop.

Lisker [9] states that the problem of the status of aspiration in voiceless stops following /s/ is one of English orthography rather than contextual variants. Unlike some environments in which the phonological status of aspiration can be determined paradigmatically, (e.g., the labial stops in "rapid" and "rabid" which contrast in voicing or the stops in "bin" and "pin" which contrasts in aspiration), the status of stops in /s/-stop clusters cannot be distinctively contrasted and are, therefore, phonologically ambiguous. Thus, it is just as plausible to attribute the lack of aspiration and voicing of stops in word-initial /s/-stop clusters to a devoicing rule since /b, d, g/, which never aspirate, are generally voiceless following any voiceless obstruent.

In summary, no physical IOA was found whose contextual variation could be described as a function of syllable structure. Instead, the physical realization of nonword-initial IOAs may be predictable largely as a function of stress. The physical realization of word-initial stops, however, form a special case. The unique status of word-initial sounds is not idiosyncratic to the present data. Indeed, word boundaries are important junctures in speech and consonants and vowels generally have been found to behave differently in word-initial versus non-word initial position (e.g., [10], [11]).

REFERENCES

[1] Halle, M., & Stevens, K. N. (1971). A note on laryngeal features. *Quarterly Progress Report, Research Laboratory of Electronics* (Massachusetts Institute of Technology), 101,198-213.

[2] Kahn, D. (1976). *Syllable-based generalizations in English phonology.* Bloomington, IN: IULC.

[3] Kiparsky, P. (1979) Metrical structure assignment is cyclic. *Linguistic Inquiry*, vol. 10, no. 3, 421-441.

[4] Lisker, L. & Abramson, A. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, 20, 384-422.

[5] Keating, P. A. (1984). A phonetic and phonological representation of stop consonant voicing. *Language*, **60**, 286-319.

[6] Kim, C. W. (1970). A theory of aspiration. *Phonetica*, **21**, 107-116.

[7] Löfqvist, A., & Yoshioka, H. (1984). Intrasegmental timing: Laryngeal-oral coordination in voiceless consonant production. *Speech Communication*, 3, 279-289.

[8] Browman, C. P.,& Goldstein, L. (1986). Towards an articulatory phonology. *Phonology Yearbook*, 3, 219-252.

[9] Lisker, L. (1984). How is the aspiration of English /p, t, k/ "predictable"? *Language and Speech*, 27, 391-394.

[10] Oller, D. K. (1973). The effect of position in utterance on speech segment duration in English. *Journal of the Acoustical Society of America*, **54**, 1235-1247.

[11] Klatt, D. H. (1975). Voice onset time, frication, and aspiration in word-initial consonant clusters. *Journal of Speech and Hearing Research*, **18**, 686-705.

Work supported by NIH Grant DC 00232 to Haskins Laboratories and by the University of Michigan.

AN INVESTIGATION OF THE ACOUSTIC CHARACTERISTICS OF THE PARANASAL CAVITIES

Jianwu DANG and Kiyoshi HONDA

ATR Human Information Processing Res. Labs.,
2-2 Hikaridai Seikacho Sorakugun Kyoto, 619-02 Japan.
dan@hip.atr.co.jp

ABSTRACT

Acoustic characteristics of the paranasal cavities were studied by a direct measurement approach. Our method evaluates the transmission functions of the nasal tracts based on the sound pressure gradient in the tract and the sound pressure at the nostrils. The results from three male subjects show that frequency ranges of the zeros caused by the paranasal cavities are from 310 to 1070 Hz for the sphenoidal sinus, 310 to 950 Hz for the maxillary sinus, and 600 to 1580 Hz for the frontal sinus. The zeros are expected to affect the shaping of nasal formants due to their stability in the low frequency region.

INTRODUCTION

Early studies of the nasal cavity have suggested that the acoustic effect of the "subsidiary cavities" within the nose play an important role in shaping appropriate nasal spectra. For directly measuring acoustical contributions of the paranasal cavities, Lindqvist-Gauffin et al. (1976) used a probe tube sound source to excite the nasal cavity while the velum was closed [1]. They found pole-zero pairs caused by the paranasal sinuses in their results by changing locations of the sound source. Takeuchi et al. (1977) estimated the volume of the paranasal sinuses and their ostia from a cadaver specimen, and observed the resonance properties of the nasal cavity when the sinuses were taken into account[2]. However, some conclusions reported in the previous studies are quite different. For this reason, more accurate observations are required for investigating the acoustic properties of the paranasal cavities. In the present study, we used a method to directly measure acoustic characteristics of the nasal and paranasal cavities for three subjects.

METHOD

Theoretical considerations

The vocal tract can be computationally represented by a cascade concatenation of small sections when the discussion is limited in the low frequency region below 4 kHz. The characteristics of the sound propagation in such a tract are easily described by drawing upon elementary electrical theory and some well-known results for one-dimension waves on transmission lines. For i'th section of l_i in length, with sending-end sound pressure P_{i-1} and volume velocity U_{i-1} , the receiving-end sound pressure and velocity P_i and U_i are given by

$$\begin{bmatrix} P_i \\ U_i \end{bmatrix} = \begin{bmatrix} \cosh\gamma_i l_i & -Z_i \sinh\gamma_i l_i \\ Y_i \sinh\gamma_i l_i & \cosh\gamma_i l_i \end{bmatrix} \begin{bmatrix} P_{i-1} \\ U_{i-1} \end{bmatrix} \quad (1)$$

where γ_i is the propagation constant depending on the length and cross-sectional area of the section. Z_i and Y_i denote the characteristic impedance and admittance of the section, respectively[3]. For a portion of the vocal tract from Section i to the radiation end, the relationship is

$$\begin{bmatrix} P_r \\ U_r \end{bmatrix} = \begin{bmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{bmatrix} \begin{bmatrix} P_{i-1} \\ U_{i-1} \end{bmatrix} \quad (2)$$

where P_r and U_r are sound pressure and volume velocity at the radiation end. The matrix of 2 × 2 is the transformation matrix whose elements depend on the geometry of the portion. Solving the ratio of the input volume velocity to the output velocity, the transmission characteristics $T(\omega)$ from Section i to the radiation end are given by

$$T(\omega)=U_r/U_{i-1}=t_{21}z_{i-1}+t_{22} \quad (3)$$

where z_{i-1} denotes the input impedance seen from Section i to the radiation end. It is seen that all of the terms on the right side of Eq. (3) are dependent only on the geometry of the portion from Section i to the radiation end. In other words, the transmission characteristics obtained from Eq. (3) are theoretically independent of the portion behind Section i, and U_{i-1} can be looked upon as a continuous current source to the portion.

An experimental study was conducted to assess accuracy of the method [4]. The results showed that this method yielded accuracy of about 4% error, a ratio of difference between measurement and theoretical values to the theoretical value, for the peaks, and 2% for the zeros of acoustic tubes of known geometry. The locations of the branches within acoustic tubes were measured as well as the frequency properties.

Experimental Procedure

In this study, we investigate acoustic characteristics of the paranasal cavities using the method described above. Figure 1 shows a schematic diagram. The external microphone M3, a B&K-4003, was set about 15 cm in front of the mouth of subjects. Two B&K-4128 probe microphones (M1, M2) were attached to each other to form parallel tubes, and were used for measuring sound pressures within the tract via two flexible tubes. The flexible tubes have an identical length of 30 cm and a matched impedance to the microphones. Outer diameter of the tubes was 0.165 cm, and inner diameter was 0.076 cm. Tip distance of the tubes was adjusted to 0.6 cm. A vinyl ball with a 0.6-cm diameter marked by B in Fig. 1 was used to keep the tips out of touch with inside surface of the nasal cavity.

Three subjects from 30 to 44 years-old participated in this experiment. The nasal tract of the subjects was treated with adrenaline (naphazoline HCl, 0.05%) in order to decongest the nasal mucous membrane. The paired flexible tubes were inserted through the nasal floor of one nasal passage into the nasopharynx about 8 cm from the nostrils. The other nasal passage was collapsed at the nostril by a finger to ensure that there was only one radiation orifice during the measurement. Measurements were taken at 0.5-cm intervals in the portion of the cavity 4 to 8 cm behind the nostrils. Both nasal passages were measured in the same experimental conditions.

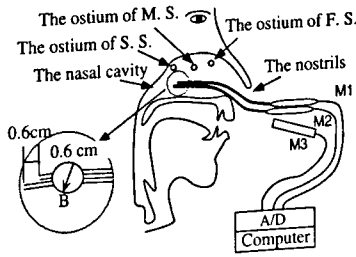


Figure 1. *Diagram of setup for measuring the acoustic characteristics of the paranasal sinuses.*

Speech materials used in this experiment were ten Japanese NV syllables and two nasal consonants. Subjects were asked to utter the speech materials twice, and instructed to prolong the utterance of the nasal consonant in the NV syllables to some extent. Measurements were made in an anechoic room. Sampling rate was 44.1 kHz, and the cut-off frequency was 3 kHz for signal analysis.

Data from a stable segment of nasal consonants were used for analysis. FFT

Table I Locations of the openings of the paranasal cavities from the nostrils. (cm)

Subjects	S. S.	M. S.	F. S.
Sub.1	6.2	5.1 (4.8)*	4.3
Sub.2	6.8	4.3	4.0
Sub.3	6.0	4.5	4.0

**The digit within the parentheses is the opening position on right side.*

with a 4096-point hamming window was applied to the selected segment, and shifted in a 1024-point interval. Frequency properties of the segment were obtained by averaging the results from each frame. The transmission characteristics of each measurement position were the average value of about 20 sound recordings.

Morphology of the nasal cavity

Morphologically, the paranasal cavities consist of four kinds of sinuses: the sphenoidal sinus (S. S.), the maxillary sinus (M. S.), the frontal sinus (F. S.) and the ethomoidal sinus (E. S.). This analysis was focused on the sinuses except for the ethomoidal sinus because the sinus has less effect on the low frequency region below 3 kHz. The locations of the ostium opening for the three sinuses are listed in Table I. Those data were obtained from volumetric MRI images [5], with reference to anatomical data [6]. The opening location of the frontal sinus has lower accuracy than the others because it could not be identified exactly with the MRI data.

RESULTS

According to the theoretical considerations, zeros caused by the paranasal sinuses are expected to appear in transmission characteristics of the frontal portions which include the openings of the sinuses. Using the method, anti-resonances pertaining to each of the sinuses can be estimated, because the openings of the paranasal sinuses are separate along the nasal tract.

The results measured from the left and right nasal passages are shown in Fig. 2 (a) and (b), respectively, for Subject 1. There are several zero patterns

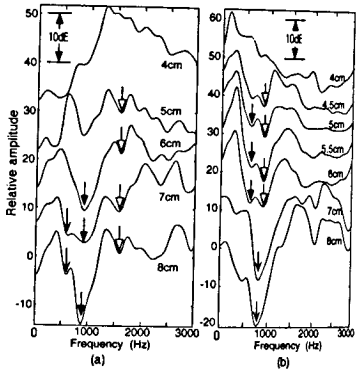


Figure 2. *Transmission characteristics the left (a) and right (b) nasal tracts obtained at a set of measurement positions for Subject 1. (The positions are shown in the right side. V-arrow shows the zeros caused by the sphenoidal sinus; black-arrow for the maxillary sinus, and white-arrow for the frontal sinus.)*

which appear in the transmission characteristics. As shown in Fig. 2 (a), there are three zeros in the frequency region below 2 kHz, which appear at about 590, 890 and 1580 Hz. A zero at 590 Hz, shown in V-arrows, appears in the transmission characteristics obtained in the measurement positions of 7 and 8 cm from the nostrils, and disappears in the measurement positions shorter than 7 cm. This implies that an opening of the sinus was in the frontal portions of 7 and 8 cm, and not included in the other frontal portions. Therefore, the opening is estimated to be between 6 and 7 cm. From Table I, it is known that the opening of the sphenoidal sinus was at 6.2 cm from the nostrils for the subject. According to the morphological data, the opening is uniquely judged to be the ostium of the sphenoidal sinus, and the zero at 590 Hz is determined to be the anti-resonance frequency of the sphenoidal sinus.

Similarly, a zero at about 890 Hz, shown by black-arrows, appears in the transmission characteristics of the frontal portions longer than 5 cm, and becomes much weaker at 5 cm, and disappears in the result measured at 4

Table II Estimated anti-resonance frequencies of the paranasal cavities (Hz).

Subs.	S. S.		M. S.		F. S.	
	left	right	left	right	left	right
Sub.1	595	788	895	612	1579	893
Sub.2	1047	1071	541	453	864	762
Sub.3	314	?	947	316	625	600

cm. With reference to the morphological data, this zero is judged to be caused by the maxillary cavity. There, the zero is more or less seen in the result obtained at 5 cm because the opening of the maxillary sinus, located at 5.1 cm, may affect the result.

A zero at about 1580 Hz, shown by white-arrows, appears in the results of the frontal portions of 5 to 8 cm, and disappears in the results measured at 4 cm. The same consideration gives the idea that an opening of a sinus exists in the region between 4 and 5 cm from the nostrils. The MRI data listed in Table I show that the ostium opening of the frontal sinus is in this region. According to the nasal morphology and the zero patterns, a conclusion can be drawn that the opening is the ostium of the frontal sinus, and the zero is caused by the sinus.

Using the same technique, estimations were made for the right side of the nasal cavity shown in Fig. 2 (b). Anti-resonance frequencies of the paranasal cavities are 788 Hz for the sphenoidal sinus, 612 Hz for the maxillary sinus, and 893 for the frontal sinus.

The anti-resonance frequencies of the paranasal cavities are shown in Table II for three subjects. Anti-resonance frequencies of the maxillary sinus are between 310 and 950 Hz. The ranges of the anti-resonances are from 310 to 1070 Hz for the sphenoidal sinus, and from 600 to 1580 Hz for the frontal sinus. For the three subjects, individual differences of the anti-resonances were generally larger than the differences due to the asymmetry of the paranasal cavities within the subjects.

CONCLUSIONS

Acoustic properties of the paranasal cavities were measured using a direct method. A set of transmission characteristics were obtained using sound pressure gradients at a series of measurement positions and sound pressure at the nostrils. Anti-resonance frequencies of the paranasal cavities were estimated by matching zero patterns of the transmission characteristics to morphological data for three subjects. The results showed that the paranasal sinuses, the sphenoidal sinus, the maxillary sinus and the frontal sinus, contribute zeros to acoustic characteristics of the nasal tract, respectively. It is expected that the zeros caused by the paranasal cavities can affect the shaping of nasal formants because the zeros appear in the low frequency region stably.

ACKNOWLEDGMENT

The authors would like to thank Hiroyuki Hirai for his helpful discussions and comments. We would also like to thank Naoki Kusakawa for his help in the experimental setup.

REFERENCES

[1] Lindqvist-Gauffin, J, and Sundberg, J. (1976). "Acoustic properties of the nasal tract," *Phonetica*, 33, 161-168.

[2] Takeuchi, S., Kasuya, H. and Kido, K. (1977). "A study on the effects of nasal and paranasal cavities on the spectra of nasal sounds," *J. Acoust. Soc. Jpn.*, 33, 4 163-172. (in Japanese).

[3] Flanagan, J., L. (1972). *Speech analysis synthesis and perception*, Springer-Verlag, New York (2nd Edition).

[4] Dang, J., Honda, K.(1994). "A new method for measuring vocal tract transmission characteristics," *Tech. Report of ATR* , TR-H-108.

[5] Dang, J., Honda, K., and Suzuki, H. (1994). "Morphological and acoustical analysis of the nasal and the paranasal cavities," *J. Acoust. Soc. Am.* , 96, 4, 2088-2100.

[6] Bunch, M. (1982). *Dynamics of the singing voice*, Springer-Verlag, New York.

THE DOMAIN OF ARTICULATION RATE VARIABILITY IN CZECH

J. Dankovičová
Phonetics Laboratory, University of Oxford, Great Britain

ABSTRACT

An experiment using three samples of spontaneous Czech speech from different speakers was conducted to find out whether variability of articulation rate (AR) is bound to a certain domain. AR was examined within three candidate units: the interpause stretch, the tone unit and the clause. In none of these units is AR constant. However, the tone unit manifests a regular pattern (a gradual slowing down) of AR.

1. INTRODUCTION

Articulation rate, i.e. a measure of rate of speaking in which pauses are excluded from the calculation, has not been widely studied. Taking into consideration only research on intra-speaker AR variability, several studies show that there is a considerable variability throughout the speech of an individual (e.g. [1], [2]). Unfortunately, due to differences in methodology (varying minimum pause duration, measures of rate used and, in particular, the unit across which AR is measured), the results across different studies are not readily comparable. A few other studies go beyond the quantification of AR variability by examining what are its determinants. Crystal & House ([3]), for instance, claim that the AR of a stretch of speech depends on its phonological structure. Eefting ([4]) states that intentions of the speaker with respect to the listener play an important role. Kaiki *et al.* ([5]) also point out the relevance of the difference between content and function words.

It should be noted, however, that an explicit justification for the choice of the unit across which AR is measured is often missing. As Eefting ([4]) notes, this may imply that AR within such a unit is assumed to remain constant. Lack of clarity in this matter indicates that articulation rate is a phenomenon whose functioning is still far from well understood and that it is not obvious how to incorporate it into linguistic theory.

This paper aims to investigate one aspect of AR variability by asking whether AR is indeed constant within some unit of speech or whether it follows some regular pattern; in other words, whether there is a fixed domain of AR variability.

2. METHOD

Three candidate units were chosen, within each of which AR was measured: the interpause stretch, the tone unit and the clause. The reasons for this choice were the following. In the case of the interpause stretch (for which the minimum criterial pause duration was 130 ms), pause characteristics and AR are known to be related (e.g. [6]). Moreover, this is the unit most often used in past research on AR variability. Tone units (defined as in [7]) are considered to be units of both rhythm and intonation in Czech ([8]), these being related notions. Clauses (syntactic stretches containing a finite verb) were chosen to cover the possibility of syntactic determination of AR variability. The minimum unit across which measurements were carried out was the phonological word, defined as a string of syllables with one stress. Within phonological words (henceforth simply 'words') AR was expressed as the number of syllables per second.

Three samples of Czech speech from three native speakers (one male and two female) were used; these were students of Charles University, Prague, aged 22 - 25 years. All samples were two minutes long and consisted of unprepared monologue on a topic chosen by the speaker. A picture from a children's book was available as a catalyst in case the subject did not know what to talk about.

The measurements were carried out with Signalyze™ 2.03 speech processing software using the time-amplitude and spectrograph displays.

3. RESULTS

Since the initial impressionistic analysis suggested that AR variability might in some way be bound to tone units, these were examined first.

The total number of tone units analysed was 179, out of which the majority were one word long (30.5%). In general, the more words tone units contained, the fewer there were of them (2-words TU: 22%; 3-words TU: 24%; 4-words TU: 14%; 5-words TU: 8% and 6-words TU: 1.5%).

3.1 Is articulation rate constant within tone units?

Comparison of the AR of the phonological words in 2-word tone units showed that these words were generally markedly dissimilar. In 85% out of the total of 39 tone units, the second word was slower than the first one.

For 3-, 4- and 5-word tone units, the coefficient of variation (i.e. deviation from the mean expressed as a percentage) was calculated. The values for all these tone units together are shown in the figure below.

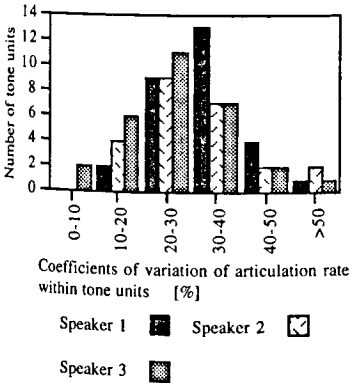


Figure 1. Frequency distribution of tone units according to their internal AR variability

Figure 1. shows that AR variability within tone units is far from constant. Variability is typically between 20% and 40%. Values below 10% and above 50% can be described as exceptional. A similar analysis was conducted separately for groups of tone units sharing the same number of words; this showed that the same range of variability holds for all groups. Such a high degree of variability is unlikely to be due solely to phrase-final lengthening; it must be distributed throughout the whole tone unit.

3.2 Internal structure of AR variability within tone units

This part of the analysis focussed on tone units consisting of 2, 3, 4 and 5 phonological words. The analysis was based on the rank ordering of component words according to their AR. Absolute values of AR were not considered at this stage.

As observed above, in the absolute majority (85%) of 2-word tone units the second phonological word was slower than the first one. Taking together tone units larger than this, in 88% of the cases, the first or the second word was the fastest. The number of tone units in which the first word was the fastest was approximately equal to the number in which the second word was the fastest. The last word tended to be the slowest; only exceptionally was it the fastest word in a tone unit.

Apart from these tendencies, other regularities occurred in all three speech samples. In particular, the majority of tone units shared a similar pattern of AR variability - a gradual decrease throughout the unit. I shall refer to it further as 'rallentando'.

For 3-word tone units, the prevailing patterns, represented iconically, are given in Table 1. (dots stand for phonological words, lines for the direction of AR change; the distance between the dots does not reflect the actual differences in AR; the percentage shows how many tone units of a certain pattern occurred in the total number of tone units of the same size):

Table 1. Prevailing patterns of 3-word tone units

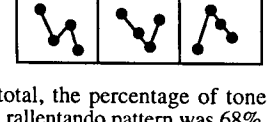
A (49%)	B (40%)
	

For 4-word tone units, six patterns were classified as rallentando, using the following diagnostic test: (i) the phonological word with the highest AR must be in the first or the second position within the tone unit; (ii) the phonological word with the slowest AR must not follow immediately that with the highest, and (iii) there must be no more than one increase in AR between successive phonological

words within the tone unit. The relevant patterns are shown in Table 2.

Table 2. Prevailing patterns of 4-word tone units

C (24%)	D (20%)	E (8%)
		

F (8%)	G (4%)	H (4%)
		

In total, the percentage of tone units with a rallentando pattern was 68%.

For 5-word tone units, the same rules defining rallentando patterns applied. This pattern occurred in 50% of cases.

This analysis leads to the conclusion that AR does vary within tone units and that the variability is systematic in the majority of cases - it gradually decreases. In the group of patterns that were not classified as rallentando, a few cases showed a gradual increase in AR within the tone unit (an 'accelerando' pattern) and the others did not form a homogeneous group and were called 'anomalous'.

In summary, the tone unit is a plausible candidate for the domain of AR variability.

3.3 The relation between the 'size' of the phonological word and its AR

At the next stage of the analysis, the possibility of explaining anomalous and accelerando pattern in terms of the size of component phonological words was considered. By (phonological) size is meant the number of syllables and long vowels in the word. Czech has distinctive vowel quantity and long vowels are claimed to be about twice of the duration of their short counterparts, thus a word with e.g. 3 syllables and no long vowel (3 /0/) is taken as equivalent to a word with 2 syllables and 1 long vowel (2 /1/).

The analysis did not demonstrate any consistent correlation between the size of component words of the tone units and their AR. Some typical examples are shown in Table 3. (AR is in syll/s).

Table 3. The relationship between the phonological size of words and their AR

	2 /1/ 5.00	4 /0/ 7.51	4 /0/ 5.73	
	3 /1/ 7.40	3 /1/ 5.25	2 /0/ 8.03	2 /0/ 5.52
	4 /1/ 7.19	2 /0/ 6.60	3 /0/ 9.06	2 /0/ 3.75

Phonological words of the same size differ in their AR considerably. There is also no consistent tendency for the smallest words to be the slowest and the largest words to be the fastest and vice versa. The patterns in Table 3. were all classified as anomalous but the analysis showed similar results for rallentando patterns too. Thus, the position of a word within a tone unit seems to be more important for the AR than the word's size.

An additional examination of anomalous and accelerando patterns in terms of their linguistic structure and pausing suggests that there might be some link between the kind of a word (functional vs. content), the character of a planning process (hesitations), etc. These speculations will, however, have to be examined in more detail in the future.

3.4 Structure of AR within interpause stretches and clauses

The first part of the analysis suggested that the tone unit is a plausible candidate for the domain of AR variability. In the second part, AR within interpause stretches and syntactic clauses was examined since these units may covary with tone units or demonstrate different systematic patterning.

3.4.1 Interpause stretches

The total number of interpause stretches in the material was 119. In all three speech samples, interpause stretches were coextensive to a considerable extent with tone units (80%, 58% and 59% for speakers 1, 2 and 3 respectively). Boundaries even of those interpause stretches in which this was not the case coincided with tone unit boundaries. They contained up to three tone units. Close examination of all interpause stretches failed to reveal any consistent regular patterning which would involve the interpause stretch as a whole. The only patterning observable was patterning within tone units. An illustration of this can be seen in Figure 2. below (points represent the AR of individual successive phonological words; successive points/words of the same colour - black or white - belong to the same tone unit; vertical lines indicate boundaries of interpause stretches).

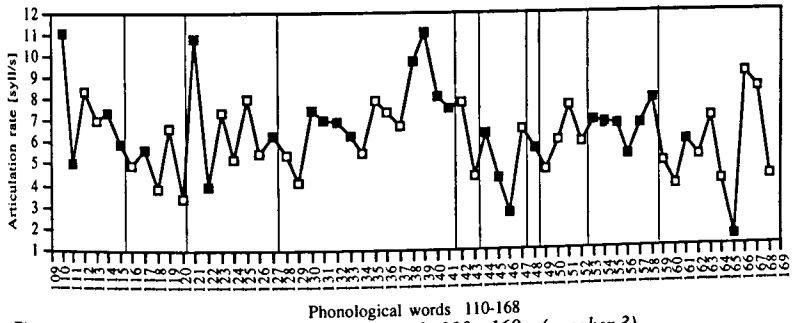


Figure 2. Interpause stretches containing words 110 - 168 (speaker 3)

3.4.2 Syntactic clauses

General findings in the analysis of clauses were similar to those for interpause stretches. Coincidence of clauses with single tone units was relatively high (69%, 52% and 55% for speakers 1, 2 and 3 respectively). In clauses which contained more than one tone unit (the boundaries normally coincided), no systematic patterning of AR variability was found. Again, the only patterning detected was that within tone units.

4. CONCLUSION & DISCUSSION

The pilot experiment described suggests that AR does have a domain of variability - the tone unit. Contrary to the implications in the literature (see above), there does not seem to be any level at which AR is constant if we accept that the phonological word is the minimum measurement unit in which it is reasonable to talk about 'articulation rate'. There is, however, a recurrent pattern of variability within the tone unit - a gradual slowing down with the first or the second word being the fastest. These regularities do not correlate with the number of syllables and long vowels in component words.

Neither the interpause stretch nor the clause demonstrate any systematic intra-unit AR patterning.

The results are in agreement with the claims about the importance of the tone unit with respect to speech rhythm. Also, they suggest that the tone unit might be a planning unit at some stage of speech processing. A larger-scale analysis will be conducted in the future to test the results in Czech and also English.

Acknowledgement

I should like to thank the Phonetics Institute at Charles University, Prague, for allowing me access to their research material.

REFERENCES

[1] Miller, J., Grosjean, F. & Lomanto, C. (1984), "Articulation rate and its variability in spontaneous speech: A reanalysis and some implications", *Phonetica*, 41, pp. 215-225.

[2] Rietveld, A.C.M. & Eefting, W. (1988), "Temporal variation in natural speech", *Proceedings*, 12, Department of Language and Speech, Phonetics section, University of Nijmegen, pp. 12-13.

[3] Crystal, T.H. & House, A.S. (1990), "Articulation rate and the duration of syllables and stress groups in connected speech", *JASA*, 88, pp. 101-112.

[4] Eefting, W. (1990), "Production and perception of temporal variation: An explorative study", *OTS Yearbook 1990*, Rijksuniversiteit Utrecht, pp. 13-38.

[5] Kaiki, N., Takeda, K. & Sagisaka, Y. (1990), "Statistical analysis for segmental duration rules in Japanese speech synthesis", *Proceedings of International Conference on Spoken Language Processing*, Kobe, Japan.

[6] Butcher, A. (1981), *Aspects of the speech pause: phonetic correlates and communicative function*, Institut für Phonetik, Universität Kiel, Arbeitsberichte nr.15.

[7] Crystal, D. (1969), *Prosodic systems and intonation in English*, Cambridge: CUP.

[8] Palková, Z. (1994), Fonetika a fonologie češtiny, Praha.

ACOUSTIC PROFILING OF GLOTTAL AND GLOTTALISED VARIANTS OF ENGLISH STOPS

G. J. Docherty and P. Foulkes
Department of Speech, University of Newcastle upon Tyne

ABSTRACT

An investigation of the acoustic characteristics of glottal and glottalised variants of stops in Tyneside English has found that the patterns of phonetic realisation which can be observed cannot straightforwardly be matched to the segmental categories [ʔ] and [ʔp, ʔt, ʔk] which are most commonly presented in accounts of these variants based on auditory analysis. The implications of these findings for phonological accounts of this aspect of English are discussed

INTRODUCTION

The consensus of work on glottal variants of voiceless stops in British English [1, 2, 3] is (i) that they have become progressively more widespread, (ii) that there are two types of glottal variant (glottaling, where the oral stop is 'replaced' by a glottal stop, and glottal reinforcement where the oral stop is doubly articulated with a glottal stop), and (iii) that there is social and geographical variation with regard to the frequency with which these variants occur, the environments in which they occur, and the extent to which stops other than /t/ are affected. Most previous work on this aspect of English has focused on describing and explaining the environments in which the variants are found, but the precise phonetic nature of these variants seems to have been taken somewhat for granted and receives scarecely a mention in the literature. The view which appears to prevail [4] is that the glottal stop produced in English glottal variants has the features of other stops, namely an onset phase, an occlusion phase with a duration similar to that found in other stops (marked acoustically by a silent gap), and a release phase. A slightly different account is presented in earlier work [5] referring to a 'distinct crack of the voice. a ceasing of the vowel sound before the consonant sets in'. In the case of glottally reinforced stops, in most accents the reinforcing gesture is timed so as to slightly precede the oral gesture [2, 4]. However, this is not universally the case, and in at least one other accent (Tyneside English) it is timed so as to mask the release of the oral stop articulation [2].

In the context of a project which aims to track phonological variation and change in British English, an auditory and acoustic study of glottal variants in Tyneside English has been carried out. Our aim is to provide a descriptive phonetic account of these variables to test existing acounts and to provide a firm phonetic foundation for subsequent phonological and sociolinguistic analyses. This paper presents the method used to construct an acoustic profile of of the glottal variants, together with the principal findings of the analysis performed to date.

METHOD

Fieldwork in the Tynside region of England has produced recordings of 32 speakers (2 social groups [WC/MC] * 2 [male/female] * 2 age groups * 4 speakers in each group). Speakers were recorded firstly in a dyad conversational exchange for around 50 minutes followed by reading a word-list designed to include a number of cases of stops in positions rendering them liable to glottaling or glottalisation. Field recordings have been supplemented by a smaller number of studio recordings of subjects reading a word-list and engaged in a map-description task .

Data from the field recordings has been analysed auditorily, with particular attention being paid to /p, t, k/ in word-medial and word-final position, revealing, as expected, numerous tokens where /t/ is perceived as glottalised and where /p, t, k/ are perceived as being glottally reinforced, with the latter occurring more frequently than the former. Rather than considering the frequency with which the different variants were encountered or the factors which led to one variant rather than another being produced, our focus here is on the phonetic nature of the glottal variants which we have observed. Acoustic analysis has been carried out so far of the data from the 4 young UWC male subjects, one of which will be used to exemplify the findings below.

With little guidance available in the literature, the acoustic analysis set out to track a range of spectrographic parameters which seem relevant for describing these variants, and which allow identification of their salient features.

Supralaryngeal Articulation

Following [6], as an index of supralaryngeal articulation, F1 and F2 have been tracked into and out of the stops undergoing analysis. Presence of transitions entering or emerging from a 'stop' indicates the existence of an oral gesture. Absence of such transitions is a little ambiguous, since it could mean that no oral gesture has been formed, or that an oral gesture is present but its existence has been masked by a reinforcing glottal gesture. A further indicator of the presence of an oral occlusion is the existence of a stop release burst. This indicates that there has been a build-up of intra-oral pressure posterior to an articulatory occlusion, but it does not, on its own, determine whether the occlusion has been fully pulmonic or whether it has been glottally reinforced (i.e. it could potentially be some form of ejective stop [2]).

Laryngeal Articulation

In this respect, the aim of the analysis has been to track changes in laryngeal articulation from the voicing during the vowel preceding the target stop through the stop and into the following vowel or sonorant. In view of the observation [7] that laryngealisation often occurs as an intermediate stage between voicing and a glottal stop, particular attention has been paid to identifying intervals of irregular vocal fold vibration. Our definition of laryngealisation is therefore based on visual and acoustic criteria and is not capable of discriminating between the different types of vocal fold configuration which might lead to laryngealisation (cf [8].), which represents one of the limitations of this study.

Presence of 'Stop Gap'

Given the conventional wisdom that English glottal variants should be characteristed acoustically by a stop 'gap' in the spectrographic trace, this is a further feature which forms part of our acoustic profile. Note that the interpretation of a stop 'gap' in a spectrographic trace is, on its own, not entirely transparent, since it could be produced by a glottal or oral occlusion or both, but it can be given a clearer interpretation in conjunction with information about any formant transitions into and out of the stop.

RESULTS

The acoustic analysis confirms the auditory impression that some stops are produced without a supralaryngeal gesture whilst many others are produced with a supralaryngeal gesture, but it is not always transparent whether that gesture is a complete occlusion or not. It would only be possible to be sure about this if a stop release burst is present, but there are very few of these in the data. In addition, there is the possibility that these speakers are producing incomplete stop gestures (spirantised articulations of this sort are common in accents of English in exactly the same environments investigated in this data). This is clearly an area where a more detailed articulatory analysis, possibly with EPG, would be beneficial.

With regard to the laryngeal gestures, there are relatively few cases where the acoustic trace resembles a [canonical] glottal stop (as described by [4, 7, 8]). In most cases where /p , t, k/ are heard to be glottalised or in the relatively fewer cases of /t/ being perceived as glottaled, speakers produce a segment with laryngealised voice throughout.

With regard to the sequencing of laryngeal-suparalaryngeal gestures, it seems that in general in Tyneside English there is a tendency for the laryngeal gesture to lag behind the supralaryngeal gesture, as revealed by the fact that transitions are often observed entering stops but not leaving them, and by the fact that laryngeal-isation has a tendency to spread into

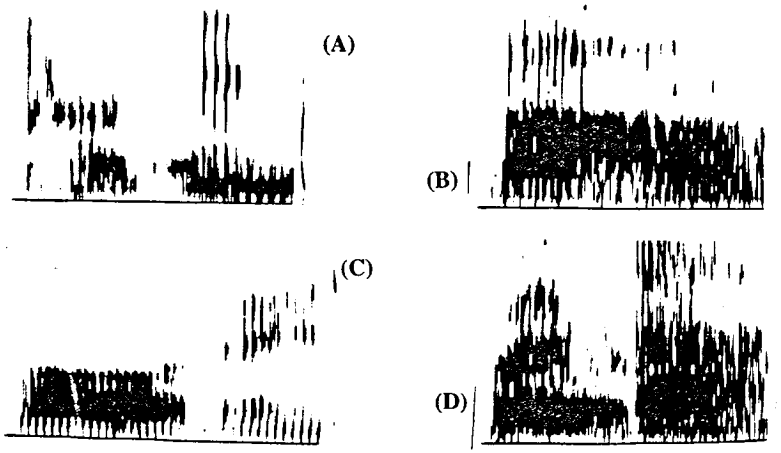


Figure 1 (A-D). Examples of the types of glottal variant observed (with 1kHz horizontal markers); (A) put in, (B) battle, (C) bought it, (D) footer. See text for further details.

following vowels, but is found less frequently on vowels preceding the stop. This corresponds to the observation made by previous auditory studies of Tyneside glottalisation which note that it is different from the patterns of pre-glottalisation found in other accents of English in that and it tends to mask the release burst of the stop as opposed to affecting the stop's onset (hence the difficulty in identifying whether a complete occlusion has been formed or not).

In general, the acoustic analysis described above reveals four types of glottal variant in Tyneside English, with (b) and (d) occurring most frequently; (a) a stop gap (although never entirely silent) with no indication of an oral gesture -- this is the variant which most closely matches descriptions of a canonical [ʔ] (Fig. 1A); (b) laryngealised voicing through the stop 'slot' with no indication of an oral gesture (Fig. 1B); (c) stop gap (although only rarely entirely silent) with clear indications of an oral articulation in the form of transitions into the stop, but note that the release burst and often the C-to-V transitions are masked by the glottal articulation (Fig. 1C); (d) laryngealised voicing through the stop phase together with clear indications of an oral articulation but with the release burst and often the C-to-V transitions masked by the glottal articulation (Fig. 1D).

Our results suggest therefore, that, in Tyneside English, it is relatively rare for voiceless stops heard as glottalised or glottaled to be produced with a glottal stop as this has traditionally been defined. It is much more common for the perception of glottal articulation to be based on the presence of laryngealised voice quality, with this present throughout the target segment. The phonetic symbols [ʔ] and [pʔ, tʔ, kʔ] do not in most cases capture what we have observed. It has been reported before [9, 10] that laryngealised voice can correlate with the percept of a glottal or glottalised stop, but the present results suggest that, in at least one variety of English, 'glottal' variants may not involve the formation of full glottal stop at all, or at least only rarely.

The second principal finding of this study is that in glottally reinforced stops in Tyneside English, the laryngeal gesture tends to lag behind the oral gesture. This is a very subtle difference between Tyneside English and that found in other regions of the UK, but it has a very significant auditory effect, namely the masking of any stop release which is formed (thus confirming [2]).

DISCUSSION

Any attempt to provide an account of the systematic phonetic characteristics of glottal variants of /p, t, k/ in Tyneside English will have to account for at least the following: (a) speakers have precise control over the timing of laryngeal and supralaryngeal articulations in cases of glottal reinforcement; (b) in the environments considered to date, there is a degree of flexibility about the glottal articulation which takes place, but it is only rarely a 'full' glottal stop, the more normal case being some form of increased glottal tension or constriction leading to laryngealised voice, and (c) in these environments, /t/ is sometimes produced with an oral gesture and sometimes without.

Whilst space does not permit a full exploration of these issues, it is likely that (a) - (c) above represent a challenge to the view held by many phonologists, and expressed in [11], that 'phonological representations should be mapped directly into speech output without passing through a buffer level of allophonic representation'. A feature representation has insufficient resolution to capture the temporal control exercised during the production of glottalised stops, and a direct mapping from feature specification (such as [+constricted glottis]) to speech output is not compatible with the variability observed in the glottal characteristics of the stops investigated in the present study. Furthermore, any account of glottaling and glottalisation as a form of lenition process must account for the fact that glottaling applies predominantly to /t/ and not to /p, k/, and that the linguistic and sociolinguistic distribution of glottaled and glottalised variants is not uniform [3]

It is possible that the present observations are part of the wide variety of fine-grained yet systematic language- and accent-specific characteristics of speech production which just cannot be governed by a phonological representation (as these are typically presented), and yet which are undoubtedly part of the grammar in the broadest sense of the term -- i.e. what it is that native speakers do, in order to be native speakers. Detailed phonetic studies such as that in which we are engaged in serve to highlight the need for greater research into these aspects of speech production and a theory of phonetic implementation which does not have the built-in limitations of a feature geometry or distinctive feature matrix.

ACKNOWLEDGEMENT

Research supported by the UK Economic and Social Research Council (Grant no. R000 234892 "Phonological variation and change in contemporary spoken British English").

REFERENCES

[1] Roach, P.J. 1973. Glottalization of English /p/, /t/, /k/ and /tʃ / - a re-examination. *Journal of the International Phonetics Association,* 3: 10-21.

[2] Wells, J.C. 1982. *Accents of English* (3 vols). Cambridge: Cambridge University Press.

[3] Milroy, J., Milroy, L. and Hartley S. 1994. Local and supralocal change in British English: the case of glottalisation. *English World-Wide,* 15,1: 1-32.

[4] Gimson, A.C. (1980) *An Introduction to the Pronunciation of English.* London: Edward Arnold.

[5] Christopherson, P. (1952) The glottal stop in English. *English Studies,* 33, 156-163.

[6] Manuel, S, & Vatikiotis-Bateson, E. (1988) Oral and glottal gestures and acoustics of underlying /t/ in English. Poster presented at the joint meeting of the Acoustical Society of America and the Acoustical Society of Japan, Hawaii.

[7] Catford, J.C. (1977) *Fundamental Problems in Phonetics.* Edinburgh: Edinburgh University Press.

[8] Henton, C., Ladefoged, P., & Maddieson, I. (1992) Stops in the world's languages. *Phonetica,* 49,65-101.

[9] Grice, M., & Barry, W. (1991) Problems of transcription and labelling in the specification of segmental and prosodic structure. *Proceedings of the XIIth ICPhS, Aix-en-Provence,* Vol 5: 66-69

[10] Kohler, K. (1994) Glottal stops and glottalisation in German. *Phonetica,* 51:38-51.

[11] Clements, G.N. (1992) Phonological primes: features or gestures? *Phonetica,* 49:181-193.

Testing a Dynamic Model of Pharyngeal Articulation

Ahmed M. Elgendy
Department of Linguistics, Stockholm University, Sweden

ABSTRACT

A dynamical model of pharyngeal articulation was designed to account for the mechanism underlying the use of the pharynx in speech production and to examine the nature of coarticulation in the back cavity of the vocal tract. Some aspects of the model are tested and its ability to predict the properties of the natural system is discussed.

INTRODUCTION

In search for the invariance units blended in the acoustic signal of speech, models have been constructed in order to account for the observed behavior of various articulators and to attempt to predict the properties of the natural system controlling the process of speech production.

Speech has been proved to be a dynamic and context-conditioned process at all levels. Articulatory dynamics involve co-ordinated movements of the articulators expressed in space and time. The question whether timing control over a speech utterance is issued externally or internally, i.e., included in the motor program, still a debatable problem (cf. e.g., [1]).

A standard model of speech production must consider activities of all parts of the vocal tract continuum together with activities of respiratory system. The nasal, oral and laryngeal portions of the pharynx constitute more than one half of the vocal tract length. Therefore, it is important to obtain exhaustive account on the physiology of the pharynx during speech. Understanding the mechanism underlying distinctive pharyngeal speech sounds would increase the efficiency of current models build up mainly for languages lacking pharyngeal phonemes and would improve our insight into the process of speech motor control.

Articulatory Dynamics of Pharyngeal Segment

In Arabic, the pharynx is used to produce distinct speech sound units both as primary as well as secondary place of articulation. The pharyngeal consonants, i.e., lower pharyngeals /ʕ, ħ/ and upper pharyngeals /q, ʁ, χ/ have the pharynx as their primary place of articulation. The pharyngealized consonants, on the other hand, use the pharynx as a secondary place of articulation with a major constriction in the oral cavity.

The production of the true pharyngeal consonants in Egyptian Arabic is characterized by a complex mechanism involving the control of co-ordinated activities of the pharynx, the epiglottis and the larynx. Sphinctric contraction of pharyngeal wall at the point of constriction occurs simultaneously with upward movement of the larynx and hyoid bone. This is accompanied by a constriction in the glottis and active bending of the epiglottis towards the arytenoids. The timing of the epiglottis movement is synchronized with a downward pull of the velum [2].

The resulting coarticulatory effect causes the jaw to sustain certain mechanical constraints realized as antagonism to the tongue movement and temporal reorganization of the syllable containing pharyngeal segment. That is, the synergies involved in controlling the production of pharyngeal segments restrict the jaw and the tongue from anticipating the articulation of the upcoming segments until the motor command is completely executed.

As a consequence, vowels are found to accommodate mandible position assigned to the pharyngeal segment intervocalically but not initially or finally in a word. The degree of jaw lowering is greater for mono-syllabic than tri-syllabic words. Acoustic analysis showed that the excessive degree of jaw lowering associated with pharyngeal segment production, compared to oral segments, is reflected as a compensatory effect on vowel duration. The degree of readjustment depends on the position of pharyngeal segment in the word and the relative degree of jaw-height of the consonant embracing the vowel at the syllable margin. That is, low vowels are longer when preceded than when followed by a pharyngeal consonant [2].

Eventually, vowels depart to their inherent position right after the gestures for pharyngeal segment are completed in initial but not in intervocalic position. Seemingly the articulators seek a rhythmic pattern among successive syllables. Furthermore, the degree of contextual overlapping between segments in non-pharyngeal environment is much greater.

The Structure of Arabic Language

The Arabic word is basically composed of three consonantal elements embedded in a finite set of vocalic patterns by which derivational forms can be generated. A vocalic pattern $|\approx V_1 \approx V_2 \approx|$ (a "tenon") is inserted into a given consonantal pattern $|C_1\text{–}C_2\text{–}C_3|$ (a "mortise") to generate a word. For example, the vocalic pattern |≈æ≈æ≈| can be tenoned to the mortise |k–t–b| so that the word /kætæb/ will mean "he wrote". Similarly, |≈i≈æ≈| when tenoned to the mortise |ʕ–t–b|, will yield the noun /ʔitææb/ "blaming". The inflectional constructions, on the other hand, are paradigmatically obtained by adding prefixes, infixes or suffixes to the derived forms. For instance the tri-consonantal mortise |k–t–b| can be used to generate /kætæb/ "to write"; /kotib/ "it was written"; /kotob/ "books"; /kættib/ "to cause to write"; /kæætib/ "a writer"; /kæætæb/ "he corresponded" etc.

There are five types of syllabic-pattern in Egyptian Arabic: CV, CVV, CVC, CVVC and CVCC in which a consonant must begin the syllable. Syllabic boundaries are located at the left of each consonant starting from the rightmost side moving to the begining of the word, e.g., /ʔis.tik.tæ æ b./. There is a strong correlation between syllabic structure, stress assingment and nucleus vowel duration in Egyptian Arabic. In addition, vowel length is phonemic in Arabic, e.g., /sæed/ "to dominate" vs. /sædd/ "to block".

The nature of Arabic word structure and the degree of compensatory lengthening or shortening exerted on vowels more than consonants strongly suggest that consonants are more "stable" than vowels in terms of their duration. Thus, consonants can be considered as "landmarks" linked together by vowels.

Vowels, on the other hand, being more flexible articulatory events than consonants, can tolerate greater amounts of compression or expansion. Accordingly, vowels are issued to preserve isochronical intervals between consonants in the syllable. They also can manage various coarticulatory effects resulting from the overlapping of successive segments due to inertial and mechanical constraints.

It is appropriate, then, to take the syllable as the basic unit of motor programming in Arabic (see [2] for an overview on the components gathered for dynamic modeling of pharyngeal articulation).

THE MODEL

Figure 1 shows schematic diagram for a dynamic model by which a word containing a pharyngeal segment can be generated. The development of the model divisions is based on current views describing strategies controlling motor programs of speech production (cf., e.g., [3]).

If the mortise |ʕ–l–m|, which bears the basic semantic unit "knowing", is selected from the lexicon and the tenon |≈ææ≈i≈| is inserted into it, the word /ʕæælim/ "a scientist" will be the concept input to the articulatory plan. The tenon is recalled from the morpho-semantic storage since the vocalic pattern partially provides the grammatical meaning of the word. Notice that |ʕ–m–l| and |l–m–ʕ| are permitted but not |m–ʕ–l|, |l–ʕ–m| or |m–l–ʕ|.

An articulatory plan must be available to decode the phonological rules before the motor plan can be executed. At this stage of high-level planning, the speech sound pattern is already designed as a series of consonants and vowels in a specific order comprises lexical and morpho-semantic items. Thus, the input for the articulatory plan takes the word as the minimal unit for phonological parameters applications.

Pharyngeal consonants prevent the anticipation of the ensuing segment to take place before their execution is completed. That, in turn, would have a perturbation effect on the temporal pattern. Hence, two articulatory strategies are presumably needed to control the timing organization.

The main strategy (the default) will fail to manage the existence of a pharyngeal segment in the utterance since it assumes that all consonants are apt to equally coarticulate with the coherent vowel(s). Accordingly, strategy (2) is proposed to

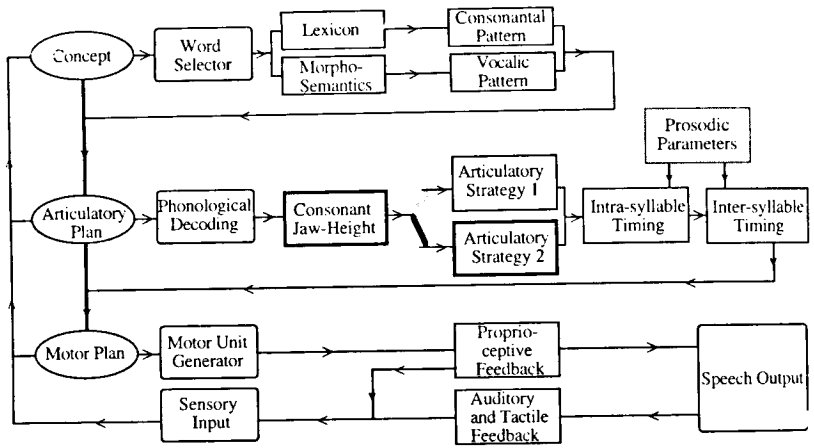


Figure 1. An articulatory model by which a pharyngeal segment can be generated.

tackle the inevitable temporal re-adjustments due to the constraints, imposed on the jaw and the tongue, characterizing the pharyngeal segment production.

The degree of re-adjustment in vowel duration, in case the final segment in a CVC is a pharyngeal segment, will depends on the degree of coarticulation between the first consonant and the following vowel. This is because the first consonant (non-pharyngeal) will allow the anticipation of the vowel. When the initial consonant is a pharyngeal, the vowel coarticulates only with the second consonant. Hence, strategy (2) will determine the degree of re-adjustment according to the degree of jaw-height of the non-pharyngeal consonant.

In both cases the value of jaw-height for both pharyngeal and non-pharyngeal consonants must be previously conceptualized. This is essential in order to maintain fixed intervals among syllables of the word. In order to keep a unitary syllable length, it is necessary that the duration of each syllable is calibrated as the combined duration of a consonant and a vowel. Hence, jaw-height must be determined for each consonant in the word since the total duration of the utterance will depend, to a great extent, on the trajectory of the jaw moving from one consonant constriction location to the next. Vowels, on the other hand, will occupy the intervals between consonants. Thus, they serve to secure the timing regulations needed to overcome the coarticulatory effects.

The module responsible for "intra-syllable timing" determines the degree of adjustment for segment duration (mainly for the vowel). The timing among various syllables in a word will be controlled by the "inter-syllable timing" module. The temporal pattern which governs the inter-relationship between syllables will be the end product of this process.

Next, intonation contour and stress assignment rules as prosodic parameters will be applicable on the sound pattern. Recall that stress position is correlated with the duration of the syllable and its position in the word. The application of other factors effecting the rate and style of speech also pertains to this stage of phonological decoding. As soon as the articulatory plan is discharged, it is fed into an articulatory buffer before the motor plan can be commenced.

TESTING THE MODEL

Phonotactics

The validity of the model can be attested by examining its ability to predict the properties of the natural system. For this purpose the distribution patterns of pharyngeal consonants with respect to all other consonants were stated as manifested in the phonotactic rules governing spoken Egyptian Arabic word structure.

It was found that the severe mechanical constraints exerted on pharyngeal segment production has a prevailing effect over the construction of the entire language system. Consonants in a given sequence are selected according to their compatibility to preserve the temporal aspects of syllable structure. That is, the organization of consonants in a sequence depends, to a great extent, on their relative degree of jaw-height. The co-occurrence of different consonants in a word is based on the consonant's inherent degree of jaw-height. The word-length can be seen as the path the jaw takes from one consonant to the next along the word (cf., [2]).

The model can successfully predict the severe restrictions on the distribution of pharyngeal segments since it contains a module for estimating jaw-height for each element of the consonantal pattern. Thus, the degree of jaw-height for each consonant in the sequence is determined prior to the listing of timing instructions for the entire utterance.

Moreover, the model could provide an explanation for the tendency observed of pharyngeal consonants to favor initial or final rather than medial position in a word. The temporal specifications in the model are highly restricted. A pharyngeal segment in medial position will demands that the execution of the motor plan must be reset. This justifies the finding that two pharyngeal consonants do not co-occur in one and the same consonantal pattern (cf. [2]). Recall that the pharyngeal segment posses an extreme degree of jaw lowering. Furthermore, the vast majority of the vocalic pattern used in Arabic language was found to be based on the low back vowel /æ/. Low vowels are more susceptible to coarticulate with pharyngeal consonants.

Acquisition of Pharyngeals

The acquisition of pharyngeal segment production takes considerably long time compared to oral consonants and has a gradual emergence. The acquisition stage is correlated with the place of constriction in the pharynx and the relative degree of jaw-displacement associated with each consonant. That is, the greater the degree of jaw displacement of pharyngeal consonant, the longer is the acquisition time [4]. This may indicate that specifying consonant jaw-height is a primary issue in the acquisition process.

The model underscores the importance of consonant's jaw-height as a primary phonological parameter by which one of the two articulatory strategies will be chosen. The delay in the acquisition time may be due to the availability of two strategies operating the temporal organization within and between the syllables of the utterance in the mature system .

Concluding Remarks

The proposed model presupposes that well specified aspects of the lexical item are available for execution from high-level planning down to the motor commands level. The positions permitted for the pharyngeal segment to occupy in the sound patterns indicate that the effect resulting from the mechanical constraints is considered by the central planning.

On the other hand, the selection of temporally compatible segments to co-occur with pharyngeal consonants in a word indicates that "timing" is issued internally, i.e., at high level in the brain. It is suggested then, that long-term feedback is essential for most of the stages of motor control process. The model's implications lend support to the view which considers coarticulation as a preplaned articulatory process. It remains, however, to test the model's ability to account for auditory perception and the speech motor commands adaptation to peripheral contexts. The model should also be re-evaluated in the light of other models of coarticulation to examine its validity to account for a universal system.

REFERENCES

[1] Fowler, C.A. (1980), "Coarticulation and Theories of Extrinsic Timing". *Journal of Phonetics*, 8, pp. 113-133.

[2] Elgendy, A.M. (1994), "Components for Dynamic Modeling of Pharyngeal Articulation". ICSLP, Proceedings of the International Conference on Spoken Language Processing. Yokohama, Japan, September 18-22, 1994, 12-2, pp. 1-4.

[3] Kent, R.D. and Minifie, F.D. (1977), "Coarticulation in Recent Speech Production Models". *Journal of Phonetics*. 5, pp. 115-133.

[4] Elgendy, A.M. (in preparation), "Delayed Acquisition of Pharyngeal Segment Production as a Result of Mechanical Constraints".

INFERRING THE COMMANDS OF AN ARTICULATORY MODEL FROM ACOUSTICAL SPECIFICATIONS OF STOP/VOWEL SEQUENCES

R. Laboissière and A. Galvan

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

E-mail: **rafael@icp.grenet.fr**

ABSTRACT

We present a part of our efforts towards an articulatory speech synthesizer capable of learn to produce articulatory gestures from acoustical description of the tasks. We concentrated on the problem of characterizing stop consonants in the formant space. We model the stop consonants targets as probabilistic models, which has advantages for both a quantitative assessment of the principle and for application to a model of speech motor control.

INTRODUCTION

A model was developed for obtaining the commands of an articulatory model of the vocal tract from acoustical targets (Laboissière 1993). Using this model, acceptable vowel–vowel transitions are obtained and typical phenomena related to coarticulation and compensation for perturbation can be replicated. These satisfactory results rely on the fact that acoustical targets (in our case the three or four lowest formants) are well defined for vowels. Problems arise when trying to use the system to infer articulatory commands for stop/vowel transitions.

Attempts to find invariant acoustical cues for stop consonants are abundant in the literature (Stevens and Blumstein 1978; Kewley-Port 1983; Sussman et al. 1991). Although the invariance at the acoustical level is still a matter of debate, we are pursuing this paradigm in order to test its validity in the context of a model of motor control for speech production.

In this paper we will describe the preliminary efforts towards our approach to vowel-consonant-vowel articulatory synthesis, and is organized as follows: we present first the principles of our inversion model; second, the technique for obtaining targets for consonants in the acoustical space will be presented as well as an preliminary assessment of the principle.

THE INVERSION MODEL

The schematic of the model we are using to invert from acoustical (distal) desired outcomes into articulatory (proximal) commands is shown in Fig. 1. This scheme is reminiscent of classical techniques in Control Theory, namely feedback control with learning of a feedforward controller.

This control model drives an articulatory model of the human vocal tract (Maeda 1988, $F(u)$ in Fig. 1), implemented as a computer program. The articulatory model was generated from cineradiographic data from a speaker uttering ten phonetically-equilibrated French sentences. From a sort principal component analysis of the mid-sagittal tongue contour it was possible to derive seven articulatory commands like jaw/tongue position, lips aperture/protrusion and larynx height (these commands compose the vector u). At the output, after computing the area function of the resulting configuration of the vocal tract, we extract the first four formants (y in Fig. 1).

As the number of inputs is greater than

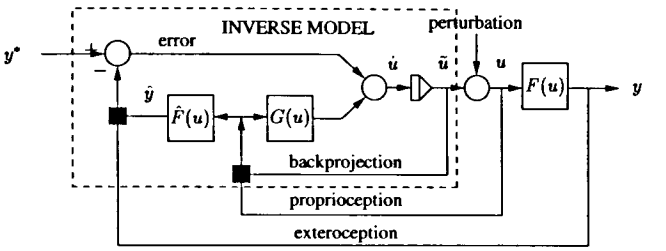


Figure 1: Architecture of articulatory controller. The system to be controlled (the plant) is indicated by $F(u)$, the articulatory inputs are u and the perceptual outputs (formants) y. The *inverse model*, capable of inferring the articulatory inputs from the desired outputs y^* contains a forward model $\hat{F}(u)$ that gives estimations $\hat{y}$ of the plant outputs from the articulatory inputs, obtained either by proprioceptive feedback or through "backprojection."

the number of outputs—i.e. the system has degrees of freedom in excess—there is no unique inverse transformation from the desired formants into articulatory commands. The proposed architecture solves this problem in two steps. First, a *forward model* of the articulatory model is learned ($\hat{F}(u)$) in order to mimic the articulatory model [see Jordan and Rumelhart (1992) for a thorough discussion on forward modelling]. More precisely, $\hat{F}(u)$ is an analytical approximation of the mapping $F(u)$. To find this regression model, we are using a mixture of linear experts trained by the Expectation-Maximisation (EM) algorithm, a technique introduce by (Jordan and Jacobs 1994). Essentially, the forward model implements a piecewise linear function.

The main interest of having the piecewise linear approximation (or any simple regression) resides in obtaining a simple expression for the controller $G(u)$. Indeed, $G(u)$ implements a piecewise constant matrix of transformation between the vector of error in the acoustical space (derived from both y^* and $\hat{y}$ or y) and the changes in the articulatory commands. As we use the pseudo inverse of the JAcobian, we ensure that minimal changes in the articulatory variables will be produced for a given acoustical error vector (Klein and Huang 1983). This means that our model can produce smooth commands without any need for planning. Another interesting feature of this architecture is that once the forward model has been learned, the combination of $\hat{F}(u)$ and $G(u)$ can act as a feedforward *inverse model*.

ACOUSTICAL TARGETS FOR STOP CONSONANTS

Let us turn now on how the controller shown in Fig. 1 actually works. In order to obtain articulatory movements, we have to present targets y^* at the input. For vowel production, this targets could be simply formant values (F_1 to F_4) and the error would be some distance between y^* and either y or $\hat{y}$. For the stop consonants there is no target in the formant space due to the occlusion of the vocal tract. The cues that convey information on the stop consonant identity are numerous, ranging from formant transitions to burst spectra [see Kewley-Port (1983) for a review].

In the present work, instead of concentrating on a dynamical description of stop–vowel production, we are asking a more fundamental question: is it possible to identify place of constriction from a kind of "intended formant configuration" that would be produced by the vocal tract just at the moment of occlusion release? Of course, this "formants" would not

exist physically in the speech signal, but could be considered as intentional target for stops.

This should relate to the locus equations (Sussman et al. 1991), but we are interested in a more general result, in which stop consonants targets could be associated with large regions in the formant space. We did a thorough exploration of the articulatory model, and were able to obtain several articulatory configurations that give the same place of constriction for the tongue. By computing the formant values for those configurations assuming a small aperture at the place of constriction, it is possible to obtain sets of points in the formant space like those shown in Fig. 2 (only F_2, F_3, and F_4 are shown, because F_1 is systematically close to 200 Hz for all configurations). The big variability observed is due to the free articulators, like lips and larynx, as well as to compensations between jaw and tongue positions.

The case shown in Fig. 2 is quite instructive. The clouds correspond to the same place of constriction (about 1.5 cm behind the upper incisors) but produced with different parts of the tongue: either the tongue dorsum or the tongue tip in a retroflex articulation. We see that the clouds are quite separable, but a more quantitative and systematic assessment of this assertion is called for. In order to do that, we model the cloud of points in the F_2–F_4 space as a probabilistic model, namely as a mixture of Gaussians. Given a vector y in that space and a model $\mathcal{M}_j$ related to a given position of constriction and mode of articulation (tongue tip or tongue dorsum), the probability of having y associated to $\mathcal{M}_j$ is given by

$$P(y|\mathcal{M}_j) = \sum_i g_{ji}|C_{ji}|^{-1/2} e^{-(y-y_{ji})^T C_{ji}^{-1}(y-y_{ji})}, \quad (1)$$

where g_{ij} are the a-priori probabilities, y_{ij} the mean vectors and C_{ij} the covariance matrices. For each of the possible locations of constriction of the articulatory model (from the alveolar to velar regions) spaced by 0.5 cm we found the best mixture of Gaussians using the EM algorithm. We observed that 4 Gaussians were in general sufficient for describing each cloud.

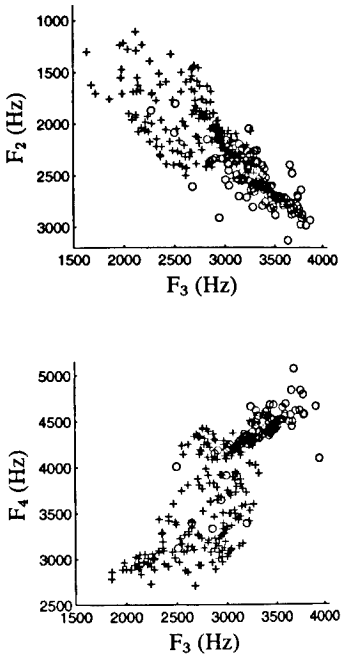


Figure 2: "Formant" values for retroflex constricitons with the tongue tip (+) and advanced tongue dorsum (◯).

Modelling targets as probabilistic models offers two advantages. First, it is possible to estimated the likelihood of each clouds being produced by the each model $\mathcal{M}_j$. This gives some measure of confusion between acoustical results of different place of constriction. Let us call y_j the points the in cloud related to the model $\mathcal{M}_j$. The log likelihood of having y_k eeing produced by $\mathcal{M}_k$ is

$$\mathcal{L}(y_j|\mathcal{M}_k) = \sum_j \log[P(y_j|\mathcal{M}_k)]. \quad (2)$$

The greater $\mathcal{L}(y_j|\mathcal{M}_k)$, the more will have confusion at the acoustical space between the related places of constriction. We compute those values for 11 clouds, three for the tongue tip articulation and 8 for the tongue dorsum. The tongue dorsum can constrict as far as 5 cm back from the incisors, which means the soft palate region. The results are summarized in Fig. 3, in which the values for the likelihoods are shown as gray levels. We interpolated the data in order to improve the presentation. Darker regions correspond to high likelihoods. It is possible to see that some regions of confusions emerge from our data: between positions d_4 and d_7 (which corresponds to the hard palate), regions t_1 and t_2 (dental and alveolar) and d_2 and d_3 (advanced tongue dorsum). Velar regions and tongue tip retroflex configurations are quite separable from the others.

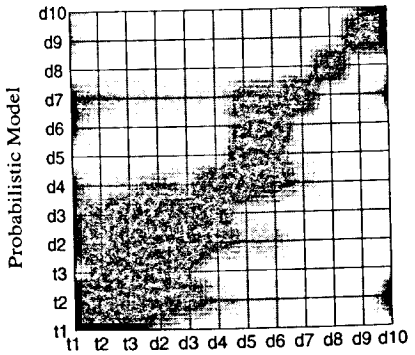


Figure 3: Log likelihhod of production of each cloud by each probabilistic model related to a place of constriction d_i stands for tongue dorsum and t_i for tongue dorsum. Darker regions mean high likelihood. Spacing of constrictions is 0.5 cm.

The second benefit of the probabilistic modelling is related to the way we can compute the error vector for the controller (Fig. 1). Indeed, for a given point in formant space y, the error vector is given simply by the gradient of the log likelihood with respect to y:

$$E = \frac{1}{P(y|\mathcal{M}_j)} \sum_{ij} g_{ij}|C_{ij}|^{-1/2} e^{-(y-y_{ij})^T C_{ij}^{-1}(y-y_{ij})}[C_{ij}^{-1}(y-y_{ij})]. \quad (3)$$

The error vectors generate a force field in the formant space which is transformed into changes in articulatory positions by the controller.

CONCLUSIONS

In this paper we showed how to obtain targets in the acoustical space for the stop components in the context of a model of motor control. We concentrated on describing the model and on how a probabilistic approach to the description of vowel–stop–vowel sequences can be useful. We showed that the description of clouds by mixtures of Gaussians yields interesting results, mainly related to the separability of the target regions for the different stop consonants produced by contact of the tongue to the hard- and soft-palate. Extensive simulations are planned for assessing the whole model.

ACKNOWLEDGEMENTS

We want to thank Jean-Luc SCHWARTZ for the insightful discussions and Louis-Jean BOË for making available the cineradiographic database. This research was supported by the European Community ESPRIT/BR grant # 6975 SPEECH MAPS and by GALILEO France-Italy project # 94290. The second author is recipient of a scholarship from CONACYT/Mexico.

REFERENCES

Jordan, M. I. and R. A. Jacobs (1994). Hierarchical mixtures of experts and the em algorithm. *Neural Computation 6*(2), 181–214.

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

Kewley-Port, D. (1983). Time-varying features as correlates of place of articulation in stop consonants. *J. Acous. Soc. Am. 73*(1), 322–335.

Klein, C. A. and C. H. Huang (1983). Review of pseudoinverse control for use with kinematically redundant manipulators. *IEEE Transactions on Man, Machines, and Cybernetics SMC-13*, 245–250.

Laboissière, R. (1993). Inversion and control of an articulatory model of the vocal tract: Recovering articulatory gestures from sounds. *J. Acous. Soc. Am. 93*(4), 2295.

Maeda, S. (1988). Improved articulatory model. *J. Acous. Soc. Am. 81*(S1), S146.

Stevens, K. N. and S. E. Blumstein (1978). Invariant cues for place of articulation in stop consonants. *J. Acous. Soc. Am. 64*(5), 1358–1368.

Sussman, H. M., H. A. McCaffrey, and S. A. Matthews (1991). An investigation of locus equations as a sources of relational invariance for stop place of categorization. *J. Acous. Soc. Am. 90*(3), 1309–1325.

MANDIBULAR MOVEMENTS AND SYLLABLES

P. Lindblad
Dept of Linguistics, Göteborg, Sweden
S. Karlsson and S. Lundqvist
Dept of Prosthetic Dentristry, Göteborg, Sweden

ABSTRACT

Jaw movements in phrases were studied optoelectronically (Selspot). The trajectories were constantly oscillating and quasi-regular. Most often, one oscillatory period corresponded to a syllable, with the dip in the vowel. This pattern was sustained also in high vowel contexts with oscillation amplitudes of only 1 mm. Integrated jaw oscillations were sometimes found before speech onset, and in silent pauses between phrases. These never-resting jaw movements constitute the basic syllabic speech structure. Cases of two syllables per mandibular period were also found. The conditions for such reductions are an urgent reseach issue.

INTRODUCTION

The syllable is a central but poorly understood phonetic entity. Evidently, it is highly important in both speech production and perception. Among other things, speech stress and rhythm patterns are closely connected to the syllabic structure.

A number of different phonetic correlates of the syllable have been proposed, both auditory and articulatory: sonority, loudness, degree of coarticulation, articulatory opening and closing, jaw movements [1, 2, 3]. However, no one-to-one, exceptionless correlate has been found. Nevertheless, the connections between speech syllabicity and these variables are far-reaching and significant.

Jaw movements are obviously closely connected with syllabic structure. These syllable-related movements are easy to observe with the naked eye. It is no coincidence that ventriloquists move the doll's jaw at a syllabic rate to give an impression of talk. Although it is possible to speak with a clenched jaw, it is highly unusual. Therefore, it is natural that hypotheses of mandibular movements as connected to syllabicity have been proposed long ago. Saussure's [2] proposal of closing movements (implosions) and opening ones (explosions) in connecting with syllabic boundaries is one example. Another, more elaborated hypothesis of the mandibular-syllabic connection was presented by Menzerath & de Lacerda [3] in the thirties. However, due to lack of technical possibilities at that time, the hypothesis could not be tested properly.

Today, technical resources exist for exploring the mandibular-syllabic connection. Strain-gauge and magnetic coil equipments, cineradiography, microbeam X-ray and optoelectronic filming have been used for tracking jaw movements [4, 5, 6]. Several studies have been reported since about 1980 of mandibular amplitudes and movements in relation to different sounds, stress, and tempo, e g [4, 5, 6]. However, the mandibular-syllabic connection has not been given much attention. The main aim of this paper is to give some fundamental, preliminary data on this issue.

Our main hypotheses are: (1) There is a far-reaching but not total correspondence between oscillatory jaw movement periods and syllables in speech. (2) The never-resting jaw movements constitute the basic syllabic speech structure.

METHOD AND MATERIAL

To elucidate the extent and nature of the mandibular-syllabic connection, the jaw movements have been analysed in natural phrases with the aid of optoelectronic filming (Selspot). This equipment consists of three basic units: light-emitting diodes, a position-sensitive detector located in two cameras, and a computer with a camera interface. For a closer description, see [7]. With this equipment, the three-dimensional spatial movements of one or more diodes can be analysed and displayed in calibrated curves and/or quantitative data, giving values of displacement amplitudes, velocities and velocity changes.

This study was based on systematic analysis of curves of mandibular movements in the vertical dimension only. The jaw movements were recorded at a sampling rate of 500 Hz by a single diode attached to the midline of the chin. Head movements were compensated for by a reference system consisting of three diodes attached to a spectacle frame worn by the subject. This registration was supplemented with a simultaneous microphone recording of the acoustic signal. This signal was registered in synchrony with the movement curve as an LP filtered acoustic waveform. In this curve, the segmentation of sound segments of the utterances were made. The acoustic signal was also perceptually analysed as concerns the different informants' degree of speech reduction and tempo.

Two substudies, A and B, have been made. In study A, 12 dental students participated, eight women and four men, mean age 24 years, range 22-27 years. (This study has been published [8].) In study B, ten other subjects participated, four women and six men, mean age 31 years, range 23-49 years. Nine of these were also dental students. All subjects were native speakers of Swedish with normal hearing.

The material spoken consisted mainly of natural phrases in Swedish, well varied as concerns vowels, consonants, consonant clusters, and stresses. The material in study A consisted of *Mississippi*, [ˌmɪsɪˈsɪpɪ], *Pappa tappar kopparna*, [ˈpapa ˈtapa(r) ˈkɔpaɳa], "Daddy drops the cups", and *Prinsessan sätter potatis* [prɪnˈsesa(n) ˌsetə(r) pʊˈtɑːtɪs], "The princess sets potatoes". In study B, the phrases were *Statsministerns sista tal måste läsas*, [ˈstasm(ɪ)ˌnɪstən(s) ˈsɪsta ˈtɑːl mɔstə ˈlɛːsas], "The prime minister's last speech has to be read", and *Cecilia tjatade att hon måste kila bort och kika på Sassas nya kjol*, [səˈsiːlɪa ˈɕɑːta (a)t (h)ʊ(n) mɔstəˌɕiːla ˈbɔʈ ɔ ˈɕiːka p(ɔ) ˈsasas ˈnyːja ˈɕuːl], "Cecilia kept saying that she had to run away and look at Sassa's new skirt". (The diacritics ´ and ` denote main stressed syllables, associated with Swedish tonal accent I and II, respectively. The parentheses surround segments that are often reduced.)

In study A, the three items were spoken one at a time. Each item was uttered five times by all 12 subjects. In study B, the two phrases were uttered with a very short pause between them. Each such sequence was uttered nine times by all 10 subjects. All informants spoke with a normal loudness and tempo, resulting in numerous expected reductions of consonant clusters and unstressed vowels.

RESULTS

The 22 subjects generally had similar mandibular movement patterns, albeit with some individual differences. Fig 1 shows a representative curve from one of the subjects. Several features found in all or most of the informants' curves will now be described.

(1) The jaw was constantly moving during the utterances. A permanent oscillating pattern with peaks and valleys characterized all utterances by all speakers. The vertical movement amplitudes were on average 5 mm and maximally 8-12 mm. A mandibular period had a mean duration of about 0.230 sec.

(2) The speed of the mandibular lowerings and raisings tended to be constant, resulting in symmetric peaks and dips in the curves.

(3) The mandibular movement periodicity showed a strong connection with the syllabic structure. Thus, there was a strong tendency for one dip to occur in each vowel and one peak in the consonant(s) in between. Generally, there was no cases of a dip connected with a single consonant or a peak with a single vowel.

However, in less than 10% in the whole material, a VCV sequence corresponded to a single curve valley. This dip was typically deeper and longer than the average. In most cases, this C was [ɳ] or [n] in phrase final VCV in *kopparna* (substudy A), and [l] in various positions in the two phrases in substudy B. The durations of these consonantal segments, measured in the waveforms, were normal.

In addition, there were a few cases of two small curve dips in a single vowel. These were only found in some speakers in the stressed, long [ɑː] in *tal* and *tjata*.

(4) In several sequencies with high vowels, mandibular oscillation amplitude was very small, only 1-2 mm, cf Fig 2. This was found for several speakers in *Mississippi, Prinsessan sätter potatis,*

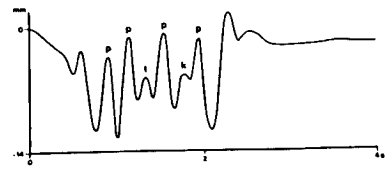


Figure 1. Mandibular movement curve of subject A reading Pappa tappar kopparna.

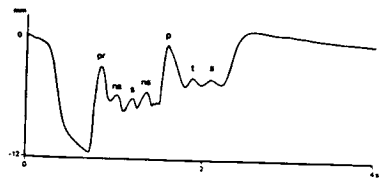


Figure 2. Mandibular movement curve of subject H reading Prinsessan sätter potatis.

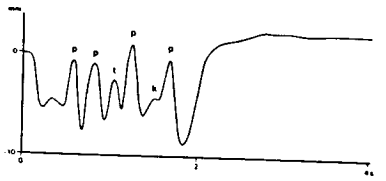


Figure 3. Mandibular movement curve of subject D reading Pappa tappar kopparna.

and in the fat type part of *Statsministerns sista tal måste läsas.* Thus, the permanent jaw oscillation, correlated with the syllabic structure, was sustained on a very small amplitude scale here.

(5) Most often, the mandibular movement pattern immediately before speech onset and in the pause between the two consecutive phrases was different from the undulating speech mode, cf Fig 2. However, in some cases an integrated oscillation of the speech type was found also in these speech-outside positions. See Fig 1 and 3. Thus, sometimes the oscillating jaw movement was started one or two periods before the speech onset and was also sustained during short silent pauses.

DISCUSSION

Method

The Selspot diode placing on the chin led to an artifactual deviation in labial consonant jaw amplitudes. Due to tissue stretching, our amplitude values were too high in these sounds, higher than in dental consonants. According to other studies with no such artifact, e g [6], alveolar/dental consonants have somewhat higher mandibular positions than labial ones. To avoid this problem, we have tested diode placings on the lower front teeth. This position is however also unsuitable, since the diode emitted light is often absorbed by the lower lip. In future investigations, we shall therefore place the diode on a small, rigid, bent metal stalk glued to a lower teeth.

In spite of this, our present data are valid for an analysis of several aspects of the close mandibular-syllabic relation.

The mandibular-syllabic connection

Obviously, speech is characterized by a quasi-regular, continuous jaw oscillation. No mandibular steady state was seen in 360 phrases uttered by 22 speakers. The movements also tended to be symmetric. This is unlike e g the tongue, which moves in a more irregular way, with intermittent steady states [9].

Furthermore, the permanent mandibular undulating pattern had a very close correspondence with the syllabic pattern of the utterances. The jaw opening movements were connected with the vowels and the closing movements with the consonants.

This undulating pattern was seen also in several sequences with high vowels, where oscillation amplitudes were only 1-2 mm, cf Fig 2. This is highly significant. The production of the individual consonants and vowels in such sequences do not demand such mandibular precision. Obviously, the very small movements have another function. Our hypothesis is that the never-resting, oscillating mandibular movements constitute the basic syllabic structure of speech.

Other articulators, e g the tongue, sometimes display steady states [9]. Therefore, it was a priori not unreasonable to suppose that the lower jaw should be kept still in a sequence of sounds like *Mississippi*, where [s] and [i] have similar, close jaw positions.

According to our hypothesis, the main reason for these permanent, oscillating movements of the lower jaw is its basic role as prime mover in connection with syllabic structure. It is however not a necessary mover. The syllabic structuring of speech is of course lead by centers in the brain. Other articulatory tools can take the place of the jaw. But this is unusual. It is also true that the mandibular-syllabic connection is not completely one-to-one. But for about 90% of over 4000 syllables in our material, there was a one-to-one connection with mandibular periods.

Almost all cases of exceptions from this general one-to-one correspondence consisted of one opening-closing mandibular movement during two consecutive syllables. (No cases of three syllables in one mandibular period was found.) The duration of the consonant C in the bisyllabic mandibular period - VCV - was normal. Since the composition of the phrases used was restricted, definite conclusions cannot at this stage be drawn about the factors that condition these reductions. However, in almost all cases, the C was a front tongue sonorant - [n, ɳ] or [l], and one or both of the Vs an open vowel - [a] or [ɛ:] - with very low jaw position.

Obviously, an urgent research task is to systematically map the conditions for these reductions. The aim of this research is to construct a model of speech syllabicity. This model must also integrate some other data. One of these is some few cases in our material of two mandibular oscillation periods within one syllable. These cases were only found in [ɑ:] in a stressed syllable. Since [ɑ:] has the longest inherent duration universally, and also in Swedish, this is more natural than if the double dip had been found in other vowels.

This double-period mandibular trajectory within a long vowel gives the impression of the lower jaw as an oscillator, moving within some frequency limits, conditioned by its size and shape. Most often, the syllable durations fit this periodicity. This oscillator concept is not new in speech articulation research [10].

Also some other of our data support the hypothesis that the mandibulum is an articulatory oscillator. They also indicate that it may be the prime mover, in accord with which the tongue and other articulators move. These data are the cases of integrated pre-phrasal jaw oscillation. Some of these cases consist of two periods, with either equal amplitudes (Fig 3) or a very small first dip (Fig 1), giving the impression of an oscillator softly starting. Also a number of cases of permanent, integrated mandibular oscillations within silent pauses between phrases support this view.

Syllables arise in baby babble around the age of six months, making it more speechlike. The cause of this is the appearance of regular lower jaw opening and closing movements [11]. In this speechlike babble, mandibular movements seem always to have a one-to-one connection with the syllables [11]. In grown-ups - and also in children aged 5-7 years [12] - this one-to-one connection is somewhat modified by some reductions mainly, but basically the same.

REFERENCES

[1] Jespersen, O. (1920), *Lehrbuch der Phonetik*, 3 Aufl, Berlin

[2] de Saussure, F. (1916) *Cours de linguistique général*

[3] Menzerath, P. & de Lacerda, A. (1933), *Koartikulation, Steuerung und Lautabgrenzung*, Berlin und Bonn: Ferd Dümmlers Verlag.

[4] Fujimura, O. (1980), "Modern methods of investigation of speech production", *Phonetica*, vol 37, pp. 38-54.

[5] Sonoda, Y. (1987), "Effect of speaking rate on articulatory dynamics and motor event", *Journal of Phonetics*, vol 15 pp. 145-156.

[6] Keating, P., Lindblom, B, Lubker, J. & Kreiman, J. (1995), "Variability in jaw height for segments in English and Swedish VCVs", *Journal of Phonetics*, vol 22, pp. 407-422.

[7] Karlsson, S. & Carlsson, G. (1989), "Recording of masticatory mandibular movements and velocity by optoelectronic method", *The International Journal of Prosthodontics*, 2:5 pp. 490-496.

[8] Lindblad, P., Karlsson, S. & Heller E. (1991), "Mandibular movements in speech phrases - a syllabic quasi-regular continuous oscillation", *Scandinavian Journal of Logopedics and Phoniatrics*, vol 16, pp. 36-42.

[9] Wood, S. (1991), "X-ray data on the temporal coordination of speech gestures", *Journal of Phonetics*, vol 19, pp. 281-292.

[10] Tuller, B., Kelso, J. & Scott, A. (1984), "The timing of articulatory gestures: Evidence for relational invariants", *Journal of the Acoustical Society of America*, vol 76, pp. 1030-1036.

[11] MacNeilage, P. & Davis, B. (1991), "Vowel-consonant relations in babbling", *Proceedings of the Twelfth International Congress of Phonetic Sciences*, Aix-en-Provence, vol 1, pp. 338-343.

[12] Dellborg, H. (1992), *Käkrörelser i barns tal*, Exam. paper in logopedics, Dept of logopedics and phoniatrics, Sahlgrenska sjukhuset, Göteborg, Sweden

VISIBLE ARTICULATORY CHARACTERISTICS OF THE ITALIAN STRESSED AND UNSTRESSED VOWELS

Emanuela Magno Caldognetto, Kyriaki Vagges, Claudio Zmarich
Centro di Studio per le Ricerche di Fonetica, CNR
Via Anghinoni, 10 35121 Padova, Italy

ABSTRACT

This research focuses on the study of the multidimensionality of the visible articulatory movements in the production of the Italian stressed and unstressed vowels. Lip and jaw movements were recorded and analysed with a fully automatic real-time system for 3D kinematics data acquisition. The data obtained show that *jaw opening* and *lower lip protrusion* were the most relevant articulatory parameters in distinguishing among the vowels.

INTRODUCTION

Integration of articulatory data with acoustic and perceptual data is fundamental in developing a phonetic theory of vowels and can find important applications in linguistic description [1,2] (cross-language comparison, articulatory features system), psycholinguistic research [3] (lip reading, bimodal perception) or technological applications [4] (audio-visual speech synthesis and recognition systems).

The Italian vowel system is considered to be very simple [1] since it may be described in terms of the features *high / low* and *front / back*. The feature *rounded* cooccurs with the feature *back* and *protrusion* always cooccurs with *rounding*.

The aim of this research is to define the visible articulatory parameters in the production of the Italian stressed and unstressed vowels, to individuate the cooccurrence of various parameters and finally to evaluate the relation of these parameters to the phonetic linguistic features.

METHOD

Lip and jaw movements were recorded and analysed with ELITE [5], a fully automatic, real-time system for 3D kinematics data acquisition. It uses small, non obtrusive, passive markers of 2mm of diameter, realised by reflective paper, attached onto the speaking subject's face. In this study the markers were placed on the central points of the vermilion border of the upper lip and of the lower lip, at the corners of the lips, and at the centre of the chin. The markers placed on the tip of the nose and on the lobes of the ears served as reference points to eliminate the effects of the head movement. The following articulatory parameters corresponding to phonologically significant features were analysed:

- *lip height* (LH), calculated as the distance between the markers placed on the central points of the upper and lower lips; this parameter may be correlated with the feature *high/low*.
- *lip width* (LW), corresponding to the distance between the markers placed at the corners of the lips, a parameter which correlates with the feature *rounded / unrounded*.
- *jaw opening* (JO), corresponding to the distance between the markers placed at the centre of the chin and the tip of the nose. This distance is primarily due to the jaw opening but it is also influenced by the movement of the skin of the chin. This parameter is correlated with the feature *high / low*.
- *anterior - posterior movement* of the *upper lip* (UP) and *lower lip* (LP), calculated as the distance between the markers placed on the central points of either the upper or lower lip and the line passing from the markers placed on the lobes of the ears. This parameter correlates with the feature *protruded / retracted*.

The visible articulatory movements of 6 subjects (4 females and 2 males), speakers of northern Italian, were recorded and analysed. All the subjects were university students, aged between 19 and 22 and were paid volunteers. They repeated five times, in random order, each of the 7 stressed /a/, /ɛ/, /e/, /i/, /ɔ/, /o/, /u/ and the 5 unstressed, /a/, /e/, /i/, /o/, /u/, Italian vowels. The vowels were in the first syllable of disyllabic (/'tasti/, /'tɛsi/, /'tesi/, /'tisi/, /'tɔsko/, /'tosko/, /'gusto/) or trisyllabic (/ta'stare/, /te'stare/, /ti'sane/, /to'skane/, /gu'stare/) words, and were preceded by a /t/ and followed by /s/ (with one exception, i.e. /gu'stare/). They occured within the carrier phrase "dico _____ chiaramente" (I say _____ clearly).

A synchronous recording of the acoustic signal was also obtained.

Portions of the articulatory signal corresponding to the vowel to be analysed were segmented on the basis of the acoustic speech signal. Since the dynamic aspects of the articulatory parameters were not taken into consideration in this study, a single point characterising the vowel was individuated for each articulatory parameter. The data were normalised subtracting the values related to the position of the lips and jaw at rest, from each parameter obtained, for each vowel and each subject. This assured the comparability of the results independently of the subjects variability in the shape and size of the articulators. The so obtained data correspond to the real extension of the lip and jaw movements and may also be connected to data relating to the internal borders of the lips [4].

RESULTS

The analysis of the normalised data showed the mutual relations between all the parameters examined: a clear correlation between the *upper* and the *lower lip protrusion* for both stressed and unstressed vowels, (r=.82 and r=.83 respectively), and a negative correlation between *lip width* and *upper lip protrusion* (r=-.81 and r=-.83), as well as between *lip width* and *lower lip protrusion* (r=-.73 and r=-.80). Stressed vowels showed also a correlation between *lip height* and *jaw opening* (r=.85). For both stressed and unstressed vowels there was no significant correlation between *lip height* and *lip width* and between *lip height* and *upper* or *lower lip protrusion*. The presence or absence of correlation observed in this study is congruous with previous results reported for English [6,7] and French [8] independently of the instrumentation or reperee points used in defining the parameters.

Stressed and unstressed vowels were analysed with two-way ANOVAs (7 or 5 vowels respectively and 6 subjects as a between factor) to assess their effect on each of the articulatory parameters examined. Post hoc Tukey multiple comparisons were carried out when the vowel effects were significant. Only the data significant at $p<.01$ will be discussed. The normalised mean values, pooled over the 6 subjects and the 5 repetitions for each parameter and each vowel are reported in Table 1. The values may be either positive or negative depending on the parameter taken into consideration. For example, LW values are negative when the distance between the corners of the lips decreases with respect to their distance at rest, as is evident for both stressed and unstressed /u/, while positive LW values correspond to an increased distance with respect to the values at rest, as is the case of the unrounded vowel /i/ in both stressed and unstressed position. UP and LP may also show both positive and negative normalised mean

Table 1. Normalised mean values (mm) pooled over subjects and repetitions for each articulatory parameter and each vowel.

		/i/	/e/	/ɛ/	/a/	/ɔ/	/o/	/u/
LH	stressed	8.6	9.6	13.6	15.0	15.1	8.7	7.7
	unstressed	7.6	8.7		10.6		8.1	7.1
JO	stressed	6.6	7.8	12.6	14.1	11.2	3.3	1.8
	unstressed	5.3	6.7		9.0		2.6	2.0
LW	stressed	0.1	0.5	1.1	0.9	-5.1	-5.1	-6.1
	unstressed	1.2	0.8		1.2		-3.3	-5.3
UP	stressed	-1.1	-0.3	-2.0	-2.1	2.6	3.9	4.4
	unstressed	-1.2	-1.0		-1.5		3.2	3.9
LP	stressed	-1.4	-1.2	-3.3	-2.9	0.9	2.8	3.6
	unstressed	0.6	-1.1		-1.9		2.2	3.4

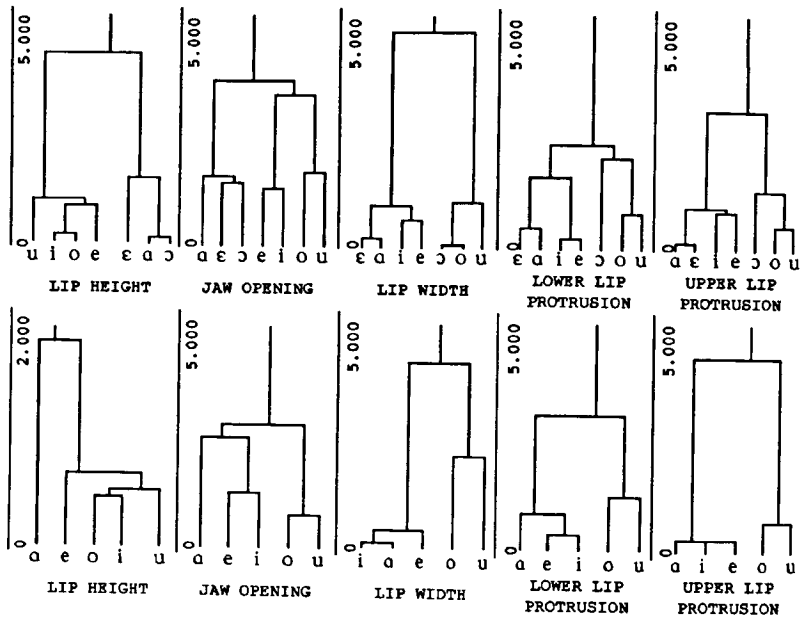


Figure 1. Hierarchical clustering of the stressed and unstressed vowels with respect to the five articulatory parameters.

values, while LH and JO are always positive.

The results of the ANOVA showed that *jaw opening* is the articulatory parameter that better distinguishes both stressed and unstressed vowels since it defines 4 degrees of jaw opening. In fact, Fig. 1 shows that stressed vowels are clustered in 4 groups, /u,o/, /i,e/, /ɛ,ɔ/, and /a/. As for the unstressed vowels, in Fig. 1, JO distinguishes /a/, /e/, /i/ and /u,o/.

LH, which is traditionally considered to be parallel to *jaw opening*, does not identify all the degrees of opening defined by JO (see Fig. 1). Moreover, the extension of its movement always shows greater values than JO, cf. Table 1. It is clear that lips not only move in synergy with the jaw, but also in an independent specific manner.

LW divides both stressed and unstressed vowels in two groups: rounded vowels and unrounded vowels, see Fig. 1.

As for the two protrusions, LP is the parameter that best distinguishes both stressed and unstressed vowels. As shown in Fig. 1, stressed and unstressed vowels are divided into 4 groups, i.e. two degrees of protrusion and two degrees of retraction. In particular, for stressed vowels, a higher degree of protrusion characterises /u/ and /o/ with respect to /ɔ/, while /a/ and /ɛ/ are more retracted than /i/ and /e/, see Table 1.

Using the parameters resulting most significant for distinguishing the vowels (*jaw opening, lower lip protrusion* and *lip width*), a three dimensional representation of the stressed and unstressed vowel space was plotted in Figs 2a and 2b respectively. As can be observed there is a tendency to reduce the values of the parameters from the stressed to the unstressed condition, even though the trend is not systematic.

DISCUSSION

Our data confirm the cooccurrence of rounding and protrusion for the Italian language. In fact, all the vowels with positive values of *lip width*, i.e. /i,e,ɛ,a/, also have negative values for both *upper*, and *lower lip protrusion*. That is, unrounded

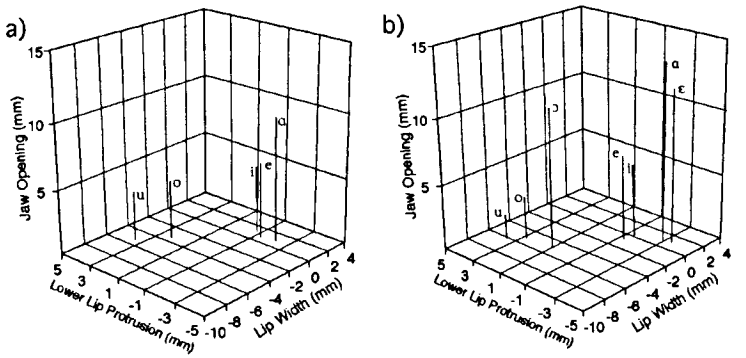


Figure 2. 3D representation of the stressed and unstressed vowel space.

vowels are always also non protruded. Similarly, vowels with negative *lip width* values, i.e. the rounded vowels /ɔ,o,u/, are characterised by positive values of *upper* and *lower lip protrusion*, that is, they are also protruded.

Jaw opening and *lower lip protrusion* are the parameters that better distinguish the vowels. It should be noted though, that differences in *jaw opening* with respect to *lip height* may be due to the marker placed on the chin: the position of this marker was influenced not only by the jaw opening but also by the movement of the skin especially during the lip protrusion.

Based on the values of the parameters analysed, the reduction of the unstressed with respect to the stressed vowels was confirmed. Moreover, the unstressed mid vowels are more similar to the stressed mid-high /e/ and /o/ rather than to the mid-low stressed /ɛ/ and /ɔ/.

REFERENCES

[1] Ladefoged P., & Maddieson I., (1990) *Vowels of the world's languages*, Journal of Phonetics 18, 93-122.

[2] Fischer-Jørgensen E., (1985) *Some basic vowel features, their articulatory correlates, and their explanatory power in phonology*, in Fromkin V.A. (Ed), Phonetic Linguistics. Essays in Honor of Peter Ladefoged, Academic Press, 79-99.

[3] Summerfield Q., (1987) *Some preliminaries to a comprehensive account of audio-visual speech perception*, in Dodd B. & Campbell R. (Eds.), Hearing by Eye: The Psychology of Lip-Reading, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 3-51.

[4] Benoît C., Lallouache T., Mohamadi T., & Abry C., (1992) *A set of French visemes for visual speech synthesis*, in Bailly G., Benoît C., & Sawallis T.R. (Eds.), Talking machines: Theories, Models, and Designs, North-Holland, Amsterdam, 485-504.

[5] Magno Caldognetto E., Vagges K., Pedotti A., Ferrigno G., (1994) *Parametri articolatori labiali e mandibolari nelle vocali cardinali dell' Italiano*, Atti delle III Giornate di Studio del G.F.S. :Le vocali, Padova 1992, 75-85.

[[6] Fromkin V., (1964) *Lip positions in American English vowels*, Language and Speech 7, 217-225.

[7] Linker W., (1982) *Articulatory and acoustic correlates of labial activity in vowels: A cross-linguistic survey*, UCLA, Working Papers in Phonetics 56, 1-134.

[8] Abry C., & Boe L.J., (1986) *"Laws" for lips*, Speech Communication 5, 97-104.

RELEASE RATES FOR [t] IN VCV SEQUENCES ESTIMATED FROM AERODYNAMIC DATA

Sheila J.Mair, Department of Electronics and Computer Science, The University of Southampton, Southampton SO17 1BJ, U.K.

Celia Scully, Speech Laboratory, Department of Psychology, The University of Leeds, Leeds LS2 9JT, U.K.

ABSTRACT

According to Stevens [1], "quantitative data must be obtained on rates of release and closure of articulators". Here, we use aerodynamic data in an orifice equation to estimate the rate of increase in the cross-sectional constriction area for [t] in different vowel contexts for 10 English speakers. Analyses of the results indicate that in most cases, the rate of release of [t] is significantly faster when an open vowel follows than when a close vowel follows.

RECORDINGS

Some data were recorded in 1987/88 for 10 adult speakers of Received Pronunciation English as part of the Alvey Project MMI 009, Speech Pattern Algorithmic Representation, and are referred to henceforth as the SPAR database. The speakers are: HB, JH, SR, GB, EA (female) and JM, MA, DH, JW, MB (male).

The recording sessions were carried out in the Department of Linguistics and Phonetics at the University of Leeds. Four channels of data were recorded onto FM tape: sound pressure (microphone signal), laryngograph signal, volume flowrate of air, interpreted as oral airflow for non-nasal sequences (measured with a Rothenberg mask) and intraoral air pressure (measured with an orally-inserted polyethylene tube). The airflow and air pressure signals were low-pass filtered at $50Hz$ before being recorded onto a mingograph along with the other two (unfiltered) signals.

MEASUREMENTS

The speech material analysed formed part of Set C2F of the SPAR database. This consisted of repeated [pəCV] sequences where C and V stand for various consonants and vowels respectively. Sequences in which C = [t] and V = [i:, ɑ:, ɔ:, u:] were selected for analysis. Repetitions 2, 3, 4, 5 and 6 of each vowel context were analysed for each speaker. Measurements of airflow and air pressure were made at 10ms intervals following the plosive release, with the time of release defined from the rapid increase in flow from zero or near-zero. Using an orifice equation, the increasing minimum cross-sectional area of the vocal tract constriction is estimated. The equation is:

$$A_c = 0.00076 \times U_c / P_c^{0.5}$$

where A_c is the minimum cross-sectional area of the constriction (in cm^2), U_c is the volume flowrate of air through it (in cm^3/s), and P_c is the pressure drop across the constriction (in cmH_20); the orifice equation is discussed in more detail in Scully [2].

RESULTS

Graphs of constriction area against time are plotted. The graphs suggest that the increase in constriction area in the initial part of a [t] release is approximately linear, and that the release is faster in the open vowel contexts ([ɑ:] and [ɔ:]) than the close vowel contexts ([i:] and [u:]). As examples, graphs for [ti:] and [tɑ:] are presented for Speaker HB in Figure 1.

Based on the area increase in the initial 50ms following the release, rates of release are calculated for each repetition. Release rates (with means and standard deviations) are presented for the different vowel contexts for each subject in Table 1.

A one-way analysis of variance indicates that there is a very highly significant effect of vowel context on the rate of release of [t] for all speakers except DH and EA ($p \leq 0.001$).

DISCUSSION

Of the vowels analysed, [ɔ:] is likely to have most lip-rounding for Received Pronunciation speakers. Lip-rounding may begin during the consonant due to processes of coarticulation and so there may be a significant pressure drop across the rounded and protruded lips. In such a case, the pressure drop across the alveolar constriction may be less than the measured intraoral air pressure suggests. Therefore constriction area values calculated with the orifice equation for this vowel context may be under-estimating the actual values.

The release rates calculated here are generally consistent with the range of $5-20cm^2/s$ estimated by Fant [3] from acoustic analyses of formant transition patterns studied from spectrograms of plosives.

Massey [4] estimated a typical release rate of $100cm^2/s$ for labial and alveolar plosives (compared to $25cm^2/s$ for velar plosives). This value seems rather high compared to the results here, even for the open [ɑ:] vowel context.

Measurements of X-ray data for [t]

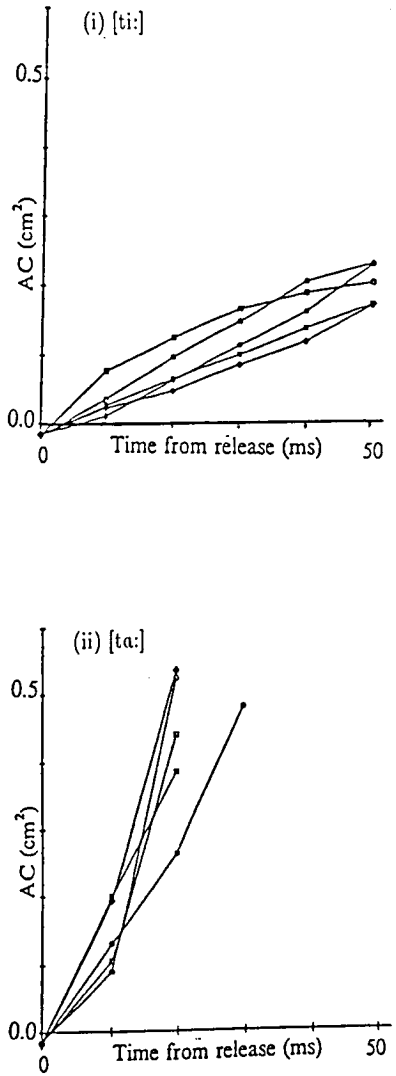


Figure 1: *Graphs of release of (i) [ti:] and (ii) [tɑ:] for Speaker HB.*

Table1. Release rates (in cm^2/s) estimated from aerodynamic data for [t] preceding different vowels, for the 10 SPAR speakers.

	JM	MA	DH	JW	MB	HB	JH	SR	GB	EA
ti:2	2.5	5.4	5.1	2.6	7.1	4.9	3.6	3.8	4.4	2.7
ti:3	4.8	5.5	5.1	3.3	6.5	3.7	3.4	3.0	4.5	1.9
ti:4	4.9	6.7	5.7	3.1	7.8	3.7	3.6	2.9	4.3	2.6
ti:5	3.6	5.5	4.3	2.8	8.3	4.9	3.2	3.4	2.7	3.1
ti:6	4.1	6.4	3.3	1.1	7.1	4.3	3.5	3.2	5.4	3.0
Mean	4.0	5.9	4.7	2.6	7.4	4.3	3.5	3.3	4.3	2.7
St.Dvn.	0.98	0.60	0.93	0.87	0.70	0.60	0.17	0.36	0.98	0.47
	JM	MA	DH	JW	MB	HB	JH	SR	GB	EA
ta:2	9.0	12.2	5.4	2.1	12.1	16.1	24.6	7.3	14.4	7.3
ta:3	4.7	13.6	5.6	8.1	11.8	19.4	36.1	6.7	13.7	4.9
ta:4	10.8	14.6	2.8	9.0	17.0	26.9	18.1	6.8	10.7	3.7
ta:5	7.1	16.0	2.2	5.6	13.3	26.3	19.4	7.0	8.6	3.3
ta:6	2.8	11.4	3.9	7.5	13.0	22.1	21.6	8.2	8.1	5.7
Mean	3.9	13.6	3.4	6.5	13.4	22.2	24.0	7.2	11.1	5.0
St.Dvn.	3.21	1.84	2.05	2.74	2.08	4.58	7.22	0.60	2.87	1.61
	JM	MA	DH	JW	MB	HB	JH	SR	GB	EA
tu:2	1.2	9.2	2.5	1.6	8.9	3.8	4.1	4.0	5.8	4.1
tu:3	1.1	8.3	4.3	1.5	8.1	4.7	5.7	4.3	5.6	3.7
tu:4	2.0	9.1	3.9	3.3	9.4	5.5	6.3	4.4	5.3	2.3
tu:5	0.6	9.0	4.0	3.7	12.0	5.7	7.5	3.5	5.2	3.1
tu:6	0.9	8.4	3.5	2.9	12.3	6.1	4.9	4.4	6.0	4.1
Mean	1.2	8.8	3.6	2.6	10.1	5.2	5.7	4.1	5.6	3.5
St.Dvn.	0.52	0.42	0.70	1.0	1.90	0.92	1.30	0.38	0.33	0.77
	JM	MA	DH	JW	MB	HB	JH	SR	GB	EA
to:2	7.9	7.0	5.7	16.7	12.6	18.6	6.0	6.3	10.0	6.2
to:3	9.6	8.6	5.7	13.8	10.9	19.3	6.3	7.3	7.1	5.4
to:4	10.7	9.1	6.3	9.6	7.8	13.1	4.9	6.1	7.4	5.4
to:5	10.4	8.7	5.3	7.3	11.3	15.4	12.8	7.9	6.9	4.3
to:6	10.4	9.8	5.4	5.5	11.3	15.3	8.7	7.9	8.0	4.1
Mean	9.8	8.6	5.7	10.6	10.8	16.3	7.7	7.1	7.9	5.1
St.Dvn.	1.14	1.03	0.39	4.62	1.79	2.6	3.15	0.86	1.26	0.87

for a male speaker of North-American English have demonstrated that the velocity of tongue movement following consonant release is "dependent on the target configuration of the following vowel" [5]. Those articulatory data are consistent with our aerodynamically-derived constriction area estimates, which have suggested that the articulatory release of the English plosive [t] in VCV sequences is faster when an open vowel follows than when a close vowel follows.

CONCLUSIONS

In the orifice equation, the measured intraoral air pressure is actually the pressure drop across the constriction, the teeth and the lips, and so the results do not necessarily indicate an actual single constriction of the vocal tract. However, the constriction area estimates derived from the aerodynamic equation do indicate consistent effects for a [t] release in different vowel contexts (faster when an open vowel follows than when a close vowel follows) and these are likely to produce consistent effects in the corresponding acoustic signal.

The shape of the vocal tract constriction and its position along the vocal tract length will also have acoustic effects which are manifest throughout the transition to a following vowel [6]. Both these parameters are likely to vary for [t] in different vowel contexts. New methods for gathering articulatory data, such as enhanced electropalatography [7], could provide invaluable information about the three-dimensional shape of the vocal tract constriction.

Simultaneous recordings of articulatory, aerodynamic and acoustic data could help our understanding of the mapping between all these different aspects, and of the enormous complexities involved in speech.

ACKNOWLEDGEMENTS

Thanks are due to Eric Brearley in Leeds for his help with the data acquisition, and to the 10 SPAR speakers. This research was supported in part by a SERC studentship award.

REFERENCES

[1] Stevens, K.N. (1991) "The contribution of speech synthesis to phonetics: Dennis Klatt's legacy", plenary lecture in the *Proceedings of the 12th ICPhS*, Aix-en-Provence, Vol.1, pp.28-37.

[2] Scully, C. (1986) "Speech production simulated with a functional model of the larynx and the vocal tract", *Journal of Phonetics*, Vol.14, pp.407-413.

[3] Fant, G. (1970) *Acoustic Theory of Speech Production*, Mouton, The Hague, second edition, p.199.

[4] Massey, N.S. (1994) *Transients at stop-consonant releases*, unpublished Masters dissertation, Massachusetts Institute of Technology.

[5] Perkell, J.S. (1969) "Physiology of speech production: results and implications of a quantitative cineradiographic study", *Research Monograph No.53*, MIT Press, Cambridge, Massachusetts, p.16.

[6] Maeda, S. (1987) Articulatory-acoustic relationships in unvoiced stops - a simulation study", *Proceedings of the 11th ICPhS*, Tallin, Estonia, USSR, Vol.5, pp.11-14.

[7] Chiu, W.S.C. and Shadle, C.H. (1992) "Use of palate shape data in an enhanced electropalatographic system", *Proceedings of the Institute of Acoustics*, Vol.14, Part 6, pp.415-422.

ARTICULATORY STRATEGIES FOR THE PRODUCTION OF /l/

A. Marchal, M. Chafcouloff and S. Lapierre
CNRS URA 261 Parole et Langage
Université de Provence, Aix-en-Provence, France

ABSTRACT

The spatio-temporal organization of lingual "gestures" for the production of /l/ is investigated in nonsense words, words and sentences. Our data reveals important differences across speech items and across speakers. The phasing of the gestures indicates that our speakers adopt different production strategies in the various contexts.

INTRODUCTION

Until recently, most of the studies on speech production have relied upon acoustic and articulatory data from nonsense words. This type of speech material allows for a fine control of linguistic and prosodic variables which interact in a speech sequence, but it is questionable whether results obtained from these carefully designed experiments bear any signifiance for the understanding of the process involved in the production of other speech items such as words, sentences... The spatio-temporal organization of lingual "gestures" for the production of /l/ is compared across speech items and across speakers.

METHODOLOGY

Data for this study has been extracted from the multilingual EURACCOR database [1]. This database consists of simultaneous digital recordings of the acoustic soundwave, of the laryngograph signal, of oral and nasal airflow and of linguo-palatal contacts. Multisensor data has been collected for the production of VCV nonsense words, isolated words matching phonetically the nonsense words, and the same words embedded in sentences. The speech items have been repeated 10 times at a normal rate. We have analysed here the production by two French female speakers ("ad", "gc"; 20 ~ 25 years old; no sociogeographic marks of pronunciation or speech defect) of the sequence /ulu/ in the nonsense word " oulou " and in " Toulouse ", the french town, as isolated word and " Toulouse " embedded in the sentence: " La cousine de Vichy épousa un hippie à Toulouse ".

The various acoustic, aerodynamic and articulatory signals available in the ACCOR database are annotated independently [2]. The present study relies on EPG data only. The following landmarks have been identified: the onset of the forward movement of the tongue, the lateral closure, the maximum constriction and the lateral release. They are annotated respectively as ACE, LCE, MCE and LRE. In addition, the beginning and the end of lingual activity is labelled GOE and GEE. The data from these annotation points is used as the basis for the subsequent spatial and temporal analyses. EPG patterns at ACE, LCE, MCE and LRE have been analysed as an indication of the amplitude of the tongue tip gesture. The temporal organization of the gestures is given by the durations between these marks. They correspond to the following phases: approach (ACE-LCE); closure (LCE-MCE) and release (MCE-LRE). EPG data has been statistically analyzed using the paired t-test and the ANOVA linear regression method.

RESULTS

Spatial Organization

On a hyper- to hypo-continuum and in the framework of the H & H theory [3], the following prediction can be made: the amplitude of the lingual gesture is expected to be larger for the nonsense words than for real words, and it should be the smallest for the sentence context

[4]. The amplitude of the tongue gesture can be estimated from the number of activated electrodes. Measurements were made at the point of maximum contact MCE and at ACE and LCE for the three speech types: Total number of linguo-palatal contacts, number of contacts by rows and by palatal areas (A=alveolar; B=prepalatal; C=palatal; prevelar, as shown in Fig. 1):

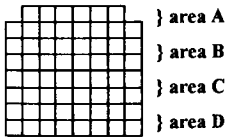


Figure 1. Area delimitations of the EPG frame

As far as the general spatial organization is concerned, two remarks can be made: 1) There is no significant difference between the number of linguo-palatal contacts between nonsense words, words and sentences for both speakers in each context. This means that they both reach similar spatial targets in terms of general amplitude of the lingual gesture.; 2) Concerning the contexts, the difference which is observed between the mean contact number for each articulatory landmark is more important for the nonsense words than for the other speech contexts. For example, the mean difference between the number of contacts of ACE and LCE in sentences for "gc" speaker is 7.3 contacts against 11.7 in nonsense words (23.3 cts - 16 cts against 23.1 - 11.4 cts), contrary to Farnetani [5], these differences are however not significant.

When we consider the number of contacts in the various palatal areas, the same tendency can be observed for the alveolar and prepalatal regions. We note that the gesture amplitude for the nonsense word is not significantly different from the other contexts, but it is suggested that the nonsense context differs most from the sentence context and suggest the following decreasing order of gesture amplitude: nonsense word/word/sentence contexts.

Temporal Organization

The first part of the temporal analysis consisted in comparing the total duration of /l/ as a function of the given contexts (Fig.2). The ANOVA analysis of variance indicates a significative difference between the total duration of the articulation for three contexts, F(27.2), p <.001 for "ad" speaker, F(26.7), p <.001 for "gc" speaker. As could have been expected, we observe the shortest duration for /l/ in the sentence, then by increasing order in the real word and in the nonsense word.

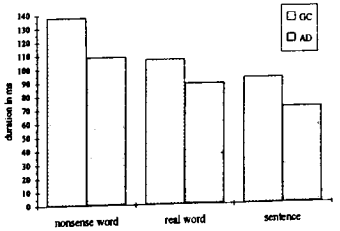


Figure 2: Mean total duration of /l/ in French for speakers "ad" & "gc" from EPG data

The question arises to know if the variation of the total duration which was observed as a function of context affects equally or not each phase. If the ratio duration of phases / total durations of /l/ is kept constant, this would imply that the internal organisation of the various gestures involved in the production of /l/ is not altered as a function of the context. The duration of each phase is proportionally increased or decreased from sentences to nonsense words. Since the amplitude is not affected (previous observations), this would imply that there exists a saturation effect and that the intended lingual gesture is in fact masked by competing demands on the articulator. An alternative hypothesis would explain the observed facts as an internal

reorganisation of the various phases. This is indeed what our analysis reveals for speaker "ad": The relative duration of phase 1 and phase 2 is primarily concerned. There is a clear shortening of the approach to constriction in the sentence context for speaker "ad" but not for speaker "gc".To illustrate this difference in articulatory organization, we have adopted a phase representation. This representation (Fig.2 & 3) uses phase/total duration ratio. Each angle indicates the relative timing of each phase translated into degrees. The circumference corresponds to the total duration of the consonant.

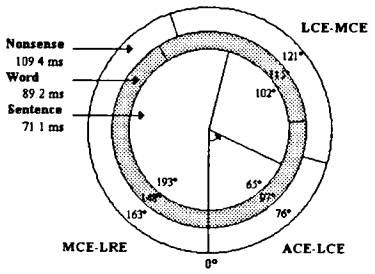


Figure 2. Articulatory phases for /l/: speaker "ad"

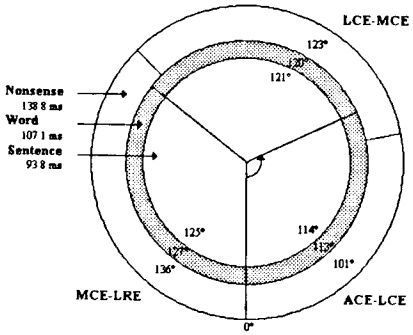


Figure 3. Articulatory phases for /l/: speaker "gc"

DISCUSSION AND CONCLUSION

We have to be extremely careful with the interpretation of these results, since the data presented here is based only on the production of 2 speakers who produced the speech items 10 times. However the fact that they differ in the reorganization of the lingual gestures across speech material contexts is in itself interesting and raises the following questions which should be addressed in a broader study: Does this apparent difference of strategy in the lingual gesture reflect a real difference in the tongue kinematics or does the EPG technique which shows only the contacts biaise the data ? The production of /l/ requires not only a lateral closure, but also a specific shape of the tongue behind the contact area. The curvature of the tongue is in part responsible for the turbulent flow conditions needed for /l/. The form of the cavity behind the closure depends also on the shape of the palate. The plastercasts of our 2 speakers show large morphological differences. The palate for "gc" is more flat than the palate of "ad".

It seems very important in addition to linguo-palatal contact patterns to obtain data on the distance from the tongue to the palate. As a first attempt to answer this question, we will record simultaneously EPG and EMMA with coils on the tongue dorsum.

Concerning the use of nonsense word material for the investigation of speech production, our EPG data would suggest a positive answer for one speaker and a negative one for the other speaker. For both speakers, there is a general tendency to shorten the articulation of /l/ in sentences with respect to words and nonsense words. However, the timing of the various phases differ from one speaker to the other. The decrease is proportionnal for "gc". In that case, results from nonsense words can be extrapolated to /l/ in sentences. This is not true for speaker "ad" where a reorganization of the timing of the approach to constriction phase can be observed. Further multisensor investigation of articulatory gestures is still needed to indicate how real is lab speech.

ACKNOWLEDGMENTS

We acknowledge financial support from The European Economic Commission DGXIII under the auspices of ESPRIT Program, Action 3279 and Action 7098.

REFERENCES

[1] Marchal, A. and W.J. Hardcastle (1993) "ACCOR: Instrumentation and Database for the Cross-Language Study of Coarticulation", Language and Speech, 2-3, pp. 137-153.

[2] Marchal, A., Nguyen, N. and W.J. Hardcastle (1995) " Multitiered Phonetic Approach to Speech Labelling", in C. Sorin, J. Mariani, H. Meloni and J. Schoentgen (Eds), *Levels in Speech Communication*, Elsevier, Amsterdam, pp. 149-157.

[3] Lindblom, B. E. F. (1990) Explaining phonetic variation: A sketch of the H&H theory. In W. J. Hardcastle and A. Marchal (Eds), *Speech Production and Speech Modeling*, Kluwer Academic Publishers, Dordrecht, pp. 403-440

[4] Farnetani, E. (1991) Coarticulation and reduction in coronal consonants: comparing isolated words and continuous speech. In *Quaderni del centro di studio per le ricerche di fonetica*, **X**, 104-107. Padova: Edizioni Librera Progetto.

[5] Farnetani, E. & A. Provaglio (1991) Assessing variability of lingual consonants in italian. In *Quaderni del centro di studio per le ricerche di fonetica*, **X**, 118-127. Padova: Ediziono Libreria Progetto.

AN ARTICULATORY DESCRIPTION OF CLICKS BY MEANS OF ELECTROMAGNETIC ARTICULOGRAPHY

G. Scharf[1], I. Hertrich[1], J. Roux[2], G. Dogil[3]
1 Department of Neurology, University of Tübingen, Germany
2 Research Unit for Experimental Phonology, University of Stellenbosch, South Africa
3 Institute of Computational Linguistics, University of Stuttgart, Germany

ABSTRACT

In this paper an articulatory description of two types of clicks which are used in the Bantu language Xhosa is presented. Although in some utterances the coils caused inadequate affrication, the main elements of the click sounds could be traced, e.g. the backward movement of the tongue body during the occlusion of [!] followed by a very fast downward movement of the tongue tip after the anterior release attaining a maximum velocity of more than three times the velocity for the tongue tip opening gesture of [t].

INTRODUCTION

Previous work on Southern African click sounds was based mainly on perceptive and acoustic analyses. So far there are only relatively few physiological data. Traill (cf. [1]) presents cineradiographic and electropalatographic data which allow for a detailed description of the articulation and a corresponding classification of the different click types. Electromagnetic Articulography (EMA) has not been used before to analyse the articulatory movements during the production of click sounds.

In producing a click sound the tongue has to form two closures: at the front and at the back of the tongue so that a body of air is enclosed in between. By a downward movement of the mid-tongue while tongue tip and dorsum maintaining contact to the roof of the mouth the air in the cavity is rarefied and a suction effect is created. Now the forward closure is released and air rushes into the mouth, producing a click sound. Then the backward closure is released (called the *click accompaniment*, cf. [2]) which does not always produce a perceptible sound or noise.

The Bantu language Xhosa uses three types of clicks which can be classified (according to the terminology of [2]) as the dental click, phonetically transcribed as [|], and the (alveo-) palatal click [!] and the lateral click [||]. However, there is a controversy on the exact places of articulation. In the present study we only examined the dental [|] and the (alveo-) palatal click [!] because Electromagnetic Articulography permits the registration of movements in the mid-sagittal plane of the oral tract only so that it was not possible to investigate the movements with laterals.

The present study is a first attempt to use EMA for describing the sequence of events during the production of click sounds.

METHODS

EMA allows for registration of movements inside the oral cavity with reasonable spatial and temporal resolution. Thus the high articulatory velocities occurring during click sound production can be registered.

In the present experiment 5 sensor coils were placed on the following positions: two reference coils on the nasion and below the lower incisors to record the movements of the head and the jaw as well as three coils on the tongue: 5 mm behind the tip of the tongue (TT), in the place of articulation of [k] in [aka] (determined by a colouring test) on the tongue dorsum (TD) and in the middle betweeen TT and TD. The kinematic recordings were made with a sampling rate of 200 Hz. Simultaneously to the articulatory recordings the acoustic signal was digitally recorded.

In addition to the click sounds the alveolar stop [t] and the dorsal stop [k] were used as reference. Every target sound was embedded in a nonsense syllable VCV (with V = [a]) and had to be produced in a target phrase: *Ndithi a_a ngoku* (I am saying a_a now). Ten tokens of each stimulus were visually presented in randomized order. Since the subject had some initial difficulties producing the palatal click with the coils attached to the tongue the sentence with the palatal click target was produced another 20 times. The sentences were produced by one male subject who learned Xhosa as a third language and has been staying in a Xhosa speaking area for the last 23 years (one of the authors, J.R.).

Since in some utterances speech was obviously disturbed by the coils leading to inadequate affrication an auditive assessment was carried out by the speaker and only those tokens were selected for further analysis which clearly sounded like adequate click sounds. In addition some utterances had to be rejected because the TD coil which obviously interfered with the click production got loose. In the end 6 tokens of [aka] and 7 tokens of [ata], [a|a] and [a!a] were included in the kinematic analysis.

ARTICULATORY DESCRIPTION

The data is not sufficient for a detailed quantitative analysis. However, a close examination of the x/y-trajectories of the tongue sensors together with the acoustic signal during the production of the sequence [a!a] provides the following articulatory description: All three tongue sensors moved simultaneously from the back, low position for the first [a] upwards and to the front and got in contact to the anterior and posterior parts of the palate. In the middle of the occlusion the highest position of all three tongue sensors was achieved. During the last part of the closure the tongue dorsum started moving downward while the mid-tongue was pulled further up and further back, moving behind the position of the tongue dorsum, which must be interpreted as a retroflex movement. For this type of click a retroflex articulation has been reported elsewhere [2]. From its highest position at the palate the tongue tip then was released and performed a very fast (vertical) downward movement attaining a peak velocity of up to 1094 mm/sec. The release allowed air rushing into the mouth producing the characteristic click sound. After that the dorsal part of the tongue moved downward with a maximum velocity of 470 - 630 mm/sec. In contrast, the stop of the dorsal plosive [k], produced by the same subject, was released with a peak velocity of 149 to 200 mm/sec and the alveolar plosive sound [t] with 290 to 442 mm/sec.

DISCUSSION

By the application of Electromagnetic Electrography for the registration of the articulatory movements during the production of dental and palatal click sounds some technical problems appeared: Some click sounds were erroneously affricated because the articulation was deranged by the coils attached on the tongue. Furthermore, in some utterances the strong friction of the tongue against the palate removed the posterior receiver coil so that these recordings could not be analysed. Besides these technical problems the subject produced slightly hyperarticulated speech with slow overall speech rate, inter-word pauses and long stop closure times. Thus it might be the case that the observed click sounds were hyperarticulated, too. Nevertheless, the basic sequences of events could be traced and the high velocity peaks occcurring in click sound production could be registered by means of the applied method.

REFERENCES

[1] Traill, A. (1985), *Phonetic and phonological studies of !Xóo -bushman*, Quellen zur Khoisan-Forschung, Hamburg: Helmut Buske.
[2] Ladefoged, P. & A. Traill (1994), Clicks and their accompaniments, *Journal of Phonetics*, vol. 22, pp. 33-64.

CONTEXTUAL INFLUENCES ON DEVOICING OF /Z/ IN AMERICAN ENGLISH

Caroline L. Smith
University of California, Los Angeles, USA

ABSTRACT

The devoicing of /z/ by speakers of American English was examined in a variety of sentence contexts using acoustic, airflow and EGG data. Although speakers differed in overall frequency of devoicing, they showed similar rank orderings for frequency of devoicing in different contexts. Both the immediate phonological context of /z/ and the prosodic strength of its position in the word influence the likelihood of devoicing.

BACKGROUND

Speakers of English often do not fully voice obstruents that are phonologically categorized as voiced. Voiced fricatives are often considered to require particularly precise conditions in the vocal tract: subglottal pressure must be higher than oral air pressure in order to produce voicing, but oral air pressure needs to be higher than atmospheric pressure to produce turbulence at the supralaryngeal constriction [1]. The term "devoicing" is used here to describe an absence of vocal fold vibration in the production of sounds normally categorized as voiced.

A number of previous studies have shown that devoicing is common in voiced fricatives in both British [2, 3, 4] and American English [5, 6, 7]. This experiment uses instrumental data to investigate two questions about the voicing of /z/ in connected speech. (1) In what environments is /z/ most likely to be devoiced? (2) How does devoiced /z/ differ from /s/?

Mechanisms of devoicing

Most previous studies have used acoustic data to identify the presence or absence of voicing. The temporal characteristics of devoicing in fricatives have been documented extensively, particularly by [3] and [7]. A greater likelihood for devoicing a voiced fricative when it is adjacent to a voiceless sound or silence has also been noted, suggesting that a kind of assimilation in voicing state is at work ([3, 4, 6, 7]). There is less information available on the physiological mechanisms involved in devoicing, such as whether devoicing is the consequence of a controlled opening movement of the glottis or is a passive consequence of the aerodynamic conditions that Ohala [1] suggests make voiced fricatives difficult to produce.

There is some evidence that the glottis does open during devoiced fricatives. Haggard [2] concludes, on the basis of F0 fall such as occurs following voicelessness, that the glottis opens during a voiced fricative in which glottal vibration ceases and then is re-initiated. In a study using transillumination [5], the majority of tokens of voiced fricatives showed evidence of glottal opening, whereas voiced stops mostly did not.

DATA COLLECTION

The present experiment was designed to investigate the devoicing of /z/ in a variety of phonological environments in natural speech. Speakers of American English produced 4 to 6 repetitions of 19 sentences. In these sentences, /s/ and /z/ occurred in contexts matched for type of neighboring sounds and position in word or phrase; the matched pairs of /s/ and /z/ occurred in different sentences.

Speakers wore a pneumotachographic mask to measure airflow and an electroglottograph (EGG) to measure vocal fold contact. These signals and the acoustic signal from a head-mounted microphone were recorded directly to disk at an 8000 Hz sampling rate. The airflow and EGG signals were low-pass filtered at 1000 Hz, the acoustic signal at 3000 Hz. A tape recording was also made, and digitized at 20000 Hz for acoustic analysis. Data from three speakers is reported here.

The EGG signal was used to identify where voicing was present in fricatives. The amplitude of one EGG cycle (maximum – minimum during one excursion) was measured at time of maximum RMS energy in the vowel preceding the fricative. The fricative was considered voiced during the portion of its duration that the amplitude of the EGG cycles exceeded one-tenth of the EGG cycle amplitude at the time of maximum energy in the preceding vowel. Voicing was considered to cease when the amplitude of an EGG cycle fell below this criterion. For each token of /z/, the percentage of fricative duration with voicing was calculated by dividing the duration of frication during which the EGG amplitude exceeded criterion by the total duration of acoustic frication.

The tokens of /z/ were categorized according to the percentage of their duration during which there was voicing. The three categories were:

0 - 25% voicing	devoiced
26 – 90% voicing	partially devoiced
90 - 100% voicing	voiced

Each category was analyzed separately and its acoustics and aerodynamics were compared with tokens of /s/ produced in matched phonological contexts.

DIFFERENCES BETWEEN /S/ AND /Z/

For each speaker, each category of tokens of /z/ was compared to an equivalent number of tokens of /s/ matched for phonological context. Paired t-tests show that /z/ and /s/ differ significantly in the following ways. For all speakers, the acoustic duration of frication is significantly shorter for /z/ than for matched /s/ for all groups of /z/. The acoustic duration of a vowel preceding /z/ is longer than a vowel preceding /s/ for all speakers when /z/ is partly or fully devoiced. However, for vowels preceding voiced /z/, only Speaker 1 had significantly longer durations; for Speakers 2 and 3 the durations of the vowels preceding voiced /z/ were not significantly different from vowels preceding matched tokens of /s/.

Measures of airflow also differed between /s/ and /z/. For Speakers 1 and 3, the mean airflow and the maximum airflow were lower for all groups of /z/'s than for the matched tokens of /s/. This was also true for Speaker 2 for partly or fully devoiced tokens of /z/; however, for this speaker there was no significant difference in the airflow measures between voiced /z/'s and matched /s/'s.

The differences in airflow between /s/'s and devoiced /z/'s could be due to a narrower glottal constriction in [z̥] than in [s], supporting the suggestion by Laver [4] that devoiced sounds may use a phonation type intermediate between the approximated vocal folds suitable for voicing and a fully open glottis. However, the substantial differences in duration between /s/ and all types of /z/ suggest that speakers are distinguishing the two sounds not only by phonation type and that, *contra* Laver, [s] and [z̥] should not be regarded as synonymous. Furthermore, given that mean airflow for devoiced [z̥] is comparable to that for [z], at least for Speakers 1 and 2, it seems unlikely that devoicing results from active widening of the glottis. Rather, it may be the consequence of a lower level of pulmonic effort.

FREQUENCY OF DEVOICING IN DIFFERENT CONTEXTS

Of the three speakers investigated so far, Speaker 1 was the least likely to devoice and Speaker 2 the most likely. Speaker 3 had the most tokens with partial devoicing. In the graphs below, the lightest shading corresponds to devoiced tokens of /z/. The darker gray corresponds to partially devoiced tokens, and the black to voiced tokens.

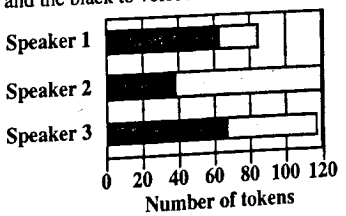


Figure 1. The number of tokens of /z/ in each of the three voicing categories: devoiced (□), partially devoiced (■), and voiced (■).

As for the influence of phonological context on the likelihood of devoicing, the known effect of the sound following a fricative [6] was confirmed in this experiment. The likelihood of devoicing by the different speakers for the /z/ at the end of "falls" is shown in Figure 2. The likelihood of devoicing differs considerably depending on whether a vowel, a voiced stop, a voiceless stop, or silence follows the /z/. (Only the relevant portion of each sentence appears in the graph labels.) Devoicing is least likely when the /z/ is followed by a vowel (the

top bar in the graph) and most likely at the end of a sentence (the bottom bar). Although the speakers differ in how often they devoice overall, they all show a similar rank ordering among contexts.

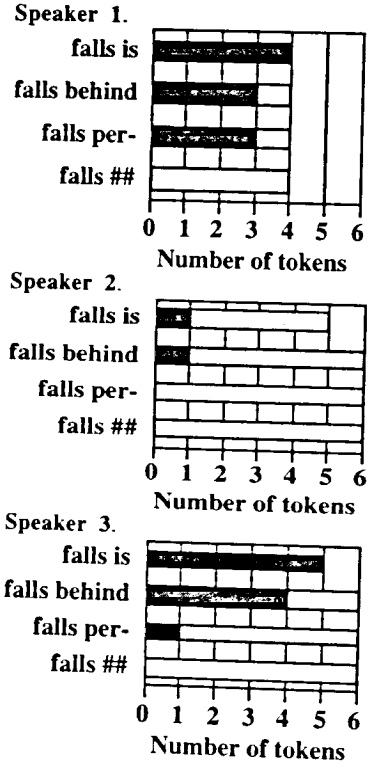


Figure 2. Number of tokens of /z/ at the end of "falls" that speakers produced as devoiced (□), partially devoiced (■), and voiced (■).

A similar pattern was observed for productions of word-final /z/ in "pause" with different following environments. This set of comparisons included the word "paused", in which the /z/ is in a syllable coda but is not word-final. In this coda position, 75% of the tokens of /z/ were devoiced and 25% partially devoiced, compared to 100% devoiced when "pause" was phrase-final. The /z/ in "pause" in all other contexts was less likely to be devoiced.

There is also some influence from the sound preceding a fricative. Syllable and word-initial /z/ were more likely to devoice when preceded by a voiced stop than a vowel. The top graph in Figure 3 shows more tokens with full voicing for syllable-initial /z/ preceded by a vowel ("dessert") than preceded by a voiced stop ("observe"). The lower part of Figure 3 shows more tokens with full voicing for word-initial /z/ preceded by a vowel ("the zinc') than preceded by a voiced stop ("red zinc").

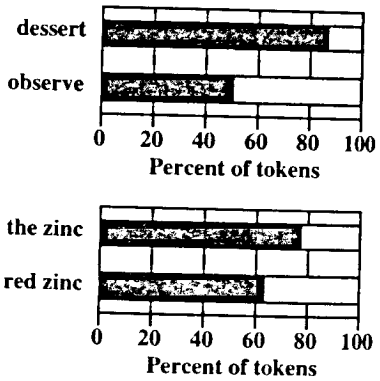


Figure 3. Effect of preceding vowel and voiced stop on likelihood of devoicing in syllable (top) and word (bottom) initial /z/.

All tokens of fully voiced /z/ in "observe" were produced by Speaker 1, the speaker who most often produced voiced /z/'s. All tokens of devoiced intervocalic /z/ in "dessert" and "the zinc" were produced by Speaker 2, the speaker most likely to devoice. Although devoicing was less frequent for the syllable- and word-initial /z/'s shown in Figure 3 than for the word-final /z/'s in Figure 2, nonetheless there were numerous tokens of initial /z/ with at least partial devoicing.

Stress also appears to play a role in determining the likelihood of devoicing a fricative, although the data are too sparse to make firm conclusions. Figure 4 illustrates the greater frequency of devoicing in word-final /z/ at the end of an unstressed syllable in "Dodgers" than at the end of a stressed syllable in "recurs". In both cases the target word was followed by a stressed syllable with initial voiced stop.

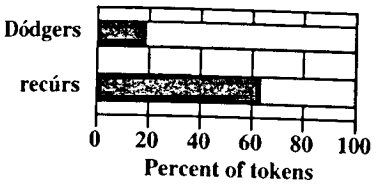


Figure 4. Percent of word-final /z/ that are devoiced at the end of an unstressed syllable (top) and a stressed syllable (bottom).

The comparisons presented here show that devoicing is more likely in positions that are generally the targets of lenition processes — in unstressed syllables, as part of a syllable coda, and at the end of a word or sentence [8]. It is not just the voicing characteristics of the immediate environment that condition the voicing of the fricative. The prosodic strength of the position in which the /z/ occurs is also very important in determining whether or not it will be voiced.

CONCLUSION

Different speakers vary as to how likely they are to devoice /z/. However, they shared similar rank ordering for frequency of devoicing in different phonological contexts. Speaker 2 rarely produced /z/ with glottal vibration during much of its duration, but nonetheless was more likely to produce at least some glottal vibration in those contexts that seemed to favor voicing.

Devoicing is most prevalent in precisely those environments where articulatory effort tends to be weaker. This pattern favors the interpretation that devoicing is a passive rather than an active process: speakers are not generating sufficient airflow from the lungs to maintain the trans-glottal pressure drop needed to maintain voicing. Much as Beckman et al. [9] model prosodic structure for temporal effects in production in terms of "sonority-time space", the occurrence of devoicing could be modeled in terms of the strength of a fricative's prosodic environment. In weaker prosodic environments, speakers may use lower airflow, resulting in a greater likelihood of devoicing. Faced with the voiced fricative 'dilemma' of maintaining a pressure drop across the supralaryngeal constriction to preserve the frication and a pressure drop across the glottis to preserve voicing, speakers of American English apparently prefer to maintain the frication.

ACKNOWLEDGMENT

This work was supported by grant NIH DC 00008 to the UCLA Division of Head & Neck Surgery.

REFERENCES

[1] Ohala, J. (1983), "The origin of sound patterns in vocal tract constraints", in P. MacNeilage (ed.), *The Production of Speech*, New York: Springer-Verlag, 189-216.

[2] Haggard, M. (1978), "The devoicing of voiced fricatives", *Journal of Phonetics 6*, 95-102.

[3] Docherty, G. (1992), *The Timing of Voicing in British English Obstruents*, Berlin: Foris.

[4] Laver, J. (1994), *Principles of Phonetics*, Cambridge: Cambridge University Press.

[5] Lisker, L., Abramson, A., Cooper, F. & Schvey, M. (1969), "Trans-illumination of the larynx in running speech", *JASA 45*, 1544-1546.

[6] Veatch, T. (1989), "Word-final devoicing of fricatives in English", presented at the Linguistic Society of America meeting, Washington, D.C.

[7] Stevens, K., Blumstein, S., Glicksman, L., Burton, M. & Kurowski, K. (1992), "Acoustic and perceptual characteristics of voicing in fricatives and fricative clusters", *JASA 91*, 2979-3000.

[8] Bell, A. & Hooper, J.B. (1978), "Issues and evidence in syllabic phonology", in A. Bell & J.B. Hooper (eds.), *Syllables and Segments*, Amsterdam: North-Holland, 3-24.

[9] Beckman, M., Edwards, J. & Fletcher, J. (1992), "Prosodic structure and tempo in a sonority model of articulatory dynamics", in G. Docherty & D. Ladd (eds.), *Papers in Laboratory Phonology II: Gesture, Segment, Prosody*, Cambridge: Cambridge University Press, 68-86.

SINGLE *VS.* DOUBLE (ABUTTED) CONSONANTS ACROSS SPEECH RATE
X-RAY AND ACOUSTIC DATA FOR FRENCH

Béatrice Vaxelaire
Institut de Phonétique de Strasbourg - USHS - ERS 125, 22 rue Descartes
67084 Strasbourg Cédex France, e-mail: vaxelair@ushs.u-strasbg.fr

ABSTRACT

The aim of this investigation is to analyze, for French, the behaviour of single as opposed to double (abutted) consonants, from X-ray and acoustic data, for two speakers (one female and one male) with increase of speech rate. If configurational constraints have been proposed for articulatory modelling of vowels [1]; [2]; [3], those related to consonant production are lacking in the literature.

INTRODUCTION

X-ray data for consonant productions [4] is rare compared with those available for vowels in the literature. So also is work on the influence of speech rate on vocal tract configurations. The present investigation attemps, hopefully, to contribute, albeit modestly, to reducing this scarcity and also shed some light on articulatory-acoustic consonantal constraints.

METHOD

The corpus consisted of 58 sentences (of 4 to 6 syllables) that embedded the target words. These words were chosen to vary consonantal length /l, p, t, k, b, d, g/ *vs.* /ll, pp, tt, kk, bb, dd, gg/. The present investigation focussed on the the following sentences:

Il **a pas** mal *vs.* /apa *vs.* appa/	Il **zappe pas** mal
Les **attabler** *vs.* /ata *vs.* atta/	La **chatte tachetée**
Très **acariâtre** *vs.* /aka *vs.* akka/	Trois **sacs carrés**
Des **abat**-jour *vs.* /aba *vs.* abba/	**Crabes bagarreurs**
Il l'**a daté** *vs.* /ada *vs.* adda/	**Pas de date** précise
Crabes **bagarreurs** *vs.* /aga *vs.* agga/	**Blagues garanties**

Note that all pairs of sentences had the same number of syllables. The data reported here is thus based on 14 out of the 58 sentences, produced by two speakers (S1 and S2) at a normal (conversational) speaking rate and at a self-selected fast rate. Thus there were 56 conditions in all: 2 speakers X 7 consonant types X 2 speech rates X two durational contrasts.

Recordings and measurements

X-ray films, together with a simultaneous audio recording of the speakers' productions were obtained.

With the help of a grid [5], measurement parameters for vocal tract configurations were determined related to lip-lip and tongue-palate (apex, body) contact-extents (mm); jaw opening (mm) and constriction width (mm), related to the preceding vowel were also measured. Temporal events were detected on the audio signal and specific timing relations between these events allowed determining, in the VC domain, acoustic durations (ms) that correspond to articulatory opening and closing gestures. Speech rate was varied as a perturbing factor of measures obtained from the different linguistic categories, thus allowing to test the resistivity of these patterns observed on both the geometric vocal tract and the acoustic timing levels.

RESULTS AND DISCUSSION

Results presented here are based on raw data and rarely on statistics as it was not possible to acquire data sufficient enough — due to experimental conditions (exposure to X-rays) — to carry out detailed statistic analyses. However, general *tendencies* will be distinguished from *systematic* observations, where the latter show

clear-cut differences across linguistic categories, rate conditions and speakers.

Measurements obtained from mid sagittal profiles, at normal speech rate, show that contact-extents (maximum value for contact) for lip-to-lip and tongue-to-palate (apex and body) productions are longer for double (abutted) consonants than for their single counterparts. This remark is valid, in an intraspeaker pairwise comparison, for all linguistic categories examined, *i.e.* bilabials, apicals and velars, and for both speakers. This difference in contact-extent is shown on Figure 1. of midsagittal tracings for /aka *vs.* akka/ and /ada *vs.* adda/. Table 1 (below) illustrates this fact.

Table 1. Contact-extents (mm) for single vs. double (abutted) consonants at a normal speech rate for S1 and S2.

Single consonants			Double consonants		
vcv	S1	S2	vccv	S1	S2
apa	8	10	appa	11	12
ata	3	8	atta	7	10
aka	7	10	akka	17	14
aba	8	9	abba	9	10
ada	6	3	adda	7	7
aga	9	10	agga	12	12

Although differences, in rare instances, may seem too minimal to be significant (1 mm), it should be noted that this obstruent strategy is always systematic (across several images) and in the same direction. Moreover, the global tendency for abutted consonants to have a longer contact-extent than their shorter counterparts is maintained in fast speech rate, thus showing the relevance of this parameter in differentiating the two categories, *i.e.* even when the linguistic system has been perturbed by speech rate increase. Table 2 (below) confirms this claim.

Table 2. Contact-extents (mm) for single vs. double (abutted) consonants at a fast speech rate for S1 and S2.

Single consonants			Double consonants		
vcv	S1	S2	vccv	S1	S2
apa	9	8	appa	9	10
ata	3	9	atta	9	14
aka	6	9	akka	22	27
aba	8	6	abba	10	9
ada	9	4	adda	14	10
aga	10	12	agga	13	13

Such differences are more or less maintained in fast speech. Increasing speech rate, leads to an increase in contact-extent for single and double consonants (compare values in Tables 1 & 2). This is especially true for the lingual consonants. The difference in behaviour between lingual consonants and bilabials as to their contact-extents has been reported elsewhere [6] for these two speakers, in an entire corpus of 58 sentences embedding varied consonants and vowels, at the two speech rates; the tendency is for contact-extent to increase with speech rate increase for lingual consonants and to remain relatively stable for bilabials. It can be hypothesized that this parameter does not only correspond to the obstruent phase of the plosive; it may also reflect, in cases where the linguistic system is perturbed, an "articulatory overshoot" phenomenon in terms of speed of lingual muscular tissue impact on the hard palate. If contact-extent does, however, reflect the obstruent phase of the consonant, then there should be some relationship between this articultory parameter and closure duration on the acoustic level (cf. *infra*). The other two articulatory parameters, jaw opening and constriction width, did not show any systematic coarticulatory behaviour. These parameters, exploited in terms of area measurements, should certainly give pertinent information in distinguishing single from double abutted consonants.

On the acoustic level, in both speech rates, mean values for closure duration for all double consonants are systematically longer than that of their single counterparts (210 ms *vs.* 90 ms respectively in normal rate; 120 ms *vs.* 70 ms respectively in fast rate). These are all clear-cut differences, indicating that closure duration is a more robust measurement than contact-extent in distinguishing single from double consonants. With speech rate increase, both single and double consonants reduce their closure duration, but it is

the double consonants that undergo a higher reduction (around 90 ms) than the singles (around 20 ms). It has been reported that under speech rate increase, long elements (vocalic or consonantal) tend to resist less to syllable compression than the already short elements even when the short element is a vowel [7]. This has been explained in terms of linguistic constraints for identity preservation, as further compression of short elements would affect their identification [8].

Is there an explicit relationship between contact-extent and closure duration? Figure 2 (left) shows that there is no strict intra-class or inter-class correlation between the two parameters. However, when all conditions are collapsed, and at a normal speech rate, double consonants with markedly longer closure durations, tend to have longer contact-extents than single consonants. This tendency is maintained in fast speech (Figure 2, right).

As concerns vowel duration of V1, mean values obtained are comparable for both classes in normal speech rate (110 ms); when speech rate is increased the vowel for the double class is reduced by 50 ms, whilst that for the singles is reduced by only 25 ms. It seems, indeed, that the longer categories are less resitant to compression provoked by speech rate.

CONCLUSION

Data for consonantal single and double (abutted) consonants have been analyzed and a relevant measure has been unveiled: articulator contact-extent. This is a robust parameter since it is also valid in distinguishing the two linguistic classes, even when speech rate is increased. Moreover, the relationship between contact-extent and closure duration has been demonstrated. Such regularities are useful in evaluating the non linear relationship between geometric parameters and the acoustic output. Thus, if contact-extent is related to closure duration, it also carries information on articulator speed and resulting impact; these factors, however, call for a more thorough analysis. Although articulatory parameters related to the preceding vowel did not show any consistent behaviour, converting the midsagittal measures we obtained to area functions should furnish relevant information for modelling.

ACKNOWLEDGMENTS

My sincere thanks go to Dr P. Perrier for his comments and encouragement, and also to Dr R. Sock for correcting the English draft of this paper. This research was partly supported by an ESPRIT Basic Research Project, n°6975.

REFERENCES

[1] Maeda, S. (1988), "Improved articulatory model.", *JASA* 81, Sup 1. S146.

[2] Perrier, P., Boë, L-J. & Sock, R. (1992), "Vocal tract area function estimation from midsagittal dimensions with CT scans and a vocal tract cast: modelling the transition with two sets of coefficients.", *JSHR* 35, pp. 53-67.

[3] Wood, S. (1979), "A radiographic analysis of constriction location for vowels.", *J. Phonetics* 7, 1, pp. 25-43.

[4] Simon, P. (1967), *Les consonnes françaises, mouvements et positions à la lumière de la cinématographie*, Klincksieck, Paris.

[5] Bothorel, A., Simon, P., Wioland, F. & Zerling, J-P. (1986), *Cinéradiographie des voyelles et consonnes du français, recueil de documents synchronisés pour quatre sujets: vues latérales du conduit vocal, vues frontales de l'orifice labial, données acoustiques*, Institut de Phonétique, Strasbourg.

[6] Vaxelaire, B. (1993), *Étude comparée des effets des variations de débit — lent, rapide — sur les paramètres articulatoires, à partir de la cinéradiographie (sujets français)*, Thèse de Doctorat Nouveau Régime.

[7] Rhardisse, N., Sock, R. & Abry, C. (1990), "L'efficacité des cycles acoustiques dans la distinction des quantités vocalique et consonantique en arabe marocain.", *18e JEP du GCP de la SFA*, pp. 108-112.

[8] Abry, C., Orliaguet, J-P. & Sock, R. (1990), "Patterns of speech phasing. Their robustness in the production of a timed linguistic task: single *vs.* double (abutted) consonants in French.", *CPC, European Bulletin of Cognitive Psychology vol. 10, n°3*, pp. 269-288.

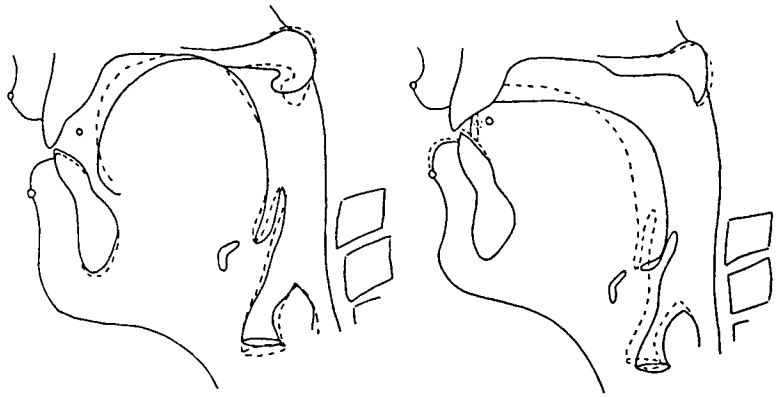


Figure 1. Vocal tract sagittal profiles for /aka/ (bold lines) vs. /akka/ (dotted lines) in normal speech rate (left) and for /ada/ (bold lines) vs. /adda/ (dotted lines) in normal speech rate (right), for Speaker 1.

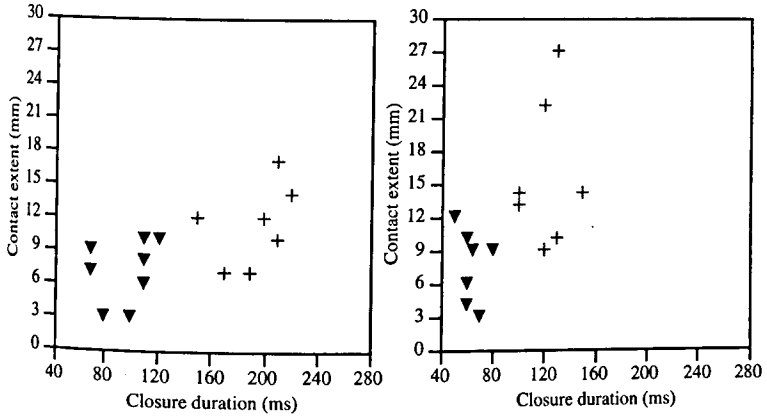


Figure 2. Scattergrams of tongue-palate (apicals and velars) contact-extents (mm) as a function of closure duration (ms) for single (▼) vs. double (+) abutted consonants in normal (left) and fast (right) speech rates. Speakers 1 and 2.

KINEMATIC AND ACOUSTIC ANALYSIS OF ARTICULATORY GESTURE PHASING

Gary Weismer and Ray D. Kent
Waisman Center & Dept. Communicative Disorders
University of Wisconsin-Madison

ABSTRACT

Kinematic and acoustic measures were made to explore the relationship between the two. Results indicate that measures of F2 onset and offset, previously hypothesized to be likely indices of extent of gesture overlap, can in some cases be predicted with good reliability from kinematic measures.

INTRODUCTION

In two recent papers (Weismer, Tjaden, & Kent, in press a,b) we have argued that acoustic measures can serve as indices of articulatory gesture overlap. Measures of onset and offset frequencies of the second formant (F2) in CV and VC sequences have, in certain situations, been shown to vary in a manner consistent with a logical analysis of gesture overlap. However, in the absence of data relating aspects of actual articulatory gestures to the acoustic output of the vocal tract, the relative goodness of these inferences will remain unknown. The purpose of this paper is therefore to report some exploratory coanalysis of x-ray microbeam data and speech acoustic measures from selected speakers and utterances. In particular, we were interested in the extent to which certain kinematic measures could be used to predict F2 onset and offset measures. The relative goodness of these predictions should point to some guidelines for the inference of articulatory behavior from acoustic measures.

METHODS

Kinematic and acoustic data were collected as part of the x-ray microbeam data base project (Westbury, 1994). The data base consists of over fifty speakers who produced a common speech sample of material ranging from simple syllables to a relatively lengthy reading passage.

Subjects

The completed analysis will be based on ten speakers. In the present report, data are presented for two speakers, including a male aged 28 years (JW7) and a female aged 20 years (JW31). Both subjects had normal orofacial structures, spoke a dialect generally described as Greater American Midwest (one speaker grew up in Iowa, the other Wisconsin), and reported no history of speech or language problems.

Speech Sample

Kinematic and acoustic measures were obtained from the sequence /ubIg/ in the utterance, *The other one is too big*. Between 15 and 20 repetitions of this utterance were produced by each subject, these repetitions including several at slower-than-normal and faster-than-normal speaking rates. The rate variation was desirable for the present analyses, because of the presumed effect of rate on the extent of gesture overlap (see, for example, Munhall & Lofqvist, 1992; Tjaden, Weismer, & Kent, 1994) and thus on variation in F2 onset and offset.

Data Collection

X-ray microbeam data for the two subjects reported herein were collected using the standard array of 11 pellet locations used throughout the data base project. For the purposes of the present project, attention was focussed on the three pellets attached at the mid-ventral, mid-dorsal, and dorsal tongue locations (see Westbury, 1994, p. 39). The approximate distances of these pellets from the tongue apex (measured along the surface of the extended tongue) were 25, 44, and 60 mm, respectively (the dorsal pellet was not tracked in subject JW7 due to technical problems at the time of data collection). The speech acoustic signal was digitized and stored synchronously with the pellet position histories. Complete technical details concerning data collection for the x-ray microbeam data base project are provided in Westbury (1994).

Measures

The measures taken in the current investigation are summarized in Fig. 1, which shows pellet time histories in the *x* dimension (antero-posterior) for the mid-ventral (TB1x), mid-dorsal (TB2x), and dorsal (TDx) markers, as well as the synchronized acoustic waveform of the /ubIg/ sequence. Downward pellet time histories reflect posterior tongue movement, and upward histories forward tongue movements. The pattern shown in Fig. 1, where all pellets describe a U-shaped path throughout the sequence of interest, characterized each utterance analyzed for the present report. This pattern is consistent with the phonetic expectation of tongue backing for the /u/ and then fronting for the /I/. The point in time at which the movement began to be directed forward usually occurred within the /b/ closure interval, but sometimes occurred prior to the acoustic evidence of the onset of the closure interval. Using this point in time, we defined DTB1x, DTB2x, and DTDx as the temporal interval between the onset of the forward-directed motion and the first glottal pulse of the vowel /I/ in *big*. The first glottal pulse for /I/ was chosen as the termination point for this temporal measure because it serves as the location of the F2on measure we have evaluated as an index of gesture overlap. We reasoned that variations in the duration of this interval of tongue movement might be correlated with F2on, if earlier onsets of the forward-directed tongue movement could be interpreted as greater overlap between the lingual and labial gestures for this phonetic sequence.

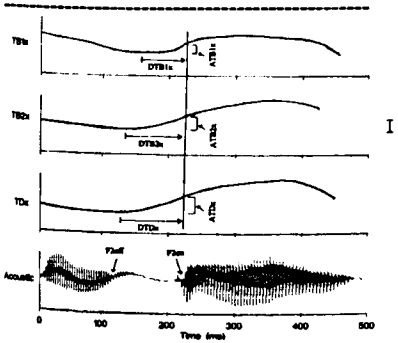


Figure 1. Graphic summary of measures.

We also measured the amount of forward movement throughout these temporal intervals, indicated in Fig. 1 as ATB1x, ATB2x, and ATDx. We reasoned that variations of extent of forward movement may also be correlated with variations in F2on.

The acoustic measures taken included the F2on value measured above, the F2off value of /u/, measured at the last glottal pulse preceding the closure interval, the /b/ closure duration and the /b/ VOT. Formant

measures were made using a combination of LPC spectra and cursor placement on digital spectrograms; temporal measures were obtained from the combined digital spectrogram/waveform display.

RESULTS AND DISCUSSION

Averaged data for selected measures described above are presented in Table 1.

Table 1. Averaged data for measures.

	DTB1x	DTB2x	ATB2x	F2on	F2off
JW31	56	104	1.39	1886	1311
JW7	103	103	7.20	1468	903

These data indicate that JW31's tongue displacements (in mm) in the *x* dimension were substantially smaller than JW7's tongue displacements. The two subjects showed very different temporal intervals (in msec) for TB1x, but essentially identical intervals for TB2x (no comparison could be made for TDx because of the technical problem noted above). Closure duration and VOT were quite similar for the two subjects, and the differences in F2 measures can probably be attributed largely to gender-based differences in vocal tract size.

Table 2 reports the significant results from exploratory analyses in which hypothesized acoustic indices of gesture overlap, F2off for the /u/ in *too* and F2on for the /I/ in *big*, were regressed on the temporal and displacement measures obtained from the pellet time histories. Because of high intercorrelations between either temporal and displacement measures, or displacement measures for two different pellets, some of these effects are mutually redundant. For example, effects *1* & *2* for

Table 2. Significant regression effects.

JW31
1. F2on=1994-1.95*DTB1x, p=.018, R^2= 29%
2. F2on=1961-117*ATB1x, p=.019, R^2= 29%
3. F2off=1445-208*ATB1x, p=.003, R^2= 45%
4. F2on=1813-2.25*OTB1x, p=.010, R^2= 34%

JW7
1. F2on=1390+11.1*ATB1x, p=.014, R^2= 33%
2. F2on=1382+12.0*ATB2x, p=.020, R^2=30%
3. F2off=1315-4.02*DTB1x, p<.001, R^2=67%
4. F2off=1124-31.5*ATB1x, p<.001, R^2=78%
5. F2off=1161-35.9*ATB2x, p<.001, R^2=80%

both JW31 and JW7, and effects *3*, *4*, and *5* for JW7 are alternate expressions of the same phenomena because of the kinds of interdependencies cited above. With this in mind, a general summary of the regression analyses is as follows. First, F2on appears to be related significantly to displacement between the onset of forward movement and the first glottal pulse of /I/ (ATBx), albeit in different directions for the two subjects. For JW31 larger ATBx was related to lower F2on, whereas the opposite was the case for JW7. Intuitively, it seems as if the latter pattern would be more consistent with the idea that earlier onsets of forward movement (and the associated larger displacements over the DTBx interval) would reflect greater gesture overlap between the lingual and labial gestures for this sequence. Second, F2off of /u/ is strongly related to the ATBx magnitude for both subjects, such that larger ATBx was associated with lower F2off. Figure 2 shows these effects as scatterplots for both subjects.

None of the analyses reported above was based on a direct measure of gesture overlap. To explore the predictive utility of a measure that might more directly express the interarticulatory phasing of lingual and labial gestures, we subtracted, token by token, the closure interval + VOT from

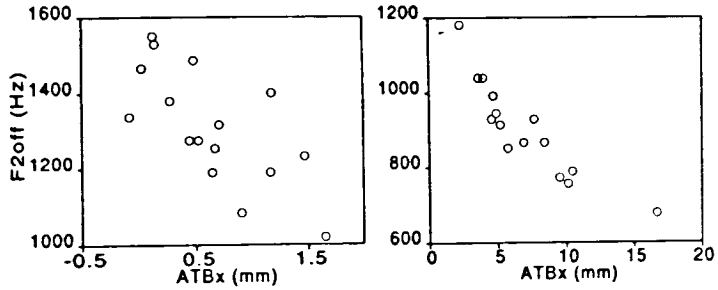


Fig. 2. Scatterplots of F2off-ATBx effects for subjects JW31 (left) and JW7 (right).

DTBx which produced a measure of the initiation of forward-directed tongue movement toward /I/ relative to the acoustically-inferred onset of labial closure for /b/. This interval was, with a single exception among 16 repetitions, always negative for JW31, indicating that the forward movement began after the labial closure; as seen in Table 2 the F2on was predicted significantly from this measure of overlap (OTBx), the nature of the effect being that the less negative the measure of overlap the lower the F2on. This is exactly what we would expect if F2on was to be used as a reliable indicator of gestural overlap, because less negative values of OTBx would imply that the forward directed tongue movement was beginning earlier in time relative to the onset of labial closure for /b/ (i.e., that there was greater gestural overlap between the labial and lingual gestures). Unfortunately, a corresponding analysis of JW7's data failed to reveal a significant function.

The data reported here show some promise in establishing reasonable links between kinematic and acoustic measures. Obviously the extent of speaker variation must be known, and exploration of better measures of interarticulator phasing undertaken, before straightforward articulatory inferences can be made from acoustic measures.

REFERENCES

MUNHALL, K., & LOFQVIST, A. (1992). Gestural aggregation in speech: Laryngeal gestures. *Journal of Phonetics, 20*, 111-126.

TJADEN, K., WEISMER, G., & KENT, R.D. (1994). *Evidence of gestural overlap across speaking rate: Acoustic data.*. Paper presented at the 128th meeting of the Acoustical Society of America, Austin, TX, USA.

WEISMER, G., TJADEN, K., & KENT, R.D.(in press). Can articulatory behavior in motor speech disorders be accounted for by theories of normal speech production? *Journal of Phonetics, 23*.

WEISMER, G., TJADEN, K., & KENT, R.D.(in press). Speech production theory and articulatory behavior in motor speech disorders. In F. Bell-Berti and L.J. Raphael (Eds.), *Studies in speech production: a festschrift for Katherine Safford Harris*. New York: American Institute of Physics.

WESTBURY, J.R. (1994). X-ray Microbeam Speech Production Database User's Handbook, Version 1.0 (131 pages). University of Wisconsin-Madison.

THE GESTURAL AND TEMPORAL ORGANISATION OF ASSIMILATION

Sidney A. J. Wood

Department of Linguistics, University of Lund, Sweden

ABSTRACT

An X-ray motion film study of the palatalisation of Bulgarian apico-alveolar stops and Swedish dorsovelar stops is presented. The gestures involved in an assimilation are initiated earlier, or are held longer, than in nonassimilated situations. This revision of gesture timing in relation to adjacent activity indicates that assimilation is preplanned and does not reflect coarticulation or vocal tract biodynamics. Gestural coordination for these examples is best described as coproduction than feature spreading.

INTRODUCTION

Ever since the 1960s it has been customary to distinguish between universal articulatory constraints and arbitrary language-specific speech habits [1]. The former were said to be intrinsic to the "speech mechanism" (i.e. the regular consequence of the speech motor system and the biodynamics of the vocal tract), the latter were said to be extrinsic (i.e. controlled from the brain). For many phoneticians, assimilation has come to be seen as an intrinsic process, often explained in terms of Öhman's model of coarticulation [2], where it is proposed that alveolars are palatalised and dorsovelars fronted by being superimposed on and summed with the underlying tongue body activity for a front vowel. This contrasts with the view of classical phoneticians like Sweet and Jespersen who saw assimilation as the result of preplanned reorganisation of articulation changing phoneme targets [3,4].

Two rival processes have been proposed for assimilation, feature spreading and coproduction [5]. With the feature-spreading approach, inherited from classical phonetics, the articulatory target of the assimilated phoneme is said to be changed by having the assimilating articulation incorporated into its own plan. With coproduction, the assimilating articulation is never actually copied to the assimilated phoneme but remains the exclusive property of the assimilating phoneme and is produced simultaneously with the assimilated phoneme.

These two issues, extrinsic vs intrinsic production of assimilation, and feature spreading vs coproduction, are investigated here.

This study is part of an investigation of the temporal organisation of coarticulation and assimilation [6,7,8,9]. Articulators were moved into position, their postures were retained for a while, and then they were withdrawn (these three phases of an articulator gesture are referred to here as the onset, the hold and the withdrawal, respectively). While different gestures were allocated simultaneously to different phonemes, any one articulator gesture was invoked for just one phoneme at a time. Whenever mutually antagonistic articulator gestures were required for successive phonemes, they were implemented sequentially, as expected by [10], rather than blended simultaneously as expected by [2].

PROCEDURES

The data has been obtained from X-ray films by procedures described in [6,8]. The films were made at 75 frames/sec (13.3 msecs per frame), with a 3 msec exposure during each frame. Background information on Bulgarian vowels is given in [11,12].

The Bulgarian utterance, spoken by an adult male from Sofia, was *Deteto xodi po pătishtata*, /deˈteto ˈxodi ˈpo ˈpătiʃtata/, [dɾˈteto ˈxɔdi ˈpo ˈpɐtiʃtətə], the child was walking along the path.

The Swedish utterance, spoken by an adult male from Helsingborg in South Sweden, was a nonsense word *kackikakor*, /ˈkaki,ka:kur/, [ˈkaki,ka:kuʁ].

Articulator movement was detected and recorded by comparing midsagittal profile tracings from one film frame to the next. The articulator gestures that are relevant for the assimilations reported here are four tongue body gestures relative to the mandible (palatal, velar, uvular and pharyngeal) as described in [13], and coronal elevation relative to the tongue

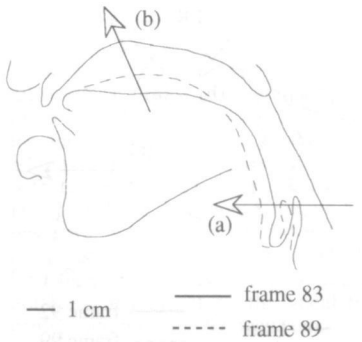


Figure 1. Two profiles from /t/ in /ătі/ (pătishtata), comparing the nonpalatalised implosion (film frame 83, final frame in /ă/ vocoid before /t/ occlusion) with the palatalised release (film frame 89, first frame of /i/ vocoid following /t/ occlusion). Note the more gradual tapering of the vocal tract posteriorly to the apico-alveolar constriction in the palatalised case. The tongue body gestures (a,b) are explained in the text.

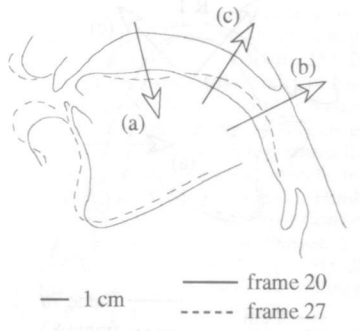


Figure 2. Two profiles from /t/ in /eto/ (deteto xodi), comparing the palatalised implosion (film frame 20, final frame of /e/ vocoid before /t/ occlusion) with the non-palatalised release (film frame 27, first frame of the /o/ vocoid following /t/ occlusion). Note the more gradual tapering of the vocal tract posteriorly to the apico-alveolar constriction in the palatalised case. The tongue body gestures (a,b,c) are explained in the text.

body [6,8]. This set of tongue gestures is available to both vowels and consonants, contrary to the view put forward by Öhman [2] and argued for by Fowler [14] that vowels and consonants are produced by independent systems. The palatal gesture is needed for front vowels and palatal consonants, the velar gesture is needed for high back vowels and dorsovelar consonants, the uvular gesture is needed for mid back vowels and uvular consonants, the pharyngeal tongue body gesture is needed for low vowels and low back consonants. The coronal gesture helps control vocal tract resonance conditions during rounded vowels [12,15].

RESULTS

Bulgarian apicoalveolar stops

The alveolar stops were palatalised on the flank that was adjacent to a front vowel. Figure 1 shows an example of a palatalised release (/t/ in /ăti/ in *pătishtata*), comparing the implosion and the release. The profiles selected are the final film frame of the vocoid segment of /ă/ and the first film frame of the vocoid segment of /i/, as these are the audible parts of the signal nearest to the silent occlusion. Figure 2 shows an example of a palatalised implosion (/t/ in /eto/ in *deteto xodi*), again comparing the implosion and the release. What is interesting here is the different timing of the articular gestures for these two contextual situations.

Gesture timing is not illustrated. For the palatalised release in Fig. 1, the pharyngeal tongue body withdrawal from /ă/ (a) and the palatal tongue body onset for /i/ (a,b) commenced already during the pre-stop vocoid segment of /ă/. They continued during the /t/ occlusion, and had progressed sufficiently by the /t/ release to palatalise it. The palatal tongue body gesture was phased relative to the coronal gesture and alveolar occlusion in such a way that the release was palatalised but not the implosion. This phasing of articulator gestures is typical of coproduction (with the palatal tongue body gesture of the post-stop vowel implemented simultaneously with the alveolar occlusion and timed to arrive at the release), rather than anticipatory feature spreading from the vowel to the alveolar stop (where the stop would get its own palatal articulator gesture copied from the vowel).

For the palatalised implosion in Fig. 2 the palatal tongue body gesture of /e/ was held until the end of its pre-stop vocoid

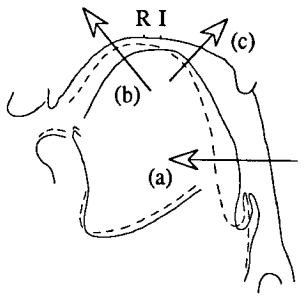


Figure 3. Two profiles from /k/ in /aki (kackikakor), comparing the nonpalatalised implosion (film frame 78, final frame in /a/ vocoid before /k/ occlusion) with the palatalised release (film frame 83, first frame of /i/ vocoid after the /k/ occlusion). The centre of the linguo-palatal contact shifted anteriorly from (I) to (R). The tongue body gestures (a,b,c) are explained in the text.

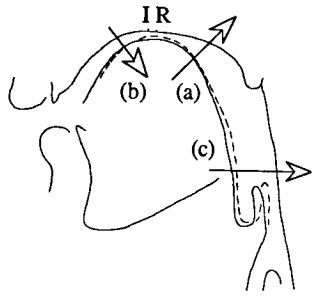


Figure 4. Two profiles from /k/ in /ika/ (kakikakor), compariing the palatalised implosion (film frame 92, final frame in /i/ vocoid before /k/ occlusion) with the nonpalatalised release (film frame 99, first frame of /a/ vocoid after the /k/ occlusion). The centre of the linguo-palatal contact shifted posteriorly from (I) to (R). The tongue body gestures (a,b,c) are explained in the text.

segment. The uvular tongue body onset for /o/ (a) and the palatal tongue body withdrawal from /e/ (a,b) did not commence until after the /t/ implosion. In this particular instance there is also a velar tongue body onset for the dorsovelar fricative /x/ of *xodi* (c) that commenced together with the palatal withdrawal and velar onset during the /t/ occlusion.

The palatal tongue body gesture of a front vowel in Bulgarian is thus phased in two different ways relative to the occlusion of an adjacent alveolar stop. To palatalise the implosive flank the palatal posture of the assimilating vowel is held until the end of the pre-stop vocoid segment before being withdrawn. To palatalise the release flank, the palatal onset of the post-stop vowel is activated already during the pre-stop vocoid segment and continues during the alveolar occlusion in order to be in place at the release. The palatal gesture of the vowel is not only locked to its own vocoid segment, it is also locked to the respective flank of an adjacent alveolar stop. The two different phasings point to preplanning of this assimilation.

Fant reports X-ray and acoustic data on palatalisation in Russian [16]. The vocal tract posteriorly to palatalised alveolar occlusion is more narrowly tapered, with consequently greater coupling of the turbulence source to high frequency back cavity resonances, producing a burst spectrum with more energy in the 8-9000 Hz region. The F2 locus is about 200 Hz higher. The narrower tapering can be seen in Figs. 3 and 4.

Swedish dorsovelar stops

Figure 3 shows an example of a palatalised release (/k/ in /aki/ in *kackikakor*), comparing the implosion and the release. The profiles are selected as before. Figure 4 shows an example of a palatalised implosion (/k/ in /ikaː/). Here again, the gestures were timed differently for each situation.

For the palatalised release in Fig. 3, the pharyngeal tongue body withdrawal from /a/ (a), the palatal tongue body onset for /i/ (a,b) and the velar tongue body onset for /k/ (c) all commenced during the pre-stop vocoid segment of /a/, continuing during the occlusion (shifting its centre about 5 mm from I to R).

For the palatalised implosion in Fig. 4, the palatal tongue body gesture of /i/ was held almost until the end of its pre-stop vocoid segment. The the velar tongue body onset for /k/ (a) and the palatal tongue body withdrawal from /i/ (b,c) commenced simultaneously just before /k/ occlusion. The retraction from I to R during the occlusion was so slight that the release was almost as palatalised as the onset, which is surprising considering that the post-stop vowel is a dark [ɑ]-like allophone of /a/ (compare this release with the more retracted implosion after /a/ in Fig. 3). The pharyngeal tongue body onset for pre-stop /aː/ (c) did not commence until after the /k/ release, which is later than the activity in Fig. 4, and did not have any effect on the vocal tract configuration for the /k/ occlusion.

CONCLUSIONS

The timing of articulator gestures for palatalisation is best described as coproduction rather than feature-spreading.

The different phasing patterns of the palatal tongue body gesture for palatalisation of implosions and releases, for alveolar stops in Bulgarian and for dorsovelar stops in Swedish point to specific gestural reorganisation for this assimilation that is different from the regular interweaving of gestures for coarticulation. They are an example of how assimilation is a preplanned process that applies to specific phonemes in defined situations distinct from the coarticulatory adjustments of all phonemes to their neighbours.

REFERENCES

[1] Wang, W. S-Y. and Fillmore, C. J. (1961). Intrinsic cues and consonant perception. *Journal of Speech and Hearing Research* Vol. 4, pp. 130-136.

[2] Öhman, S. (1966). Coarticulation in CVC utterances: spectrographic measurements. *Journal of the Acoustical Society of America* Vol. 39, pp. 151-168.

[3] Sweet, H. (1877). *Handbook of Phonetics*. Oxford: Clarendon.

[4] Jespersen, O. (1897). *Fonetik*. Copenhagen: Det Schubotheske Forlag.

[5] Fowler, C. A. (1980). Coarticulation and theories of extrinsic timing control. *Journal of Phonetics* Vol. 8, pp. 113-133.

[6] Wood, S. A. J. (1991). X-ray data on the temporal coordination of speech gestures. *Journal of Phonetics* Vol. 19, pp. 281-292.

[7] Wood, S. A. J. (1993). Crosslinguistic cineradiographic studies of the temporal co-ordination of speech gestures. *Working Papers* Vol. 40, 251-263, Dept. of Linguistics, University of Lund.

[8] Wood, S. A. J. (1994). Syllable structure and the timing of speech gestures: an analysis of speech gestures from an X-ray film of Bulgarian speech. In R. Aulenko and A-M. Korpijaako-Huuhka (eds), *Proceedings of the Third Congress of the International Clinical Phonetics and Linguistics Association*, pp. 191-200. Publications of the Dept. of Phonetics Vol. 39, University of Helsinki.

[9] Wood, S. A. J. Assimilation and coarticulation - evidence from the coordination of tongue gestures for the palatalization of Bulgarian alveolar stops. Proceedings of the Barcelona symposium on modelling tongue articulation. *Journal of Phonetics*, forthcoming.

[10] Kozhevnikov, V. A. and Chistovich, L. A. (1965). *Speech, Articulation and Perception*. Washington: Joint Publications Research Service.

[11] Pettersson, T. and Wood, S. A. J. (1987). Vowel reduction in Bulgarian and its implications for theories of vowel production. *Folia Linguistica* Vol. 21, pp. 261-279.

[12] Wood, S. A. J. and Pettersson, T. (1988). Vowel reduction in Bulgarian: the phonetic data and model experiments. *Folia Linguistica* Vol. 22, pp. 239-262.

[13] Wood, S. A. J. (1979). A radiographic analysis of constriction locations for vowels. *Journal of Phonetics* Vol. 7, pp. 25-43.

[14] Fowler, C. A. (1983). Realism and unrealism, a reply. *Journal of Phonetics* Vol. 11, pp. 303-322.

[15] Wood, S. A. J. (1986). The acoustical significance of tongue lip and larynx maneuvers in rounded palatal vowels. *Journal of the Acoustical Society of America* Vol. 80, pp. 391-401.

[16] Fant, C. G. M. (1960). *The Acoustic Theory of Speech Production*. The Hague: Mouton.

This work was supported by the Swedish Council for Research in the Humanities and Social Sciences

EMERGENT SYLLABLE USING ARTICULATORY AND ACOUSTIC PRINCIPLES

A.R. Berrah, L.J. Boë and J.L. Schwartz
Institut de la Communication Parlée, Grenoble, France

ABSTRACT

Speech sequences can be regarded as a succession of syllable gestures composed of consonants and vowels. The choice of these basic units is not arbitrary: languages have a tendency to optimize their sound structures. This paper deals with the prediction of CV syllables. In order to bring about these gestures, we have suggested syllabic prototypes defined acoustically and articulatorily. Then, we have defined new optimization criteria.

1. INTRODUCTION

"Where do phonological universals such as segments and features come from?" In 1984, Lindblom et al. [1] addressed this ambitious question by trying to simulate the emergence of a self-organized model of phonological structure. They have restricted their attention to articulation involving transitions from a closed (stoplike) to an open (vowellike) state, in other words, CV syllables: the most frequent in the languages of the world [2].

With a 7D articulatory space (lip height and protrusion; jaw; body, dorsum and apex of the tongue; height of the larynx), and a 4D acoustic-perceptive space (F1-F4 in barks) our research has consisted in predicting the rank of "efficiency" of CV syllables among all the 20 possible combinations stoplike [b d g ɢ] (with [ɢ] an hypothetic but not observed pharyngeal stop) and with vowellike [i 'e' a 'o' u]. The CV prototypes have been designed with Maeda's articulatory model [3]. For designing CV prototypic transitions, we have taken into account X-Ray data for coarticulation [4], and Sussman's locus values for formant transitions [5]. Each syllable is characterized by a global efficiency: ratio of acoustic efficiency and articulatory cost. We define a criterion of maximization of intersyllabic distances by using F'_2 (in Barks) evaluation in the perception of transitions. For a given system of syllables, we propose a global criterion taking into account intersyllabic distances and efficiency of each syllable. The emergence of a syllable in a set of syllables is then simulated and its rank is discussed.

As a by-product a 3D space is proposed for the 20 syllables; it has been derived from a Kruskal analysis [6] on intersyllabic acoustic distances of our prototypes.

2. PROBLEM SOLVING

A great number of questions still remain to try to explain how linguistic sounds are made up. Topological studies have shown the occurrence in the systems of a group of phonetic properties which are found in most languages of the world.

The first questions asked in the face of the universal tendencies in the systems are notably:

- Which factors entail these restrictions?
- What are the causes which provoke the tendency to use only one little set of sound qualities for communication?

Assumptions are made by examining universals from a phonetic point of view and point towards the notion of ***functional efficiency***. The systemic nature of sounds is shown: a sound is examined as the constituent of a system.

Two very important fields of research will stem from these hypotheses in the seventies: The *Quantal Theory* of Stevens [7] and the *Dispersion Theory* of Liljencrants & Lindblom [8].

The initial hypothesis is based on two principal ideas: articulatory simplicity and perceptive distinction.

3. SYLLABIC PROTOTYPES

We have restricted our research to the twenty syllables obtained by combining the following 4 consonants: [b d g ɢ] and the 5 vowels [i 'e' a 'o' u]. The choice of these vowels is not arbitrary: Typology reveals that most languages of the world with 5 vowels have a vowel system with [i 'e' a 'o' u]. We want to obtain an articulatory and an acoustical

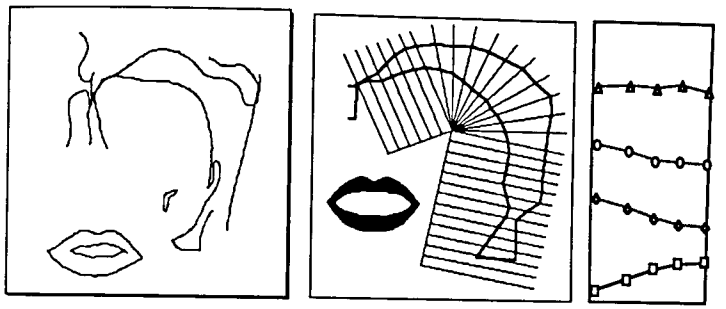


Figure 1. From left to right: X-Ray photography, vocal tract shape and the formant transitions from [d] to [a].

target for every syllable to determine which configurations are typical for syllabic systems.

In fact, to select syllabic standards in the acoustic and articulatory space, we must characterize a typical form of the vocal tract for a certain number of acoustic parameters (the formants).

Implementation

The articulatory and the acoustic targets of every syllable are obtained from the prototype of the vowel and the coarticulated occlusion of the consonant.

(i) We have used the Vallée et al.'s vowel prototypes [9].

(ii) Thanks to Maeda's model, we have generated a syllabic prototype from the vowel prototype by adjusting the parameters of some articulators [10] until occlusion is reached.

Prototypic syllables

Thanks to the X-Ray data of Bothorel et al., we have proposed [10] a typical form of the vocal tract for the fifteen most frequent syllables CV [b d g] x [i 'e a 'o' u] and five hypothetical CV syllables [ɢ] x [i 'e' a 'o' u] [11].

Locus notion

To validate the syllabic prototypes, we have computed the locus of consonants [b], [d], [g] and [ɢ], [g] presents two loci associated to front and back vowels. Figure 2 presents the locus for our prototypic syllables in a 3D space in agreement with Sussman's data.

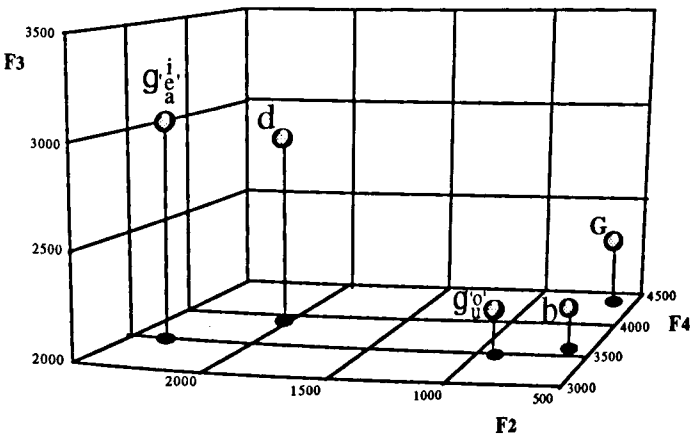


Figure 2. Consonantal locus of our CV syllables in a 3D space F2-F4.

4. THE PREDICTION MODEL

Our syllable model consists in predicting the most frequent syllables used in languages. We hypothesize that they are the most efficient for communication. The general postulate of this study is that phonetic signals evolve in such a way that their production and perception is more easy for training, production and perception.

Acoustic efficiency

The acoustic space considered is 4D which represents the first 4 resonances of the vocal tract. Our predictive model is inspired of Liljencrants' & Lindblom's model [8] and the DFT model [12]. This model is based on a bidimensional vowel space defined by the first (F_1) and the second effective formant (F'_2). In order to take into account physiological and perceptual constraints [13], The contribution of F_1 is increased.

Intrasyllabic efficiency

Intrasyllabic efficiency (or perceptual salience) is a characteristic of individual transitions. It is defined in terms of the extent of syllable trajectory, i.e., the distance between the initial and the final auditory spectra. A syllable is more salient if the resonance frequencies of the consonant and the vowel are different. The acoustic efficiency, Acoust-Eff_{CV}, is given by:

$$\sqrt{(F_{1_c} - F_{1_v})^2 + \lambda^2 * (F'_{2_c} - F'_{2_v})^2}$$

λ^2: ponderation between F_1 et F'_2.

Intersyllabic efficiency

Intersyllabic efficiency (or perceptual distance) of two arbitrary CV transitions is a dimension used to rank all possible pairs of CV events in order to minimize the risk of confusion between the items of the lexicon.

The intersyllabic distance ds_1s_2 (S_1=C_1V_1, S_2=C_2V_2) is described by:

$$\sqrt{\begin{aligned}&(F_{1_{c1}} - F_{1_{c2}})^2 + \lambda^2 * (F'_{2_{c1}} - F'_{2_{c2}})^2 \\ &+(F_{1_{v1}} - F_{1_{v2}})^2 + \lambda^2 * (F'_{2_{v1}} - F'_{2_{v2}})^2\end{aligned}}$$

Articulatory cost

The preference for less extreme articulations introduces a ranking of both static configurations and movements. The articulators can have not the same weight [12] in the evaluation of the articulatory cost.

The expression of the articulatory cost of a CV syllable produced by m articulatory parameters P is given by:

$$Art\_Cost_{CV} = \sqrt{\sum_{k=1}^{m} w_p^2 * (P_{kc} - P_{kv})^2}$$

System and energy

The contrast properties are determined by the relations between the syllables inside a system, and not the own acoustic and articulatory characteristics of each one. The prediction principle of syllabic systems consists to accomplish a research of all optimal systems.

Global efficiency

Each syllable is identified by its acoustic and articulatory characteristics. The emergence of a syllable depends, in part, on the ratio of the acoustic efficiency and the articulatory cost. This ratio constitutes the *global efficiency* described by:

$$Glob\_Eff_{cv} = \frac{Acoust\_Eff_{cv}}{Art\_Cost_{cv}}$$

Maddieson & Precoda [2] have done the observation according to that there is as many occurrences of [di] as of [du]. The computed global efficiency of [di] and [du] are roughly the same. Thus, our model has been adjusted to verify this strong observation basis.

Energy

The stability of a system is appreciated in the acoustic level by a criterion of minimal energy. The optimal systems minimize the following expression:

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{1}{d_{s_i s_j}^{\ 2}} + \beta \sum_{i=1}^{n} \left(\frac{1}{Glob\_Eff_{s_i}} \right)$$

n: number of system syllables.

The parameter β weights the intrasystemic values with regard to the intersystemic values.

5. RESULTS

Acoustic criterion

The emergence order of the twenty syllables according to the acoustic criterion is: [ba ga da be bo gi do ge ga de go ge bi du go bu gi di gu gu].

It is not strange that salient syllables as [ba] and [da] are in the top of the list.

Articulatory criterion

The emergence rank according to the articulatory criterion is: [bu bo gi gu de di ge bi be go du ga go ba da ga do gu ge gi]. We have penalized the pharyngeal syllables by an important weight for the backward displacement of the tongue body.

Global efficiency

Once, the perceptive and articulatory characteristics computed, we can evaluate the global efficiency. The list of the syllables according to their value is: [bo bu de ge be ba ga go da gi bi gu du di ga do go gi ge gu].

Syllable space

We have applied Kruskal's analysis [6] on intersyllabic acoustic distances of the prototypes to obtain a 3D space [10]. We have noted a predominance of the vowel qualities on the location of the syllables.

System energy

At this step, we want to select among the twenty syllables, a lexicon of nine syllables highly discriminable and easy to produce. For this purpose, we must generate the optimal syllable lexicon, i.e., the system of nine syllables presenting the minimum of energy. After different simulations, allowing to adjust the intrasyllabic weight (β), the system [bu ba ga da gi bi gu du di] has emerged.

6. CONCLUSION

Our *substance-based* approach can be used to predict the main tendancies of emergence of lexicons. This first step has allowed to predict the universally favoured syllables taking into account constraints of production and perception.

REFERENCES

[1] Lindblom, B., MacNeilage, P. & Studdert-Kennedy, M. (1984), *Self-Organizing Processes and the Explanation of Phonological Universals*, In *Explanation of Language Universals*. B. Butterworth et al. (Eds.), Mouton, The Hague, 181-203.

[2] Maddieson, I. & Precoda, K. (1992), Phonetic models and syllable structure, *Phonology*, vol. 9, 45-60.

[3] Maeda, S. (1989), *Compensatory Articulation during Speech: Evidence from the Analysis and Synthesis of Vocal-Tract Shapes using an Articulatory Model*, In *Speech Production and Modelling*, W.J. Hardcastle & A. Marchal (Eds.), Academic Publishers, Kluwer, 131-149.

[4] Bothorel, A., Simon, P., Wioland, F. & Zerling, J. (1986), *Cinéradiographies des voyelles et des consonnes du français*, Institut de Phonétique, Strasbourg, France.

[5] Sussman, H.M., Hoemeke, K.A. & Farhan, S.A. (1993), A Cross-linguistic Investigation of Locus Equations as a Phonetic Descriptor for Place articulation, *J. Acoust. Soc. Am.*, vol. 94, 1256-1268.

[6] Kruskal, J.B. (1977), *Multidimensional Scaling and other Methods for Discovering Structure*, In *Statistical Methods for Digital Computers. Mathematical Methods for Digital Computers*, K. Enslein et al. (Eds.), John Wiley, New-York, vol. 3, 296-339.

[7] Stevens, K.N. (1972), *The Quantal Nature of Speech: Evidence from Articulatory-Acoustic Data*, In *Human Communication: a Unifed View*, P.B. Denes & J.R. Davis (Eds), McGraw-Hill, New-York, 51-66.

[8] Liljencrants, J. & Lindblom, B. (1972), Numerical Simulation of Vowel Quality Systems: the Role of Perceptual Contrast, *Language*, vol. 48, 839-862.

[9] Vallée, N., Boë, L.J., Payan, Y. (1995),Vowel prototypes for UPSID'S 33 phonemes, In *these Proceedings*.

[10] Berrah, A.R. (1994), L'émergence des structures sonore, D.E.A. Sciences Cogntives, INP, Grenoble.

[11] Maddieson, I. (1986), *Patterns of Sounds*, Cambridge University Press.

[12] Schwartz, J.L., Boë, L.J. & Vallée, N. (1995), Experimenting the dispersion focalization theory: Phase spaces for vowel systems, In *these Proceedings*.

[13] Lindblom, B.E.F. & Lubker, J. (1985), *The Speech Homonculus and a Problem of Phonetic Linguistics*, In *Phonetic Linguistics*, V. Fromkin (Ed), Academic Press, Orlando, 169-192.

DAF EFFECTS ON STUTTERERS VOICE QUALITIES AND VOWELS SYSTEMS

B. Harmegnies, M. Bruyninckx
University of Mons-Hainaut, Mons, Belgium

D. Poch-Olivé
Autonomous University of Barcelona, Barcelona, Spain

ABSTRACT

Five Frenchspeaking normal subjects and five Frenchspeaking stutterers have been recorded in two conditions (reading task and map task), under four DAF delays. The formants frequencies of vowels [i], [a] and [u] have been measured. Their statistical treatment suggests important articulatory changes in stutterers between the two conditions under the 80 ms delay.

INTRODUCTION

In 1950, Lee found out that a delayed auditory feed-back (DAF) can induce a speech trouble ("artificial stuttering") in normal subjects. This effect is now known as the *Lee-effect*. On the other hand, other studies showed that DAF can improve stammerers'fluency [1,2,3]. Some authors suggested that the delay inducing the maximal trouble in normal subjects'speech is within the 120-160 ms range. It is generally estimated that a DAF delay of ca. 80 ms causes the most spectacular effects on stammerers' speech.

It has been argued that these beneficial effects might be related to speech modifications induced by DAF, such as extension of the production length, slackening of speech, and increase of the fundamental frequency [1,2,3,4,5].

Most research in this field have nevertheless involved English language; very few information is therefore available about French. Moreover, topics like influence of the speaking style, or variations in the vowels systems structures have not been extensively investigated in this field.

In this paper we therefore study the repercussions of speaking styles changes and DAF delay variations on the vowels systems of Frenchspeaking subjects.

EXPERIMENTAL SETTING

Ten Frenchspeaking male subjects have been recorded. Five were normal, although the other five suffered from stuttering.

Two kinds of tasks were presented to the subjects. In the first one ("reading task"), the experimenter had the speakers read an extract drawn from a modern French novel. In the second one ("map task"), the subjects were asked to explain to a remote interlocutor how to travel from a given city to another. For this purpose, they were given a map indicating the names of the cities to go through, the types of roads, the special details (bridges, rivers,..), etc.. The imaginary travel and the map were specially conceived to have the subject speak as much as possible.

Each task was carried out under varying conditions of auditory feedback: normal (condition 1) and delayed auditory feedback (DAF conditions). Three delays were used, i.e., 80 ms (condition 2), 120 ms (condition 3), and 160 ms (condition 4).

Each subject had therefore to read 4 texts and to describe 4 journeys.

The delay was obtained by means of the Kay CSL 4300 DAF routine. All the recordings were performed in a sound proof room at the Phonetics Laboratory of the University of Mons, by means of a Neumann U87 P 48 microphone, connected to a Sony 501 ES PCM coder. The digitized sounds were stocked on a Panasonic VHS video recorder.

A sample of 240 vowels was extracted from the recorded corpus. Only the French vowels /i/, /a/ and /u/ are taken into consideration in this paper. For each subject, each task, each condition, one vowel was analyzed. They were selected in order to minimize context effects: the only common anterior context was /k/ for the collected [i], and /p/ for [a] and [u].

RESULTS

We measured, for each vowel, its first and second formants at its center, by means of a Kay 5500 DSP sonagraph.

As suggested by figure 1, the formants values tend to be more central in the reading task than in the map task. This could be due to the fact that, in the reading task, speakers behave in more a humdrum way, since the speech act can, in this case, appear as rather useless; for the map task, on the contrary, they have to communicate information perceived as usefull to an interlocutor, who, moreover, is believed to be far from the place where the speakers are. The map task could, therefore, provoke hyperarticulated speech movements although the reading task could provoke hypoarticulation.

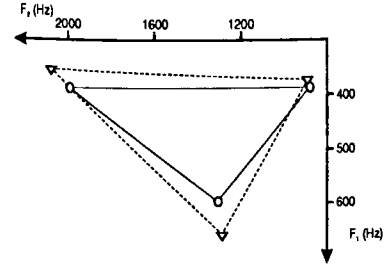


Figure 1. Average formants frequencies of the vowels in the reading- (circles) and the map- (triangles) tasks.

In order to evaluate the centralizing tendencies, and to assess their dependances upon the speakers types (normal vs stutters) and the DAF durations (0 ms, 80 ms, 120 ms, 160 ms), we processed them in a way inspired of Harmegnies and Poch [6,7]. Each vowel from the reading task was paired with the same vowel from the map task, drawn from the same subject, under the same DAF condition. We therefore obtained 120 pairs of vowels (10 speakers x 3 vowels X 4 DAF delays), each pair beeing characterized by two first formants frequencies and two second formants frequencies.

The formants values were converted to mels, prior to the statistical processing, by means of formula (1):

$$f_{mels}=2595(1+\frac{700}{f_{Hz}}) \qquad (1)$$

where f_{Hz} is the frequency in Hertz, and f_{mels} is the frequency in mels.

We thereafter computed a centralization index, δ, following equations (2) to (4):

$$ED_{map}=\sqrt{(F1-\bar{\Phi}1)^2+(F2-\bar{\Phi}2)^2} \qquad (2)$$

$$ED_{read}=\sqrt{(f1-\bar{\Phi}1)^2+(f2-\bar{\Phi}2)^2} \qquad (3)$$

$$\delta=ED_{map}-ED_{read} \qquad (4)$$

where ED stands for Euclidean Distance, Φ 1 and Φ 2 are the grand means of the first and second formants across the whole data base, F symbolizes formants values for the map task, f for the reading task, and overlining denotes averaging.

As can be observed in table 1, which gives a statistical summary of the computed values, δ is positive in all cases, confirming an overall tendency of the formants values to centralize in reading speech, relative to speech under the map task.

Back vowels seem, on the whole, less affected than front- and medium vowels. In normal speakers, [u] average δ value is close to zero, suggesting the existence of quasi invariant

articulatory gestures, whatever the speaking style. For stutterers, [u] δ value is nevertheless hardly 6 times as great as the one for normal speakers; it seems therefore that articulation of the back vowels can be affected by speaking style in stutterers, although it is not the case in normal subjects. This could be related to efforts involving the pharyngo-laryngeal area, that stutterers mobilize in their attempts to compensate for difficulties encountered in controlling laryngeal production.

Table 1. Values (means: "m" and standard deviations: "s") of the δ centralization index in normal speakers ("norm") and stutterers ("stut").

	[i]		[u]		[a]	
	m	s	m	s	m	s
norm	34.4	45.7	3.5	94.7	58.6	45.4
stut	47.9	64.9	24.1	261.0	43.1	49.9
both	41.2	55.8	13.8	194.1	50.9	47.8

The open vowel [a] is the most influenced in normal speakers, but not in stutterers. In order to try to interprete those findings, it is important to notice that inter style differences in formants values of [a] involve F1 (average difference of 65 Hz) quite more than F2 (average difference of 12 Hz). Greater centralization of [a], relative to the other vowels, could therefore be interpreted in terms of more important differences in aperture degrees. In other words, normal subjects seem able to change their degrees of [a] aperture more than the stutterers, under the effect of varying speaking styles.

As figure 2 shows, DAF in normal speakers does not sensibly affect the overall variation profile of δ. In stutters, on the contrary, a striking difference between centralization profiles under various DAF conditions is to be found. The 80 ms delay seems

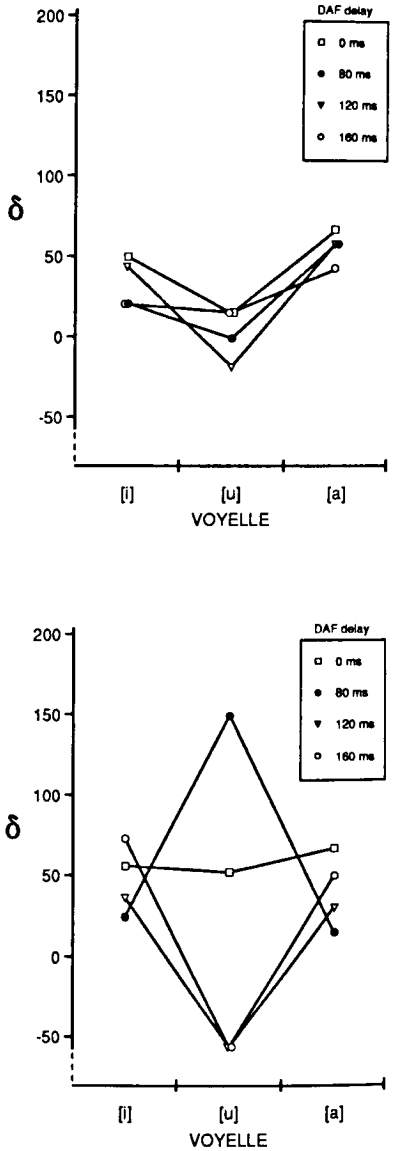


Figure 2. Average values of the δ centralization index in normal speakers (up) and stutterers (down), by vowel and DAF delay.

to arouse spectacular differences, specially involving vowel [u]. This finding could be in agreement with previous research on DAF usefulness in stammer treatment: it is generally found that a DAF delay in the 50 ms - 100 ms range improves stutterers fluency. In the case of this research, inter style increased variablity of the centralization of [u] might be associated with more variable articulatory gestures in the pharyngo-laryngeal region.

CONCLUSION

The data presented in this paper confirm previous observations [6,7] about the effects of speaking styles on the speech signal. As in those works, the characteristics of formants spaces suggest variations along an hyper-hypoarticulated speech axis, correlated with functionnal aspects of speech in the communication situation.

Here, reduction of the F1/F2 space is associated with the reading task, although speech under the conative task is characterized by increased spreading of the formants values.

In normal subjects, the variation in DAF delays does not change the overall inter style variation profile. In stutterers, on the contrary, the relationship between centralization values is modified by the delay. The 80 ms delay provokes phenomena that are not to be found with other delays.

Further reseach should investigate the reasons for that specific effect and try to relate the findings with claims that 50-100 ms DAF delays help stutterers improve their speech production.

ACKNOWLEDGEMENT

Data presented in this paper have been collected by F. Leeuwerck under the direction of Profs. A. Landercy and B. Harmegnies, at the Laboratory of Phonetics, University of Mons-Hainaut, Belgium.

REFERENCES

[1] ANDREWS, G., HOWIE, P.M., DOZSA, M. & GUITAR, B.E. (1982), "Stuttering: speech pattern characteristics under fluency-inducing conditions", *J. S. H. R.*, 25, pp. 208-216.

[2] BORDEN, G.J., DORMAN, M.F., FREEMAN, J.J. et RAPHAEL, L.J. (1977), "Electromyographic changes with delayed feedback of speech", *J. of Phonetics*, 5, pp. 1-8.

[3] STEPHEN, S.C.G. and HAGGARD, M.P. (1980), "Acoustic properties of masking/delayed feedback in the fluency of stutterers and controls", *J. S. H. R.*, 23(3), pp. 527-539.

[4] BURKE, B.D. (1975), "Variables affecting sutterer's initial reactions to delayed auditory feedback", *J. of Com. Disorders*, 8, pp. 141-155.

[5] KNAUSS LECHNER, B. (1979), "The effects of delayed auditory feebback and masking on the fundamental frequency of stutterers and nonstutterers", *J. S. H. R.*, 22, 7a, pp. 343-353.

[6] POCH, D. et HARMEGNIES, B. (1992), "Variations structurelles des systèmes vocaliques en français et espagnol sous l'effet du style de parole", *J. Physique, IV, C1, pp. 283-286.*

[7] HARMEGNIES, B. et POCH, D. (1992), "A study of style-induced vowel variability: laboratory versus spontaneous speech in Spanish", *Speech Communication*, 11, pp. 429-437.

NUMBER OF POSSIBLE BASIC VOWEL QUALITIES AND THEIR PSYCHOACOUSTICAL DISTANCE MEASURE

Antti Iivonen
University of Helsinki, Department of Phonetics

ABSTRACT

Spatial representation for the psychoacoustical vowel space and vowel resolution are discussed. A simple, but proper spatial approximation for the vowel space of the basic (major) vowel types is an F2/F1 vowel chart with Bark scales in which the F1 dimension is enlarged 60% in relation to F2. For other vowel types additional parameters are needed. If we want to display the spatial vowel resolution, a good approximation is achieved by a Critical Band Window (CBW-F1=1 Bark sized circle according to the F1 scale).

PSYCHOACOUSTICAL VOWEL SPACE (PVS)

Several suggestions have been presented in the literature for a simulation of listener's *psychoacoustical vowel space (PVS)*. Psychoacoustical scales used include musical, Koenig, full logarithmic, mel, and Bark scale. The most usual parameter combination applied is F2/F1, but additionally several modifications of the F2/F1 space have been suggested: F2'/F1, F2–F1/F1, F2–F1/F1–F0. Some suggestions involve a three-dimensional vowel space (F1/F2/F3).

A successful spatial simulation of a PVS implies that every equal spatial distance corresponds to an equidistant psychoacoustical distance. This issue concerns the *scale problem.* The perceptual role of F3 in some vowel types should be discussed. Some parameter combinations (e.g. F2–F1/F1–F0) involve the question, whether computations are really carried out in the perceptual processing. One problem concerns the *vowel resolution.*

THE SCALE PROBLEM

In order to elucidate the scale problem, I made some experiments with the set of vowels proposed by Lindblom [5]. He calculated the first four formant values of 19 "quasi-cardinal vowels" representing psychoacoustically equal quantization steps. Lindblom used among other things the whole spectrum approach and Plomp's auditory distance metrics for calculations. For [i] and [y], Lindblom has the same F2 value probably due to the natural Swedish vowel formants. Many other languages like Finnish, French, and German have separate F2 values for those vowels. It seems to be plausible to add an hypothetical [i] to the set (cf. Fig. 3 below).

The 19 vowels are presented in Fig. 1 according to full Bark scales. They are connected with lines. Equal psychoacoustical distances would imply that the nearest three vowels should in all cases form spatially equilateral triangles. Fig. 1 shows that this is not the case. The major observation is that the distances are on the average longer concerning F2. There are additionally some minor irregularities.

The conclusion is that F1 must be enlarged in relation to F2 in order to get a better spatial representation for (averaged) equal distances. The F1 dimension of an F2–F1/F1 space has been enlarged 100% in relation to F2 in [4] in order to achieve a better correspondence between the display and the phonetic experience. The calculations of the average distance needed for the enlargement in F2/F2 space showed that the proper enlargement should be approx. 60%. The corresponding modification of the vowel space has been carried out in Fig. 2.

Additional criteria for a PVS are the number of qualitative degrees in IPA vowel set. The maximal number of "horizontal" vowels is three which corresponds acoustically to six because of the effect of rounding on F2. The "vertical" dimension is problematic. There are four main degrees in IPA, but the intermediate degrees confuse the interpretation. Fig. 3 (below) seems to suggest that there exists room for 6 horizontal and 5 vertical (psychoacoustical) degrees.

It must be noticed, however, that the influence of F3 has been neglected so far. Its perceptual contribution is that the front vowels become brighter by means of the combined effect of F2 and F3 (cf. the discussion of perceptual integration in [7]).

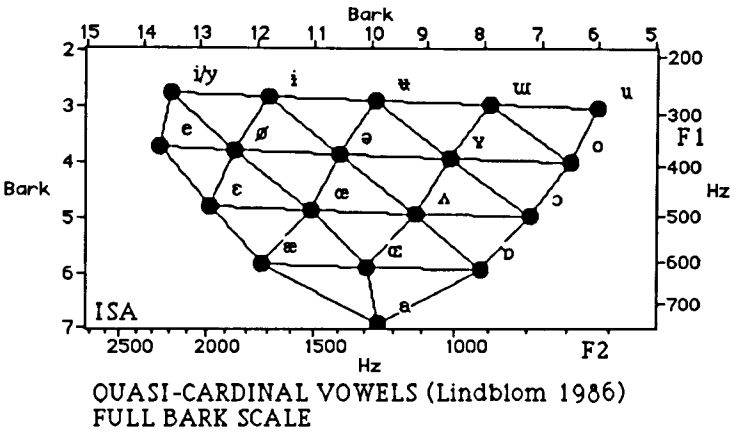


Figure 1. Lindblom's (1986) 19 quasi-cardinal vowels presented in a F2/F1 space according to full Bark scales. (The figures are produced by means of Intelligent Speech Analyser (ISA) developed by Raimo Toivonen.)

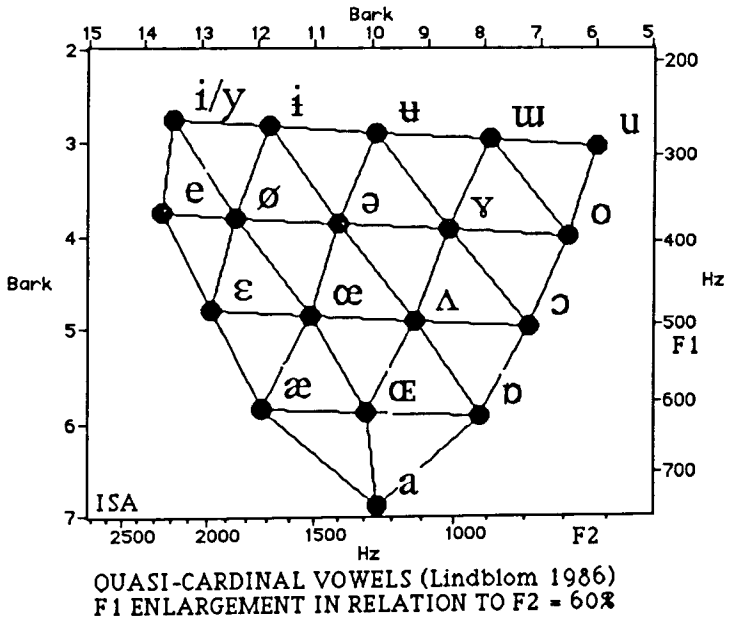


Figure 2. The same vowels as in Fig. 1 presented in an F2/F1 space according to full Bark scales with the enlarged F1 dimension After the enlargement (60%). the spatial distances correspond on the average better the psychoacoustical equi-distancies between the vowel points.

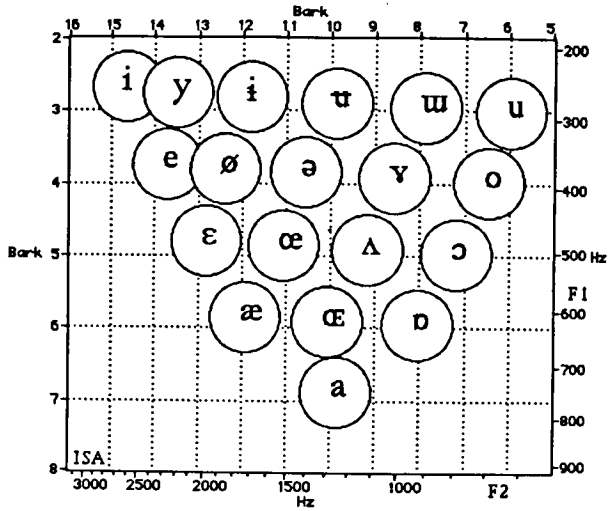


Figure 3. The same vowel set as in Fig. 1–2 presented as 1 Bark sized circles (Critical Band Windows according to F1 dimension). A hypothetical [i] added. Note the overlapping of [i] and [e] with [y] and [ø].

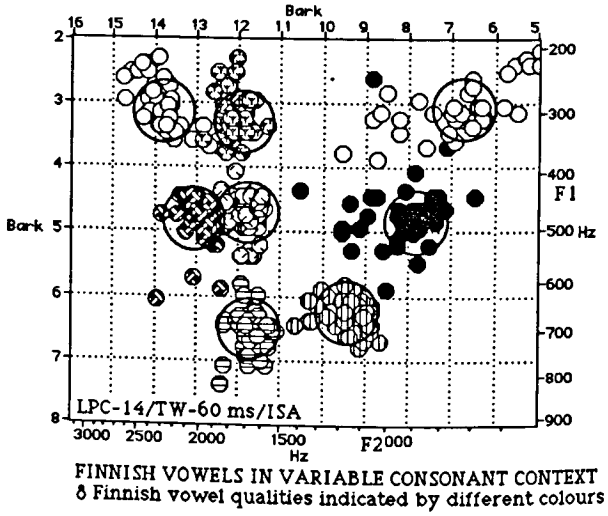


Figure 4. A scattergram of Finnish stressed vowel occurrences (N=352) representing 8 phonemic qualities produced by a male speaker AA in two and three syllable words. The vowels occurred in 11 different symmetrical consonant contexts. A Critical Band Window (CBW-F1) is drawn on the densest accumulation of the single occurrences of each phoneme.

If the F3 is not taken into consideration, the spatial distances between [e] and [ø] (cf. Fig. 1 and 2) as well as between [i] and [y] become too short (these effects can been seen also below in Fig. 3).

The conclusion is that F2/F1 representation (with enlarged F1) can be characterized as an *approximate framework* for basic or *major* vowel types (cf. the notion in [4]), but it cannot be a proper psychoacoustical framework for *all possible* vowel qualities. For [i], [e], [y], and [ø] F2' could be used for display. Besides, the F2/F1 charts must be understood as relative maps, because the absolute vowel positions depend on the vocal tract length. Other articulatory factors should be discussed, too. It is, however, another issue.

PHONETIC VOWEL RESOLUTION

According to Flanagan's experiments [1], the difference limen (DL) for formant perception of *synthetic* vowels is some 3-5%. That small DL would imply that there exist more than 400 perceptually different vowel qualities in the F2/F1 space [2]. The number of possible phonetic major vowel types must be much smaller. Flanagan's result [1] seems to reflect more the *discrimination* of vowels than their phonetic *perception*. Besides, the perception of *natural* vowels is another issue. Nagakawa *et al.* [6] observed 6–13% DLs in the perception of the F2 differences.

Instead of the 3–5% resolution, I suggested [2] that the Critical Band of the ear could be a proper measure for phonetic vowel resolution. If the whole F2/F1 vowel space (with *full Bark* scales) is filled with one-Bark-sized windows (= Critical Band Window, CBW; each comprising one critical band, i.e. 1 Bark), we get about 45 different vowel points [2]. If we fill the vowel space, which has an *enlarged F1* dimension, we get the illustration of Fig. 3. It shows the 19 vowel points discussed above, which comprise 1 Bark each (CBW-F1), according to the F1 scale. A hypothetical vowel [i] has been added. There remains empty space for some additional CB windows. A fronted [a] and a back [ɑ] could be added. The number of CB windows is near that of the IPA vowels (the new IPA vowel chart in JIPA 23 (1) 1993 contains 28 basic vowel symbols). The suggested representation implies that F1 resolution is better than that of F2.

A Critical Band Window (specifically, CBW-F1) can be understood as an area that scans its surrounding space to check, if there exists a perceptually distinct psychoacoustical distance from other vowels. If two CBWs overlap, it can be assumed that the listener may have difficulty distinguishing the vowels considered.

The relationship of speech production and speech perception is a complicated issue. An example of the CBW application is shown in Fig. 4. The scattergram of Finnish vowels and a CBW-F1 on each vowel phoneme show that the major part of single occurrences of a phoneme are covered by a CBW-F1 window, but the distribution is larger especially in /u/ and in /o/ (data from [3]). The [u] and [o] variants to the left of the CBW-F1 represent mainly short vowels and dental contexts.

ACKNOWLEDGMENT

I am grateful to Eli Fischer-Jørgensen for inspiring discussions.

REFERENCES

[1] Flanagan, J. L. (1955) A difference limen for vowel formant frequency. *JASA*, vol. 27, pp. 613-617.

[2] Iivonen, A. (1987) The Critical Band in the Explanation of the Number of Possible Vowels and Psychoacoustical Vowel Distances. Mimeogr. Series of the Dept. of Phonetics, Univ. of Helsinki 12.

[3] Iivonen, A. & Laukkanen, A.-M. (1993) Explanations of the qualitative variation of Finnish vowels. *Publications of the Department of Phonetics, University of Helsinki*, Series B: 5, 29–54.

[4] Ladefoged, P. & Maddieson, I. (1990) Vowels of the world's languages. *J. Phonetics*, vol.*18*, pp. 93–122.

[5] Lindblom, B. (1986) Phonetic universals in vowel systems. In Ohala, J. J. & Jaeger, J. J. (eds.), *Experimental Phonology*, pp. 13–44. Orlando: Academic Press.

[6] Nagakawa, T. Saito, T. & Yoshino, T. (1982) Tonal difference limens for second formant frequencies of synthesized Japanese vowels. *Annual Bulletin of the Research Institute of Logopedics and Phoniatrics*, vol. 16, pp. 81–88.

[7] Schwartz, J.-J., Beautemps, D., Abry, C. & Escudier, P. (1993) Inter-individual and cross-linguistic strategies for the production of the [i] *vs.* [y] contrast. *J. Phonetics*, vol. 21, pp. 411–425.

A DYNAMIC APPROACH OF VOWELS SYSTEMS IN ITALIAN

D. Poch-Olivé
Autonomous University of Barcelona, Barcelona, Spain

B. Harmegnies
University of Mons-Hainaut, Mons, Belgium

ABSTRACT

A native speaker of Italian was first recorded on the occasion of a natural conversation. After transcription of the whole corpus, 210 words, each containing one token of a given vowel, were selected. They were read in laboratory conditions by the speaker. A preliminary statistical treatment, focused both on the [e]/[ɛ] and the [o]/[ɔ] differences, confirms that whereas tendencies to merge those categories exist in Italy, the recorded informant does use a 7-vowel system. The formants frequencies are further processed both by means of the δ index and of discriminant analysis. The statistical treatment confirms the existence of phenomena similar to those observed in other languages of the Romance group, from both qualitative and quantitative points of view.

INTRODUCTION

Harmegnies and Poch have been carrying out a joint study project focused on the dynamics of vowels systems under the effect of speaking style. Their experiments have dealt with languages of the Romance group, i.e., Castilian Spanish [1], Catalan [2], Belgian French [3] and European Portuguese [4]. Languages in this group are very interesting, because whereas they derive from the same origin, they are characterized by vowel systems differing one from another in quite a large variety of ways, i.e., 1. by the number of units in the system (from 5: Spanish, up to 14: French); 2. by the presence of phonological reduction in the system (Catalan and Portuguese) or its absence (Spanish and French); 3. by the existence of central vowel(s) in several of them (Catalan, French and Portuguese) although Spanish has no such vowel.

Laboratory speech (words lists reading) and spontaneous speech (spontaneous conversation), at least, have been considered in all those languages. When compared to laboratory speech, spontaneous speech may be characterized by 1.schwa-like tendency (not in all the languages); 2.reinforced timbres variability (not in all the languages); 3.increased overlapping of the vowels clusters (in all the languages).

Further research requires to use other languages both in order to evaluate the universality of the phenomena observed, and to seek a wider understanding of them.

Particularly, among the languages involved up to now, Spanish is the only one without central vowel; the study of another language characterized by the same feature therefore appears quite desirable.

In this paper, we apply the previously developped methodology to Italian, which possesses a seven-unit system, without central vowel, and free of phonological reduction.

EXPERIMENTAL SETTING

The *spontaneous speech* sample was drawn from a natural conversation held with a male Italian speaker. This was born and lived in Napoli, and spoke standard Italian. The talk lasted about one hour and took the form of a semi-directive interview led by an experimenter, where the subject was suggested to evoke various themes, such as Italian food, his birth place, or current occupation.

On the basis of an exhaustive transcription of the whole recording, a set of 210 vowels (i.e., 30 samples of each of the 7 Italian vowels) was randomly selected. Each word containing a selected vowel was randomly put in a list that the speaker was afterwards asked to read, in order to produce the *laboratory speech* sample.

RESULTS

The sampled vowels

Finally, 420 vowels, organized in pairs with invariant status in spontaneous- and lab speech, were retained. Their first and second formants frequencies were measured both in spontaneous- and laboratory speech by means of a DSP 5500 KAY analyzer, which provided a 20 Hz resolution on the whole frequency span. These frequency values are summarized in tables 1 and 2.

Table 1. Averages (m) and standard deviations (s) of the first formants frequencies in laboratory- ("Lab") and spontaneous ("Spt") speech.

	Lab		Spt	
	m	s	m	s
[i]	311.23	37.37	391.50	62.58
[e]	396.40	34.21	439.23	51.83
[ɛ]	442.50	33.14	505.76	34.20
[a]	591.26	65.52	595.33	69.71
[ɔ]	459.50	21.10	510.13	38.21
[o]	429.30	40.17	457.60	59.08
[u]	347.46	44.05	394.73	52.40

Table 2. Averages (m) and standard deviations (s) of the second formants frequencies in laboratory- ("Lab") and spontaneous ("Spt") speech.

	Lab		Spt	
	m	s	m	s
[i]	1951.53	71.17	1807.00	94.15
[e]	1697.60	100.66	1651.30	99.18
[ɛ]	1717.33	73.78	1603.13	96.22
[a]	1305.66	109.45	1326.70	104.88
[ɔ]	931.36	75.33	1092.03	158.00
[o]	988.86	150.47	1103.96	140.23
[u]	916.43	180.47	980.16	179.32

The vowel system

Our procedure of vowel selection was based upon the idea that the Italian vowel system is composed of the 7 open vowels [i], [e], [ɛ], [a], [ɔ], [o] and [u]. Nevertheless, as speakers of Italian tend nowadays to merge the realizations of /e/ and /ɛ/, as well as the ones of /ɔ/ and /o/, we tried to determine whether the realizations of the corresponding vowels by our informant were to be considered as belonging to different categories.

The 30 presumed [e] were compared to the 30 presumed [ɛ] in both styles. The comparison revealed significant differences along the first formant axis (laboratory speech: Mann-Whitney's U = 151, p<.0001; spontaneous speech: Mann-Whitney's U = 126, p<.0001), but not along the second formant axis (laboratory speech: Mann-Whitney's U = 405, p=.5; spontaneous speech: Mann-Whitney's U = 333, p=.0845).

The 30 presumed [ɔ] were, in turn, compared to the 30 presumed [o]. The comparison revealed significant differences along the first formant axis, as well, (laboratory speech: Mann-Whitney's U = 232, p=.0012; spontaneous speech: Mann-Whitney's U = 205, p=.0003), but not along the second formant axis (laboratory speech: Mann-Whitney's U = 395, p=.4155; spontaneous speech: Mann-Whitney's U = 394, p=.4113).

In other words, it seems reasonable to consider that our informant uses a 7-vowel system, since the /e/-/ɛ/ and the /ɔ/-/o/ pairs appear as significantly differenciated, at least relative to the aperture dimension.

The dynamics of the system

Previous research [1-4] has showed that vowels tend to centralize, i.e., to appear closer to the F1/F2 plane center, when uttered in spontaneous speech. In order to test this hypothesis, we computed, for each inter style pair of vowel, a

centralization index, δ [1]. This one is defined as the difference between the Euclidean distance from the observed vowel to schwa in laboratory speech and the Euclidean distance from the observed vowel to schwa in spontaneous speech, i.e., the amount of displacement towards schwa:

$$ED_{Lab}=\sqrt{(F1-500)^2+(F2-1500)^2} \quad (1)$$

$$ED_{Spt}=\sqrt{(f1-500)^2+(f2-1500)^2} \quad (2)$$

$$\delta=ED_{Lab}-ED_{Spt} \quad (3)$$

where **ED** stands for Euclidean Distance, **F** symbolizes formants values for the laboratory speech sample, and **f** for the spontaneous speech sample. As can be predicted from equation 3, positive values only of δ denote centralization, the magnitude of which is measured by δ. Moreover, as δ is a difference index, its significance can be tested by means of paired two-sample inferential procedures; in this case, the null hypothesis is that the difference between ED_{Lab} and ED_{Spt} does not significantly differ from zero.

Values of the δ index are presented in table 3, together with the results of inferential tests. The paired Student t test has been used, as well as the Wilcoxon matched pairs T test. As can be seen from table 3, centralization turns out to be significant for all vowels, but [a]. It is to be noticed that since both the parametic and the non-parametric procedures deliver the same conclusions, possible artifacts caused by the shapes of the δ distributions should be considered very unlikely. Also, the results of the ranked-based procedure would be held constant under any monotonic transform of the frequencies (such as mel transform, e.g.).

The centralizing tendency revealed by our treatment goes together with increase of the surface occupied by the vowels in the F1/F2 plane, in spontaneous speech, relative to laboratory speech. This can be observed from the relationships between the formant variabilities. As shown in tables 1 and 2, the standard deviations of the first formants values are systematically greater in spontaneous speech than in laboratory speech. The tendency is less obvious in F2, where the standard deviations are less different one from another.

Table 3. δ values, student paired t-test statistic ("t") with probability under the null hypothesis ("p"), and Wilcoxon matched pairs test normal approximation ("z_w") with probability under the null hypothesis ("p").

	δ	t	p	z_w	p
[i]	158.6885	8.58	<.001	4.72	<.0001
[e]	47.0357	2.99	.006	2.56	.0104
[ɛ]	102.6575	7.45	<.001	4.65	<.0001
[a]	11.5071	.68	.499	.54	.5857
[ɔ]	152.3713	6.52	<.001	4.62	<.0001
[o]	115.7178	4.29	<.001	3.43	.0006
[u]	72.1478	2.80	.009	2.48	.0132

The combination of the decreased formants differences caused by centralization, together with the increased formant variability within each vocalic category decreases differenciations in the whole F1/F2 spontaneous speech system: not only are the vocalic clusters closer one to another in the F1/F2 space, they are moreover less homogeneous. The spontaneous speech system therefore seems to have reached a more pronounced degree of disorganization than the laboratory speech one.

As a general rule, increased entropy in any system implies that the system is less informant: one may therefore expect the lab speech F1/F2 system to convey more information than the spontaneous speech one. Recognizing the elements of the system should therefore be more a hazardous task in the spontaneous- than in the lab speech sample.

In order to test this hypothesis, we performed 2 discriminant analyses (one in spontaneous speech and one in lab speech), with the vowels as *a priori* categories and the formant values as discriminant variables. Once computed the discriminant functions, a recognition task was simulated in each subsample. Their results are presented in tables 4 and 5.

Table 4. Confusion matrix from the simulated vowel recognition task in laboratory speech. Actual groups are in rows, and predicted group membership in columns.

	/i/	/e/	/ɛ/	/a/	/ɔ/	/o/	/u/
/i/	*93.3*	3.3	3.3	0	0	0	0
/e/	6.7	*70.0*	23.3	0	0	0	0
/ɛ/	0	16.7	*83.3*	0	0	0	0
/a/	0	0	3.3	*90.0*	0	6.7	0
/ɔ/	0	0	0	0	*86.7*	13.3	0
/o/	0	0	0	0	40.0	*46.7*	13.3
/u/	0	0	0	0	3.3	6.7	*90.0*

Table 5. Confusion matrix from the simulated vowel recognition task in spontaneous speech. Actual groups are in rows, and predicted group membership in columns.

	/i/	/e/	/ɛ/	/a/	/ɔ/	/o/	/u/
/i/	*76.7*	16.7	6.7	0	0	0	0
/e/	16.7	*60.0*	23.3	0	0	0	0
/ɛ/	0	10.0	*86.7*	0	0	3.3	0
/a/	0	0	6.7	*80.0*	6.7	6.7	0
/ɔ/	0	0	3.3	6.7	*73.3*	13.3	3.3
/o/	0	0	3.3	3.3	33.3	*33.3*	26.7
/u/	0	3.3	0	0	3.3	20.0	*73.3*

The recognition procedure clearly appears safer when performed on the basis of laboratory speech samples. The overall correct recognition is, in this case, 80%, although it decreases to 69% in spontaneous speech. This observation thus confirms that vowels in the F1/F2 plane are more differenciated in laboratory- than in spontaneous speech.

CONCLUSIONS

This study confirms, for Italian, tendencies already pointed out for other languages of the Romance group. Vowels in spontaneous speech are realized closer to the F1/F2 plane center; they moreover are better differenciated one from another in laboratory- than in spontaneous speech. Although drawn from a single-speaker experiment, those findings constitute an interresting account to the study of the universality of the reported phenomena. Further research should nevertheless seek to confirm the findings, and to refer observed variabilities to the ones caused by interindividual differences and sociolinguistic factors.

ACKNOWLEDGMENT

The recordings have been collected by David Puigvi, at the CIRASS (University of Napoli), under the supervision of Dr P. Maturi. Transcriptions have been performed by M. Piccaluga.

REFERENCES

[1] HARMEGNIES, B., POCH-OLIVE, D., "A study of style-induced vowel variability: laboratory versus spontaneous speech in Spanish", *Speech Communication*, 1992, 11, 429-437.

[2] BLECUA, B., POCH-OLIVE, D., HARMEGNIES, B., "Variaciones en la organizacion de los sistemas vocalicos del espanol y del catalan en funcion del estilo de habla", *Actas de las Jornadas Internacionales de Linguistica Aplicada*, Universidad de Granada, 1993, 22, 19-31.

[3] POCH-OLIVE, D., HARMEGNIES, B., "Variations structurelles des systèmes vocaliques en français et espagnol sous l'effet du style de parole", *Journal de Physique*, 1992, 4, 283-286.

[4] DELPLANCQ, V., HARMEGNIES, B., POCH-OLIVE, D., "Effets du style de parole sur la réduction vocalique en portugais", to be printed in *Verbum*.

TESTING THE DISPERSION-FOCALIZATION THEORY: PHASE SPACES FOR VOWEL SYSTEMS

J.L. Schwartz, L.J. Boë, N. Vallée
Institut de la Communication Parlée, Grenoble, France

ABSTRACT

The Dispersion-Focalization Theory (DFT) attempts to predict vowel systems thanks to a competition between two perceptual costs, namely global dispersion vs local focalization. The competition is controlled by two parameters, α and λ. We describe a new methodology for testing the DFT predictions: for a given number of vowels, phase spaces allow to determine the DFT winner in the (α, λ) space. We derive an (α, λ) region for which the theory predictions fit quite well with the phonological inventories.

1. INTRODUCTION

Substance-based theories of linguistic systems propose a deductive approach, which looks at the primary language-specific facts from an external point of view by considering non-linguistic constraints on possible speech sounds. Stevens [1] and Liljencrants and Lindblom [2] introduced the basic categories of arguments about the nature of the listener-speaker interaction and its role in shaping phonological systems, namely Lindblom's Dispersion Theory (DT) and Stevens' Quantal Theory (QT). The principle of the Dispersion-Focalization Theory (DFT) is to set a competition between a structural dispersion cost based on inter-vowel perceptual distances and a local focalization cost based on intra-vowel perceptual salience [3].

2. IMPLEMENTING THE DFT

2.1. Cost of a vowel system

The DFT assumes that for a given number of vowels, namely n, the preferred system (i.e. the most frequent in phonological bases) is obtained by minimizing a global cost summing two components, namely a structural Dispersion cost and a local Focalization cost, both applied on acoustic parameters characterizing each vowel. Vowels are described in our work by four formants (F1, F2, F3, F4), with F4 fixed at 3560 Hz, and all values expressed in bark, as computed by the formula proposed in [4]: bark = 7 ArgSh (Hz / 650)

The energy function of a given system with n vowels V^i, $i \in \{1, .., n\}$, is given by:

$$E_{DF} = E_D + E_F$$

with E_D a dispersion cost and E_F a focalization cost. E_D is defined, as in the DT, by :

$$E_D = \sum_{\substack{i = 1 \,..\, (n-1) \\ j = (i+1) \,..\, n}} (1/d_{ij})^2$$

with d_{ij} the perceptual distance between vowels V^i and V^j. To compute this distance, we use an Euclidian distance in the (F_1, F'_2) space, where the "second perceptual formant" F'_2 is computed from F_2, F_3 and F_4 on the basis of a model we have developed in the 80s [5]. In order to deal with the excessive number of high non-peripheral vowels in the DT predictions, we introduce Lindblom's proposal of a "stretching" of the acoustic space in the F_1 dimension [6] by using an (F_1, F'_2) weighted Euclidian distance, namely:

$$d_{ij} = [(F1^j - F1^i)^2 + (\lambda F'2^j - \lambda F'2^i)^2]^{1/2}$$

where λ can be chosen at any value lower than 1, assuming that higher formants play a lesser part in vowel phonetic quality than do lower ones.

The DFT discards from the DT by the introduction of a second energy term, called focalization cost, diminishing the energy of configurations with vowels with close F_1 and F_2, F_2 and F_3 or F_3 and F_4 ("focal vowels", [7]) and hence making such configurations more stable. This cost is defined by:

$$E_F = \alpha (E_{12} + E_{23} + E_{34})$$

with

$$E_{12} = -\sum_i 1/(F_2^i - F_1^i)^2$$

$$E_{23} = -\sum_i 1/(F_3^i - F_2^i)^2$$

$$E_{34} = -\sum_i 1/(F_4^i - F_3^i)^2$$

where α is a second free parameter. We thus obtain an energy function depending respectively on the parameter λ, which sets the weighting between F_1 and F'_2, and the parameter α, which determines the weighting of the additional focalization cost.

2.2. Selection criterion

Various criteria have been proposed in the literature for selecting vowel configurations. Whatever the criterion, a crucial methodological point concerns the way one deals with the well-known impossibility to analytically derive the solution of a non-linear minimization process (namely with non-quadratic energy landscapes, resulting in local minima). The original solution we have adopted here is based on what we call the "phase space" in reference to a classical procedure in Chemistry. In this method, we a-priori define a number N of "prototypes" covering all the vowel space, and for each number n we attempt to determine, according to the values of the free parameters λ and α, which is the system, made of n vowels selected among the N prototypes, which minimizes the total energy E_{DF}. Hence the problem becomes tractable: it consists in choosing one between a finite number (in theory, C_N^n) of systems, thanks to an associated variable E_{DF}. We use 33 prototypes with positions as "regular" as possible, in terms of distances in the (F_1, F'_2) space. The methodology may be summarized in the following way: For each value of n, determine the "phase space ", namely the regions in the (λ, α) space in which a given system of n vowels chosen among our 33 prototypes "wins", in the sense that it has the minimal E_{DF} cost in respect to all his concurrents. These "phase spaces" are then compared with phonological inventories, derived from the UPSID Database [8].

3. PHASES SPACES FOR SYSTEMS FROM 3 TO 7 VOWELS

3.1. Simulation results

The methodology described in Section 2 allowed us systematically determine the "phase spaces" for all values of n between 3 and 9. We shall concentrate the discussion on values of n from 3 to 7, which provide the most significant trends in the UPSID basis. The results are given in Fig. 1 to 5, which display, respectively for n = 3, 4, 5, 6 and 7, the best system for a given value of the pair (λ, α) in the square region [0, 1] x [0, 1]. These results provide the following trends.

Decreasing λ favours peripheral systems, which is exactly what we expected, since it results in vertically shrinking the vowel space. A more unexpected consequence is that a too small value of λ leads to either reduced (Fig. 1, 2) or asymmetrical (Fig. 3, 5) peripheral configurations, since the interactions between the front and the back side of the peripheral systems become important.

Increasing α favours focal vowels, namely first [i] and [y], then front unrounded vowels with the highest focalization benefit for the highest vowels, and finally back rounded vowels, which have all the same E_F cost. This mainly results in switching from a central high vowel (be it [ɨ] or [ɯ]) to an [y] . It may also produce a switch from a peripheral vowel to an [y].

Increasing n increases the dispersion cost of peripheral systems, hence it decreases the λ boundary necessary for making these systems optimal. Conversely, large n values favour systems with one or even two non-peripheral high vowels (for n=6 or 7).

3.2. Comparison with UPSID data

UPSID inventories provide the following trends. Symmetrical peripheral systems are the great winners for an odd number of vowels, namely [i, a, u] for 3-vowels systems, [i, 'e', a, 'o', u] for 5-vowels systems and [i, e, ɛ, a, ɔ, o, u] for 7-vowels systems. For 4-vowels systems the dominant acoustic structure is

[i,'e',a,u], while for 6-vowels systems, if we discard systems including schwa, which is according to us a "special" vowel [9], the situation is balanced between [i, e, ɛ, a, 'o', u] and [i, 'e', a, 'o', u, i].

If one considers all these constraints together, the correct (λ, α) region is roughly defined by:

$0.2 \leq \lambda \leq 0.3$ and $0 \leq \alpha \leq 0.4$

Additional constraints based on the stability for systems which contain an [y] as the single non-peripheral high vowel, namely an unbalanced [i, y, u] structure for high vowels, provide a floor value for α higher than 0 (see [3, 10, 11]), namely:

$0.2 \leq \lambda \leq 0.3$ and $0.3 \leq \alpha \leq 0.4$

4. CONCLUSION

This study shows that we are able to define a region for the two DFT parameters for which theoretical predictions are quite in line with experimental data coming from the UPSID database.

The DF Theory provides some kind of generalization of the D Theory. Indeed, the first simulations by Liljencrants and Lindblom (1972) should correspond more or less to the results displayed in Fig. 1 to 5 for a value λ equal to 1 (same weighting for F1 and F'2, or M1 and M'2 in their terms, with formants expressed in mels) and a value α equal to 0 (no focalization). However, our simulations clearly show that λ must be much lower than 1 in order to solve the problem of peripheral vowels and α higher than 0 in order to solve the problem of front rounded vowels. This second point is crucial. It confirms that the focalization term *is* necessary for understanding the [i, y, u] structure for high vowels, which is not negligeable in the UPSID base, since it represents 4.5 % of the whole base, and more than 25 % of the structures with three high vowels (namely two peripheral and one non-peripheral). Therefore the DFT provides a good basis for understanding vowel systems in detail.

REFERENCES

[1] Stevens, K.N. (1972). The quantal nature of speech: Evidence from articulatory-acoustic data. In (E.E.Davis Jr. & P.B.Denes, eds.) *Human Communication: A unified view*, 51-66. New-York: Mc Graw-Hill.

[2] Liljencrants, J., & Lindblom, B. (1972). Numerical simulations of vowel quality systems: The role of perceptual contrast. *Language*, 48, 839-862.

[3] Boë, L.J., Schwartz, J.L., & Vallée, N., (1994). The prediction of vowel systems: perceptual contrast and stability. In E. Keller (Ed.) *Fundamentals of Speech Synthesis and Speech Recognition* (pp. 185-213). John Wiley.

[4] Schroeder, M.R., Atal, B.S., & Hall, J.L. (1979). Objective measure of certain speech signal degradations based on masking properties of human auditory perception. In (B. Lindblom, S. Ohman, eds.) *Frontiers of Speech Communication Research*, 217-229. London: Academic Press.

[5] Mantakas, M., Schwartz, J.L., & Escudier, P. (1986). Modèle de prédiction du 'deuxième formant effectif' F'2 – application à l'étude de la labialité des voyelles avant du français. *15th JEP, Société Française d'Acoustique*, 157-161

[6] Lindblom, B. (1986). Phonetic universals in vowel systems. In (J.J. Ohala, ed.) *Experimental phonology*, 13-44. New-York: Academic Press.

[7] Badin, P., Perrier, P., Boë, L.J., & Abry, C. (1991). Vocalic nomograms: Acoustic and articulatory considerations upon formant convergences. *Journal of the Acoustical Society of America*, 87, 1290-1300.

[8] Maddieson, I. (1984). *Patterns of sounds*. Cambridge studies in speech science and communication, Cambridge: Cambridge University Press.

[9] Schwartz, J.L., Boë, L.J., Vallée, N., & Abry, C. (1995). The Dispersion – Focalization Theory of vowel systems. Submitted to the *J. of Phonetics*.

[10] Vallée, N. (1994). *Systèmes vocaliques: de la typologie aux prédictions*. Thèse de Doctorat en Sciences du Langage, Université Stendhal, Grenoble.

[11] Jomaa, M., & Abry, C. (1992). *La base de systèmes vocaliques RHÔNSON*. Rapport PPSH 78B, ICP.

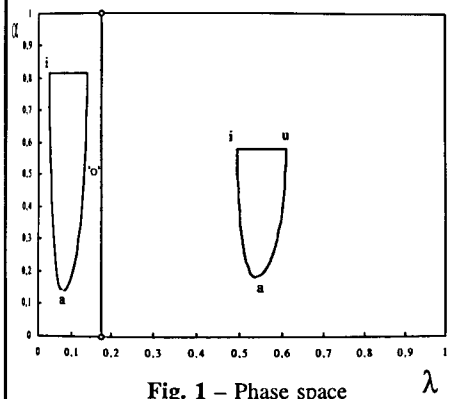


Fig. 1 – Phase space for 3-vowels systems

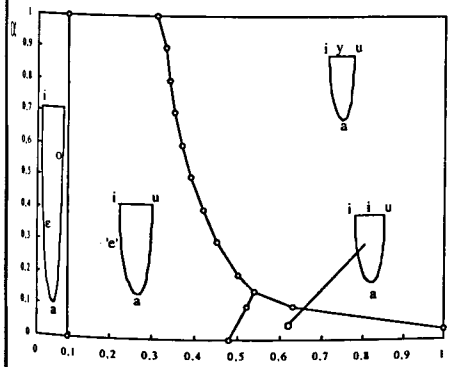


Fig. 2 – Phase space for 4-vowels systems

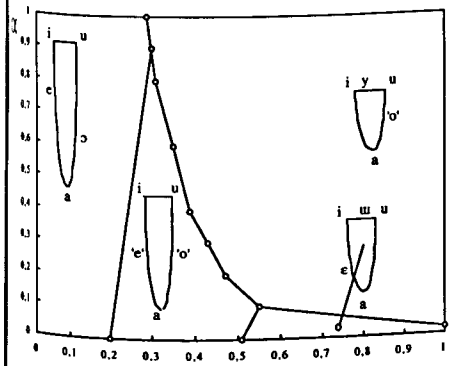


Fig. 3 – Phase space for 5-vowels systems

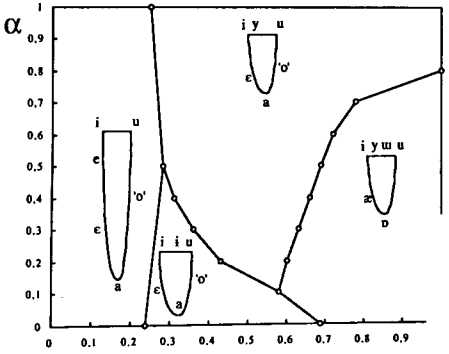


Fig. 4 – Phase space for 6-vowels systems

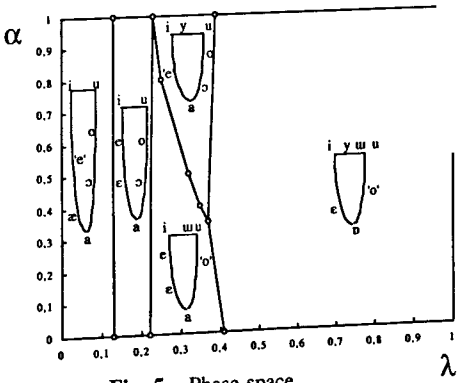


Fig. 5 – Phase space for 7-vowels systems

PHONETIC EVIDENCE FOR THE GREAT MONGOLIAN VOWEL SHIFT

Jan-Olof Svantesson
Dept. of Linguistics, Lund University

ABSTRACT

This paper presents acoustic phonetic evidence for the phonological shift that restructured the vowel and vowel harmony systems of East Mongolian.

BACKGROUND

In previous work [1] I have shown that East Mongolian (Khalkha and various Inner Mongolian dialects) has gone through a vowel shift resulting in a change of the vowel harmony system: while Old Mongolian had palatal (front-back) harmony, East Mongolian has pharyngeal (ATR) harmony.

The evidence in [1] for this vowel shift was acoustic data on the vowels of Khalkha and some other Mongolian dialects, as well as available descriptions of other Mongolian languages. Some uncertainty remains about the exact quality of the Old Mongolian vowels, however. The modern language whose vowel system is closest to Old Mongolian is Kalmuck (West Mongolian), but acoustic evidence for the Kalmuck vowel qualities was not available in [1].

PROCEDURE

During a visit to the Kalmuck republic in 1992 I recorded speakers of the two main Kalmuck dialects, Dörbed and Torgud, and I have also made further recordings of Khalkha and other East Mongolian dialects. Here I will present formant measurements for two Dörbed (Elst, Ovata) and two Torgud (Jaškul´, Astraxan´) speakers as well as for four Khalkha speakers (two from Ulaanbaatar, one from Bajanxongor and one from Zawxan), and for one speaker each of the Cahar and Baarin dialects, spoken in Inner Mongolia in China.

Each speaker read a list of words illustrating the vowels of his dialect (only male speakers were recorded). The words were read five times in isolation. The recordings were made on an analogue cassette recorder with fairly high quality. There is contrasting vowel length in Mongolian, but only long vowels were analyzed (except for Cahar and Baarin [ɪ] which only occurs short). The relevant words are given in Table 1. The first three formants were measured using the automatic formant tracking facility of the Soundscope program. The results are shown in Table 2, and F1-F2 diagrams for some of the speakers are given in Figure 2.

Table 1. Wordlists (the vowel in the initial syllable was analyzed).

Kalmuck	*Khalkha*	*Cahar*	*Baarin*	
biːr	biɤrte	piːr	piːr	
yːl			yːl	
		bɪr	bɪr	
			yːləx	öːl
eːrx	deːlte			
øːr				
ɛːrg			ɛːl	säːl
			œːr	öːl
baːlx	baːltai	dʒaːl	dʒaːl	
		dəːl	dəːl	
uːl	suːlte	suːl	suːl	
	dzʊɤtai	oːl	oːl	
boːdg	boɤrte	oːld	oːld	
	boːltɔi	bɔːl	bɔːl	

THE VOWEL SHIFT

The Mongolian vowel shift is illustrated in Figure 1, an F1-F2 diagram showing simultaneously the vowels of (Dörbed) Kalmuck and Cahar. The Kalmuck vowels are encircled, and arrows point towards the etymologically corresponding Cahar vowels. The Kalmuck vowel system is unchanged compared to Old Mongolian, except that a vowel phoneme /ɛ/ has developed by palatalization. Figure 1 thus illustrates the diachronic change from Old Mongolian to East Mongolian. Two different processes have reshaped the vowel system, backing (*y*>*u*, *ø*>*o*, *e*>*ə*) and pharyngealization (*u*>*ʊ*, *o*>*ɔ*), exemplified by:

Old M	*Kalm*	*Khalkha*	*Inner M*	
yge	yg	ug	ug	'word'
køl	køl	xol	xol	'foot'
degere	deɤr	deːr	dəːr	'top'
ula	ul	ʊl	ʊl	'sole'
tomo	tom	tɔm	tɔm	'great'

The main acoustic effects are F2 decrease and F1 increase, respectively. These processes are less consistent in Khalkha, where backing has not affected *e*. Vowel harmony in Old Mongolian and Kalmuck is manifested by the vowel alternation pairs *y*~*u*, *ø*~*o*, *e*~*a*, which differ in the front–back (palatal) dimension. In Inner Mongolian, these vowel pairs have become *u*~*ʊ*, *o*~*ɔ*, *ə*~*a*, still alternating in the same way in vowel harmony, which has thus become based on the feature pharyngeal (or ATR), which distinguishes the vowels of these pairs in East Mongolian [1].

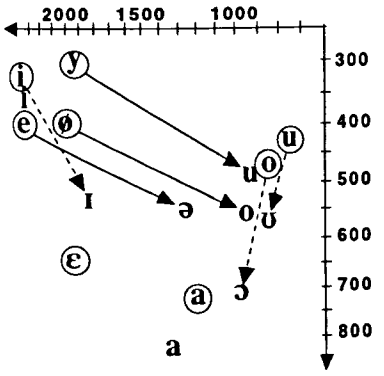


Figure 1. The Mongolian vowel shift.

The vowel *i* was neutral in Old Mongolian, in the sense that it could cooccur more or less freely with both front and back vowels in the same word. It has remained neutral in Kalmuck and Khalkha, but in Inner Mongolian it has split into two phonemes: it became *ɪ* in words containing a back vowel (which may have been lost, as in *bɪr* < *bira* 'strength'), and has remained *i* elsewhere. In this way, a fourth alternation pair *i*~*ɪ* was created and the neutral vowel was eliminated.

PALATALIZATION

East Mongolian lost the front rounded vowels by the vowel shift, but some dialects, including Baarin, have reintroduced them through palatalization [2]. Two different palatalization processes have affected the vowels in East Mongolian. One is due to old *i*-diphthongs (written V*ji* in the Classical Mongolian script), and the second is the result of the development of old VC*i* groups:

Old M	*Khalkha*	*Baarin*	
ajil	ail	ɛːl	'family'
ujila	ʊil	ɤːl	'cry'
ojira	ɔir	œːr	'near'
yjile	uil	yːl	'act'
sagali	säːlʲ	säːl	'milk'
uguli	öːlʲ	öːl	'owl'
ogoli	ɔːlʲ	ɔːl	'adze'

The two palatalization processes produced different results in Khalkha: the first one resulted in diphthongs, and in the second one, the original *i* palatalized the consonant and disappeared, resulting in a number of palatalized consonant phonemes contrasting with plain consonants (e.g. *bʲar* 'strength'; *bar* 'tiger'). The palatalized consonant affected the preceding vowel phonetically, indicated by an umlaut in the table above (see [2] for phonetic data), but because of the contrasting consonants, the palatalized vowels *ä, ö, ɔ̈* can be regarded as allophones of *a, ʊ, ɔ*.

The situation is different in Baarin, where consonant palatalization was lost, creating a contrast between umlauted and plain vowel phonemes. It also appears that the umlauted vowels merged with the vowels which developed from old diphthongs in this dialect so that, for instance, *öːl* 'owl' and *ɤːl* 'cry' became homophones. This was tested by comparing F1 and F2 simultaneously using Mahalanobis' D^2 test with the formant frequencies converted to the mel scale. This test was performed for the three pairs *ɛ*~*ä*, *œ*~*ɔ̈* and *ɤ*~*ö* with the result that there was no significant difference for the first two pairs ($F(2,7)=2.06$ and 0.74), while there was a significant difference between *ɤ* and *ö* ($F(2,7)=19.29$, $p<0.001$). It is necessary to investigate this question further before a final analysis can be made, but it is clear that at least four new vowel phonemes, /ɛ/, /œ/, /ɤ/ and /y/, have appeared in Baarin as a consequence of the palatalization processes (cf. Figure 2).

REFERENCES

[1] Svantesson, Jan-Olof (1985), "Vowel harmony shift in Mongolian", *Lingua*, vol. 67, pp. 283-327.

[2] Svantesson, Jan-Olof (1991), "Vowel palatalization in Mongolian". *Actes du XIIème Congres International des Sciences Phonetiques*, Vol. 5, 102-105. Aix-en-Provence: Université de Provence.

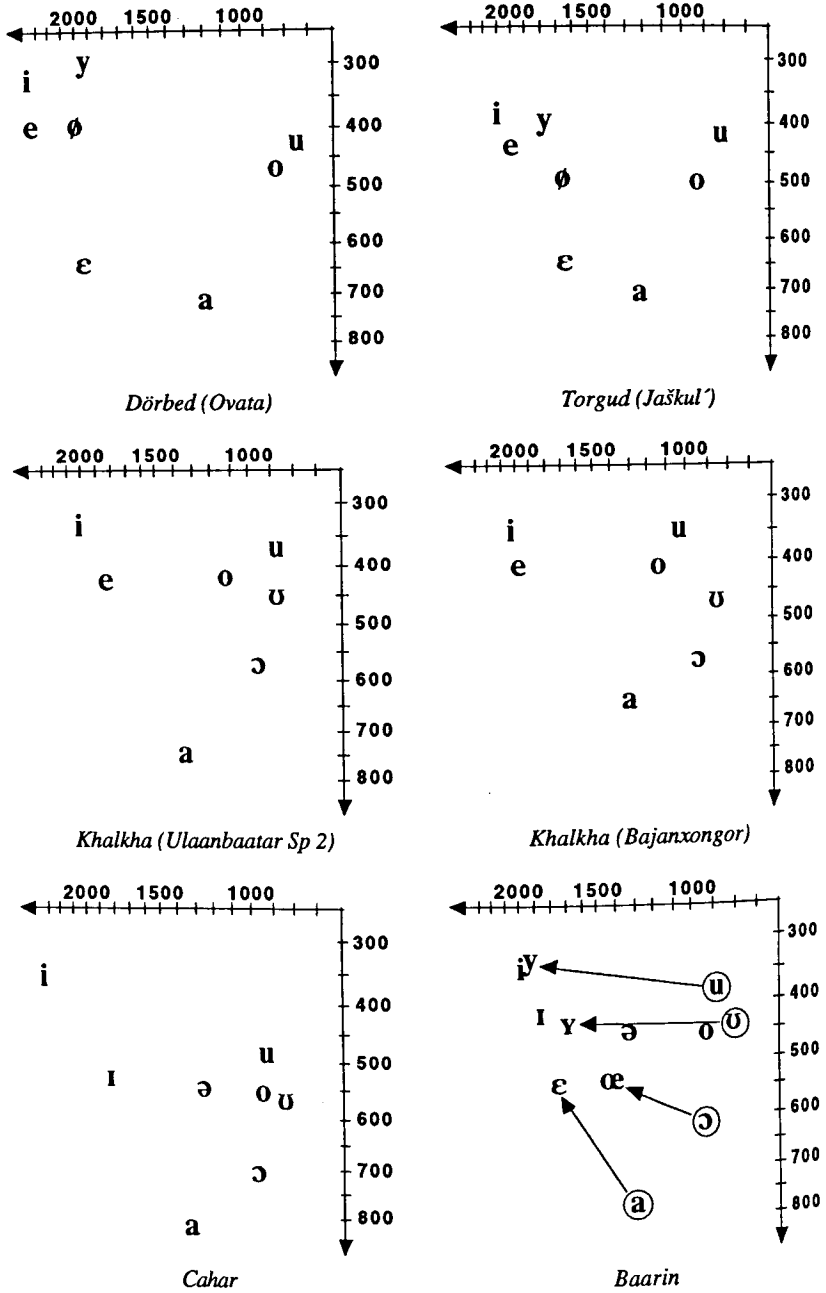


Figure 2. F1-F2 diagrams. Each vowel symbol represents the mean of five tokens.

Table 1. Formant values. For each vowel, the mean and standard deviation of F1, F2 and F3 are given, based on 5 tokens of each vowel.

Dörbed (Elst):

	F1		F2		F3	
i	347	0	2016	24	2711	39
y	347	0	1686	155	2242	66
e	434	31	1790	71	2503	129
ø	478	0	1460	24	2112	66
ɛ	642	19	1660	20	2373	39
a	669	24	1217	31	2486	36
u	349	26	661	38	2588	129
o	504	24	773	36	2625	132

Torgud (Jaškul'):

i	382	20	2103	50	2634	79
y	400	19	1790	71	2329	79
e	434	31	1999	0	2521	31
ø	495	24	1677	24	2443	19
ɛ	651	31	1660	36	2390	97
a	712	39	122	19	2008	128
u	417	24	825	31	2607	126
o	495	24	930	24	2164	78

Khalkha (Ulaanbaatar Sp 1):

i	382	23	2112	50	3042	82
e	460	24	1973	121	3008	64
a	782	0	1295	57	2668	95
u	391	0	1156	129	2503	66
ʊ	512	19	1017	24	2934	128
o	487	19	947	48	2625	109
ɔ	617	19	964	36	2216	53

Khalkha (Bajanxongor):

i	347	36	2010	22	2379	182
e	417	24	1973	50	2616	57
a	651	0	1303	31	2625	50
u	347	0	1020	41	2312	94
ʊ	469	20	869	69	2495	109
o	408	24	1147	24	2338	36
ɔ	574	19	938	24	2329	113

Cahar:

i	356	20	2347	81	3016	73
ɪ	539	39	1842	66	2755	24
a	825	0	1329	24	2851	90
ə	547	24	1269	36	2738	43
u	486	36	930	50	2712	58
ʊ	573	36	842	39	2886	143
o	556	48	947	78	2929	133
ɔ	704	19	982	24	2268	57

Dörbed (Ovata):

	F1		F2		F3	
i	321	24	2381	22	2972	39
y	295	20	1921	78	2503	24
e	400	19	2347	31	2938	73
ø	391	31	1992	78	2659	36
ɛ	642	19	1947	71	2973	109
a	717	25	1195	25	2846	202
u	426	48	751	62	2694	61
o	469	20	860	71	2834	48

Torgud (Astraxan'):

i	347	0	1921	40	2755	39
y	313	19	1521	97	2147	24
e	434	0	1756	39	2382	36
ø	443	20	1512	36	2216	31
ɛ	591	24	1512	20	2225	78
a	625	24	1130	44	2356	19
u	340	16	618	31	2378	47
o	452	24	851	73	2155	100

Khalkha (Ulaanbaatar Sp 2):

i	338	19	2016	50	2920	84
e	425	19	1825	134	2694	69
a	747	20	1347	31	2642	191
u	373	24	886	39	2356	158
ʊ	452	24	878	48	2069	175
o	417	24	1121	142	2407	314
ɔ	573	47	973	39	2190	117

Khalkha (Zawxan):

i	313	19	2129	0	3268	84
e	443	20	2008	19	2877	57
a	695	31	1225	36	2295	155
u	330	24	799	24	2329	39
ʊ	460	24	790	48	2094	84
o	426	36	956	31	2312	57
ɔ	547	39	930	59	2094	113

Baarin:

i	347	0	1981	39	2746	36
y	338	19	1938	73	2295	57
ɪ	425	19	1869	44	2651	31
ʏ	443	20	1686	37	2329	39
ü	425	48	1808	24	2408	66
ɛ	548	39	1764	50	2582	90
ä	582	24	1719	69	2686	56
œ	539	39	1425	128	2443	99
ɔ̈	513	36	1434	0	2416	58
a	773	36	1295	47	3068	58
ə	452	24	1321	58	2295	36
u	382	20	895	50	2399	78
ʊ	443	20	808	66	2416	117
o	452	24	930	24	2295	178
ɔ	617	19	940	10	2129	107

ON THE LEXICAL ASPECTS OF VOWEL DISPERSION THEORY: DUTCH CASE

Louis ten Bosch
Inst Phonetic Sciences, University of Amsterdam, Amsterdam, The Netherlands

Abstract
The 'vowel disperion theory' states that the structure of the vowel inventory in a language can be explained by optimization of acoustic inter-vowel contrast, given articulatory boundary conditions for each vowel. In this paper, the primacy of the acoustic properties is questioned by considering the possible effect of the lexicon on vowel dispersion. Here, the need for acoustic contrast between two vowels is assumed to be determined by the functional load of the vowel opposition in the lexicon. The results for Dutch indicate that the functional 'load' explains a part of the acoustic structure of the Dutch vowel inventory. Since the model is tested for one language only, we emphasize the used methodology, rather than the language-specific results.

1 Introduction

The set of phonemes in a language shows a large variety across languages. Universal trends in the structure of phoneme inventories (known as 'phonological universals') have been observed for a long time and attempts have been made to formulate them explicitly (e.g. Ruhlen, 1976; Crothers, 1978; Maddieson, 1984, 1991; Liljencrants & Lindblom, 1972; Lindblom, 1986 and later; Quantal Theory: Stevens, 1989; c.f. Ten Bosch & Pols, 1989). In general, the phonetic models of the structure of vowel systems start from two principles: (a) the reduction of articulatory effort, and (b) the optimization of inter-vowel acoustic contrast. There is much debate about the adequacy of these principles and their relative weighting. It is well known (see e.g. Ten Bosch, 1991) that a specification of the weighting is essential for the outcome of the optimization, but also less attention has been paid to the relation between the principle of acoustic vowel contrast and the functionality of this contrast (see Lindblom, 1972, 1986; Ten Bosch, 1991, chapter 4; Vallée, 1990). Moreover, with respect to the implementation of the contrast and effort principle, more elaborate models are available now and the vowel dispersion model as well as a general segment inventory model could now be based on articulatory synthesis models and advanced auditory models (An example of the use of more elaborate models is given by the SPEECH MAPS project, 1994).

In this paper, we want to address the point that the principle of 'acoustic contrast' is not based on the 'functional load' of vowel oppositions. For example, if a language has three vowels /a/, /i/ and /u/ that are spectrally specified by three target positions and many minimally pairing words with /i/ and /u/ and only a few with /a/, the need for acoustic contrast between /a/ and both other vowels is less that the need for contrast between /i/ and /u/. The 'need' for acoustic contrast between two vowels is (also) related to the structure of the lexicon and the frequency of words. Important spects of the model are focussed onto in the three following sections. Next, results will be presented for the Dutch case. The results are discussed in the concluding section.

2 Influence of lexical structure

Let us assume there are N vowels. For each vowel pair (v_1, v_2) we can select those words from the lexicon that form phonemically minimal pairs with respect to $v1$ and $v2$, resulting in a list L_1 consisting of words containing v_1 that each has one corresponding minimal opposing word containing v_2 in the list L_2. Additionally, the lists L_1 and L_2 are constructed so as to contain words with the same grammatical category to allow word confusion that is syntactically possible. Our basic assumption here is that the need for contrast between v_1 and v_2 is determined by the probability of confusion between L_1 and L_2, in other words, by the (token) frequency of each word in L_1 and in L_2. Denote the token frequency of word w by $f(w)$. The probability of word confusion due to vowel confusion is given by

$$\sum_w f(w).\frac{P}{\text{lexicon size}}$$

P denoting the probability of confusing a word with a minimal pair. This can be rewritten as

$$\sum_{v_1,v_2}\left(P(v_1 \rightarrow v_2)\sum_{w_1,w_2} f(w_1).f(w_2)\right)/NF$$

where the word lists L_1 and L_2 correspond to the distinct vowel pair (v_1, v_2) and NF denotes a normalisation factor depending on the size of the lexicon. The above expression is symmetric in v_1 and v_2, since the 'donor' word w_1 and the 'receiver' word w_2 play an equal role. The psycho-linguistic interpretation of this equal role is that the confusion between a certain given word containing v_1 and a minimal pair containing v_2 depends on the token frequency of w_2. It is known that, broadly speaking, the 'accessability' of words increases with its token frequency; in the above expression it is assumed that this relation is linear.

The consequence is that the former expressions for D are exchanged by the new expression

$$D = \sum_{v_i,v_j} A_{ij}P(v_i \rightarrow v_j) \qquad (1)$$

where A_{ij} are constants that are entirely determined by the structure of the lexicon:

$$A_{ij} = \sum_{w_1-in-L_i, w_2-in-L_j} f(w_1).f(w_2)/NF$$

Writing $A_{ij}P(vi \rightarrow vj) = e_{ij}$, $D = \sum e_{ij}$ can be approximated by $1-(1-e_{12})(1-e_{13})...(1-e_{(N-1),N})$ in other words $D = \prod_{v_i,v_j}(1-e_{ij})$ is to be maximized. This latter expression is approximated by

$$\prod_{v_i,v_j}((1-P(v_i \rightarrow v_j))^{A_{ij}})$$

which reveals a lexically-determined weighing of the expression

$$\prod_{v_i,v_j}(1-P(v_i \rightarrow v_j))$$

which returns the probability of v_i not being confused by any other vowel from $v_1,\ldots,v_N$, given the confusion probabilities $P(v_i \rightarrow v_j)$ and uniform distribution of the vowels. The exponents A_{ij} that are determined by the lexicon modify the unbiased case into the lexically-balanced case.

1 Inter-vowel confusion

The second aspect of the model is the relation between inter-vowel confusion and inter-vowel acoustic distance.This aspect is a common feature of each vowel dispersion model. Many models have been proposed (Lindblom, 1972; psychological categorization models, c.f. Smits & ten Bosch, 1994, statistical models). Here we will use $P(v_1 \rightarrow v_2) = \exp(-C.d_{12})$. By substitution in (1) this implies that the following expression is to be minimized: $D = \sum_{v_i,v_j} A_{ij}\exp(-C.d_{ij})$, in which C denotes a constant that is related to the overall scaling of the acoustic space.

2 The definition of acoustic distance

The distance d_{ij} between vowels v_i and v_j is here determined by the Euclidean distance between the first two formant frequencies in ERB. The ERB-transformation is performed in order to agree with the frequency selectivity of the human auditory system (Patterson, 1976; Glasberg & Moore, 1990). The formant representation is chosen for two reasons: to allow a match between model predictions and phonologically specified vowel systems, and the findings (e.g. by Kewley-Port & Atal, 1989) that Euclidean distances based on bark-transformed formants may highly correlate with judged dissimilarities between vowels.

3 Experimental set-up and results

On the basis of the previous sections, the experiment was set-up as follows. Lists of all lexical items of the same grammatical category in Dutch have been extracted from the CELEX database (CELEX, 1990). The twelve Dutch monophthongs (denoted a, i, u, e, o, E, O, I, A, y, U, OE, the last two vowels figuring in 'put' and 'peut') in Dutch were selected for comparison. Diphthongs were not taken into account. For each vowel pair (v_1, v_2), two list where constructed with corresponding phonematically minimal word pairs with the same grammatical category. For example, the two vowels /O/ and /E/ yield two lists with /bOt/ (Eng. 'bone') and /bEt/ ('bed') figuring in it. The minimal pair /rOt/ - /rEt/ ('rotten' - 'save') is not included since they differ in grammatical category.

On the basis of expression (1), all coefficients A_{ij} were determined. Next, optimal vowel positions were looked for that minimized expression (1). This was done by Kruskal's algorithm, by searching positions in a two-dimensional space, such that $P(v_i \rightarrow v_j) = exp(-Cd_{ij})$. For the application of Kruskal's algorithm, $C = 1$ was taken. The optimal systems were found by minimization of the 'stress' which could be defined in a linear or monotonic fashion. Vowel systems were determined for eight

combinations of three binary factors (stress: linear versus monotonic; receiver freq.: token versus lexical; lexical lists: nouns + pronomina only versus all categories). The latter factor refers to the construction of the lists L_i, whether these consist of nouns and pronomina only, or of all categories. This exception is based on the following table presenting relative lexical and token frequencies for 10 syntactical categories (indicated in the first column). Among the PREP, there are hardly any minimal pairs. The VERB category is excluded since it only contains infinitives.

CATEG.	rel. lex. fr.	rel. token fr.
A	13.8	9.5
ADV	1.4	8.2
ART	0.0	10.7
C	0.1	6.6
EXP	0.1	0.0
N	72.3	19.1
NUM	0.2	1.0
PREP	0.1	13.1
PRON	0.1	13.3
V	11.6	18.0

In the following table, the results obtained from Kruskal's algorithm are summarized. For each combination, these results were rank correlated (Spearman) with the actual formant data (derived from Koopmans-van Beinum, 1980 and from Van Son & Pols, 1990).

	combi	Spearman
1	mtn	0.75
2	mtf	0.70
3	mln	0.68
4	mlf	0.66
5	ltn	0.63
6	ltf	0.64
7	lln	0.53
8	llf	0.54

Combinations are indicated by a three-letter combination, referring to the combination monotonous - linear, token - lexical, and (noun+pronomina) ('noun') - all categories ('full'). The difference between combination number 6 and 7 is significant, as well as is the difference between 1 and 4, 2 and 5, 3 and 6, and larger differences. The results are optimized across many (> 200) random start configurations.
Among the monotonic options, the 'mtn' option yields the optimal Spearman correlation with actual data (token frequency, nouns + pronomina). The corresponding vowel system is shown in figure 1. The contour lines connect the formant positions corresponding to 'equal articulatory effort' as proposed in ten Bosch (1991). The 12 monophtongs are plotted in the figure in such a way that the resulting configuration resembles the actual situation (Kruskal's data are specified up to an overall factor, up to rotations, and up to line reflections in the formant space). Among the linear options, the 'ltf' combination yields the highest Spearman correlation. In this setting, Kruskal's algorithm attempts to optimally match the inter-vowel distances on the basis of the inter-vowel confusion probabilities, based on token frequencies and all syntactical categories. The corresponding optimal vowel system in the 'ltf'-case is shown in figure 2.

4 Discussion.

The table presented above shows that the match between predicted and actual vowel system is larger in the monotonous case than it is in the linear case. In fact, the condition in the linear case is harder to meet. Given the monotonic and linear option, the results for the token frequency (slightly) outperform the results obtained with the lexical frequency. This is in line with our expectation. The differences between the options (noun+pronomina) ('noun') - all categories ('full') are small and in fact not significant.

Both figure 1 and 2 show that the lexical structure of Dutch explains a part of the structure of the Dutch vowel system. There are, however, a few remarkable errors. In the monotonic option (figure 1), the position of the short /I/ and /A/ are remarkable. Globally, the triangle-like structure is preserved, but especially the short vowels are not located in coherence with their known acoustic specification. The distance between /A/ and /O/ is larger than expected. This is related to the fact that the number of minimally opposing words for these vowels is large (ten Bosch, 1991). Also in figure 2 (referring to the linear option), the /i/, /a/ and /u/ do not span the vowel triangle any more. The short /A/ lies further from the center than /a/ does. Also here, the distance between /A/ and /O/ is larger than expected.
In general, the localisation of the vowels /U/ from Dutch 'put') and /OE/ (from 'peut') is not precise. Nevertheless, the triangle-like structure of the vowel system, at least for the monophthongs, is clearly visible.
Apart from the question how to integrate diphthongs (that are excuded entirely here), there is another issue to be addressed here, viz. the distinction between long and short. In fact, we studied the 12 monophthongs without any reference to length differences.

The integration of the length opposition into an acoustic contrast measure based on spectral *and* durational contrasts is troublesome (see e.g. ten Bosch, 1991). How duration is to be included remains unclear.

ACKNOWLEGDMENT
This research is sponsored by the University of Amsterdam and by the Dutch foundation NWO.

REFERENCES

Bosch, L.F.M. ten (1991). *On the structure of vowel systems. Aspects of an extended vowel model using effort and contrast.* Ph.D. thesis, University of Amsterdam.

Bosch, L.F.M. ten, and Pols, L.C.W. (1989). 'On the necessity of quantal assumptions'. Journal of Phonetics, vol. 17, pp. 63–70.

CELEX (1990). *A program for retrieval of lexical information (for Dutch, English, German).* Centre for lexical information, University of Nijmegen, The Netherlands.

Crothers, J. (1978). 'Typology and universals of vowel systems'. In: Universals of human language. Vol. 2: Phonology (J.H. Greenberg, ed.). Stanford, Cal., Stanford Univ. Press. pp. 93–152.

Glasberg, B.R., and Moore, B.C.J. (1990). 'Derivation of auditory filter shapes from notched-noise data'. Hearing Research 47, 103–138.

Hayes, W.L. (1981). *Statistics.* CBS College Publishing.

Kewley-Port, D. and Atal, B. (1989). 'Perceptual differences between vowels located in a limited phonetic space.' J. Acoust. Soc. Am. 85, pp. 1726–1740.

Maddieson, I. (1984). Patterns of sound. (Cambridge studies in speech sciences and communication). Cambridge Univ. Press.

Liljencrants, J. and Lindblom, B. (1972). 'Numerical simulation of vowel quality systems: the role of perceptual contrast'. Language 48, pp. 839–862.

Lindblom, B. (1986). Phonetic universals in vowel systems. In; Experimental Phonology (J. Ohala and J. Jager, eds.). Academic Press, Orlando, Florida. pp. 13–44.

Patterson, R.D. (1976). 'Auditory filter shapes derived with noise stimuli'. J. Acoust. Soc. Amer., vol. 59, pp. 640–654.

Ruhlen, M. (1976). *A guide to the languages of the world.* Language Universals Project, Stanford Univ. Press.

Smits, R. and Ten Bosch, L. (1995). 'The multi-layer perceptron as a model of human categorization behavior I. Theory.' (subm. to *J. Math. Psychology*)

SPEECH MAPS (1994). Mapping of Action and Perception in Speech. (C. Abry and P. Badin, eds.). ESPRIT project nr. 6975.

Van Son, and Pols, L.C.W. (1990). 'Formant frequencies of Dutch vowels in a text, read at normal and fast rate'. J. Acoust. Soc. Am. pp. 1683–1693.

Vallée, N. (1990). *Typology des systemes vocales.* Report, Institut de la Communication Parlée, Grenoble (Fr.).

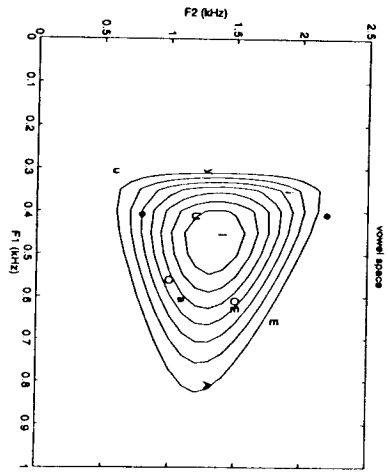


Fig 1

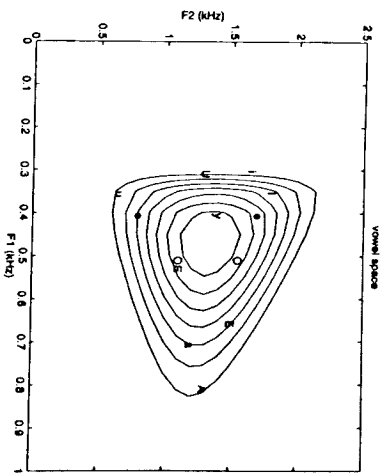


Fig 2

VOWEL PROTOTYPES FOR UPSID'S 33 PHONEMES

N. Vallée, L.J. Boë and Y. Payan
ICP URA CNRS n° 368 INPG/ENSERG – Université Stendhal
BP 25 38 040 Grenoble Cedex 9 France

ABSTRACT
UPSID is a phonological database made up of 33 primary vowel symbols. In this paper we propose 33 vowel prototypes at articulatory and acoustic levels. We can generate them with an anthropomorphic articulatory model. A wide-ranging bibliographic study has enabled us to (i) establish a classification of values for articulatory input parameters, geometrical values of the midsagittal section and crucial values of the area functions; and (ii) specify F1-F2-F3 formant values. These prototypes have been used for a prediction model of the sound systems of world languages.

1. INTRODUCTION

The research of vowel prototypes has been done in the frame of a substance-oriented phonology. Tradionally, the functional efficiency hypothesis, which is looked for at the substance level, is widely used to explain the contents of phonological systems. In this hypothesis, distinctiveness plays a fundamental part: each component is defined in relation to the other components of the system. In advocating a *"substance-based"* analysis, Liljencrants & Lindblom [1] have proposed to reverse the trend by stressing the importance of the "lower levels" in the emergence of systems. The selection of sound units would rest on articulatory and perceptive physiological constraints, which would allow us to explain and predict system structures. The theory of maximal contrast and therefore the design of predictive models both rely on the principle of sound discrimination. To improve modelization, Schwartz & al. [2] have proposed the DFT model *(Dispersion- Focalization Theory)*: They have added auditory pregnancy criteria to distinctiveness [3] (see Schwartz, Boë, Abry & Vallée in these Proceedings). To simulate typical vowel configurations in an acoustic space with a model, we need prototypes, specified both by articulatory and acoustic characteristics [4]. Consequently, we will switch from *form* to *substance* by establishing a relationship between some linguistic units of a representative sample of the phonological inventory of world languages (UPSID [5]), and their physical shape (articulatory and acoustic parameters). In a typological study [4], we have listed 33 vowel qualities which permit to describe the set of symbolic constituant elements of UPSID's 317 systems. From a normalization of the acoustic vowel space [6], we have surveyed and estimated the corresponding values for articulatory input parameters and vocal tract geometry (e.g. location and dimension of the constriction, upper point of the tongue body, furthest back point of the tongue root in the pharynx). For this task, we have used macro-variations [7], – interface functions between articulatory input and acoustic output –, which allow to provide for (i) the acoustic consequences of gestures, (relationship between formants and articulators); (ii) the influence of crucial geometrical parameters [8] (location and dimension of the constriction, and lips area) on the acoustic output. The task has been executed with SMIP, software developed at the ICP within the framework of a European project (ESPRIT/BR N°6975) whose central core is made up of Maeda's articulatory model [9].

2. PROBLEMS TO SOLVE

The traditional description [10], which provides a position for each vowel in terms of height and advancing tongue arching in the buccal cavity, is inadequate. The highest point of the tongue is an operational descriptive parameter whereas the location of the constriction can be directly linked to the acoustic output. Recently, Boë & al. [11] have attempted to unify the traditional description *"lips, tongue arching"* and the acoustic oriented description *"troat-tongue-lips"*.

2.1. INVERSION:

It consists in deriving the vocal tract shape from the acoustic output. Several static articulatory configurations of the vocal tract constitute what is called "a fiber" of the articulatory space, i.e. they make up a set of configurations which supply the same acoustic output [12] [13]. We then need to select a single configuration of the vocal tract and get rid of the rest of the fiber by imposing articulatory and acoustic constraints on these prototypes with the help of experimental and theoretical publications available.

2.2. ARTICULATORY-ACOUTIC RELATIONSHIP:

Secondly, we have to deal with the non-linear and discontinuous relationship between articulation and acoustics. Thanks to the study of macro-variations we are able to foresee the relationship between formants and articulators (Boë, Badin & Perrier, in these Proceedings).

2.3. VARIABILITY:

Different strategies of tongue and jaw allow to produce acoustically identical vowels. Experiments such as "bite-block" [14] show that the vocal tract is capable of using articulatory compensations to produce the same vowel under different conditions.
However, research on invariance [8] [13] [15] has shown an important regularity of the location of the constriction in vowel articulation, whatever the language.

These 3 fundamental issues have led us to collect results of acoustic surveys as well as data on articulatory descriptions [4].

3. METHOD

3.1. VOWEL SPACE:

There are two fundamental constraints: (i) any prototype must be included into maximal vowel space [6]; (ii) the configuration proposed in that space must not fall into the "gap" observed in natural language systems around 300 Hz for F1 and 1,000 Hz for F2, which corresponds to formant area linked to the nasal-pharyngeal tract [16][17].

3.2. PROTOTYPES FOR FRENCH:

All prototypes have been elaborated by calculating dispersion ellipsoids, at the acoustic level, of the 10 oral vowels of French [i e ɛ a ɔ o u y ø œ] for which numerous data were available [4]. Thanks to 60,000 sagittal views of the vocal tract, generated by Maeda's articulatory model [9], we have looked for the ones which fitted our dispersion ellipsoids. With sagittal views, the model supplies us with the values of 7 control parameters determining the position of articulators: lips (retracted and protruded), tongue (dorsum, body and tip), jaw and larynx. Thanks to a sagittal section, we can calculate with SMIP: (i) the area functions whose crucial zones are: Xc the position of the narrowing, its aperture Ac, and Al the lips area; (ii) the transfer function of the vocal tract and formants.
In the same way as Majid & al. [7], and thanks to SMIP, we can infer articulatory data from acoustic targets.

3.2. OTHER PROTOTYPES:

Acoustic targets of the remaining 23 vowels have been positioned in vowel space thanks to surveys of formant data from work done on modelling (synthetic vowels), and various acoustic studies on over 30 languages [4]. A database has thus been constituted. It contains vowel systems of various sizes.
For choosing the value of articulatory parameters, we have also worked with macro-variations of French oral vowels and a wide bibliographic survey. Comparing the various data has enabled us to find a coherence between articulatory control parameters, crucial values of the area function and position in the space formant, even though the variability of sources has sometimes forced us to make compromises in adjusting parameters (Figures 1 & 2).
The 33 acoustic prototypes retained can be synthetized, allowing an auditory control.

4. PUTTING PROTOTYPES TO GOOD USE

More than a stage between form and substance to evaluate predictions, prototypes are the raw material for a whole field of research:

- First, vowel prototypes remind us of the first definitions of "standard vowel quality" of phoneticians [10] or the Jones's cardinal vowels [18], whose primary objectives were to be a reference for the IPA user.

- The hierarchical classification of articulators for all prototypes allow to address again the issue of traditional

articulatory description of vowel and its relationship with acoustic production [11].

• Prototypes are used as a preliminary phase in any attempt at predicting vowel systems. It is now common knowledge that psycho-acoustic parameters are not sufficient for all types of prediction and we must look into articulatory production process for criteria that could improve simulations. Results of this type of research look promising in order to associate articulatory dimension to acoustic and perceptive criteria of distinctiveness – e.g. a description of the articulatory distance (Berrah & al., in these proceedings).

5. REFERENCES

[1] LILJENCRANTS, J. & LINDBLOM, B. (1972). "Numerical Simulation of Vowel Quality Systems: the Role of Perceptual Contrast". *Language 48*, 839-862.

[2] SCHWARTZ, J.L., BOË, L.J., PERRIER, P., GUÉRIN, B. & ESCUDIER, P. (1989). "Perceptual Contrast and Stability in Vowel Systems: A 3-D Simulation Study". *Eurospeech 89*, Paris, Vol. 1/2, 63-66.

[3] BOË, L.J., SCHWARTZ, J.L. & VALLÉE, N. (1994). "The prediction of Vowel Systems: Perceptual Contrast and Stability". *Fundamentals of Speech Synthesis and Speech Regognition*, Ed. by Keller E., Wiley & Sons Ltd, London, England.

[4] VALLÉE, N. (1994), *"Systèmes vocaliques : de la typologie aux prédictions"*. Thèse de Doctorat en Sciences du Langage, Université Stendhal, Grenoble.

[5] MADDIESON, I. (1986). *"Patterns of Sounds"*. 2nd edition, Cambridge University Press, Cambridge (1st edition: 1984).

[6] BOË, L.J., PERRIER, P., GUÉRIN, B. & SCHWARTZ, J.L. (1989). "Maximal Vowel Space". *Eurospeech 89*, Paris, Vol. 2/2, 281-284.

[7] MAJID, R., BOË, L.J. & PERRIER, P. (1986). "Fonctions de sensibilité, modèle articulatoire et voyelles du français". *15e Journées d'Étude du GALF-G.C.P.*, Aix-en-Provence, 59-63.

[8] BOË, L.J., PERRIER, P. & BAILLY, G. (1992). "The Geometric Vocal Tract Variables Controlled for Vowel Production: Proposals for Constraining Acoustic - to - Articulatory Inversion". *Journal of Phonetics 20*, 27-38.

[9] MAEDA, S. (1989). "Compensatory Articulation During Speech: Evidence from the Analysis and Synthesis of Vocal-Tract Shapes using an Articulatory Model". *Speech Production and Speech Modelling*, Ed. by Hardcastle W.J. & Marchal A., Academic Publishers, Kluwer, Netherlands, 131-149.

[10] BELL, A. (1867). *"Visible Speech"*. Ed. by Simpkin & Marshall, London.

[11] BOË, L.J., GABIOUD, B., PERRIER, P., SCHWARTZ, J.L. & VALLÉE, N. (1994). "Vers une unification des espaces vocaliques". *Levels in Speech Communication: Relations and Interactions*, Ed. by Beekmans R., Jospa P., Schoegen J., & Serniclaes W., Elsevier Science Publishers B.V., Amsterdam, Hollande.

[12] BOË, L.J., & PERRIER, P. (1988). "C.F. Hellwag 200 ans après ou les éléments d'une fibre conductrice". *17e Journées d'Étude sur la Parole*, S.F.A., G.C.P., 200-205.

[13] STEVENS, K.N. & HOUSE, A.S. (1955). "Development of a Quantitative Description of Vowel Articulation". *J. Acous. Soc. Am. 27*, Vol. 5, 484-493.

[14] GAY, T., LINDBLOM, B. & LUBKER, J. (1981). "Production of Bite-Block Vowels Acoustic Equivalence by Selective Compensation". *J. Acous. Soc. Am. 69*, 802-810.

[15] WOOD, S.A.J. (1982). "X-Ray and Model Studies of Vowel Articulation". *Working Papers 23*, Lund University, Department of linguistic, Lund.

[16] MAEDA, S. (1984). "Une paire de pics comme corrélat acoustique de la nasalisation des voyelles". *13e Journées d'Étude du GALF-G.C.P.*, Bruxelles, 223-224.

[17] FENG, G. (1986). *"Modélisation acoustique et traitement du signal de parole : le cas des voyelles nasales"*. Thèse de Docteur Ingénieur, INP Grenoble.

[18] JONES, D. (1918). *"An Outline of English Phonetics"*. 1st edition (9th edition: 1960), Heffer W. & Sons L.T.D., Cambridge.

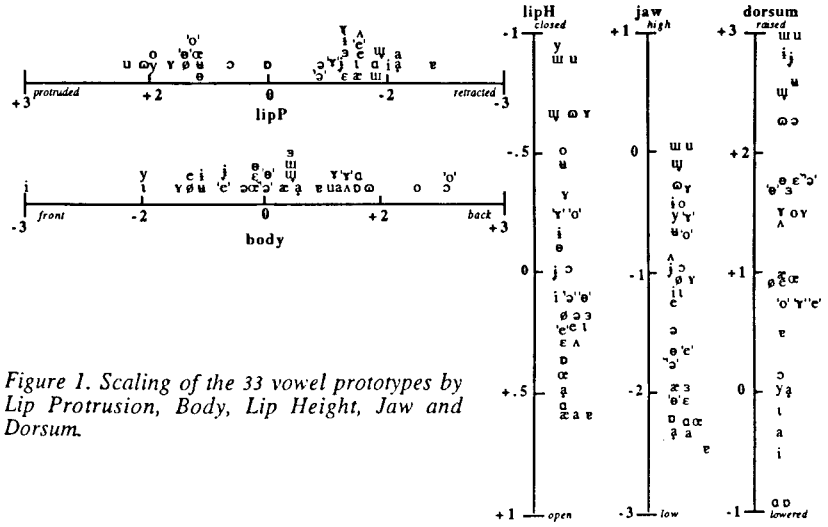


Figure 1. Scaling of the 33 vowel prototypes by Lip Protrusion, Body, Lip Height, Jaw and Dorsum.

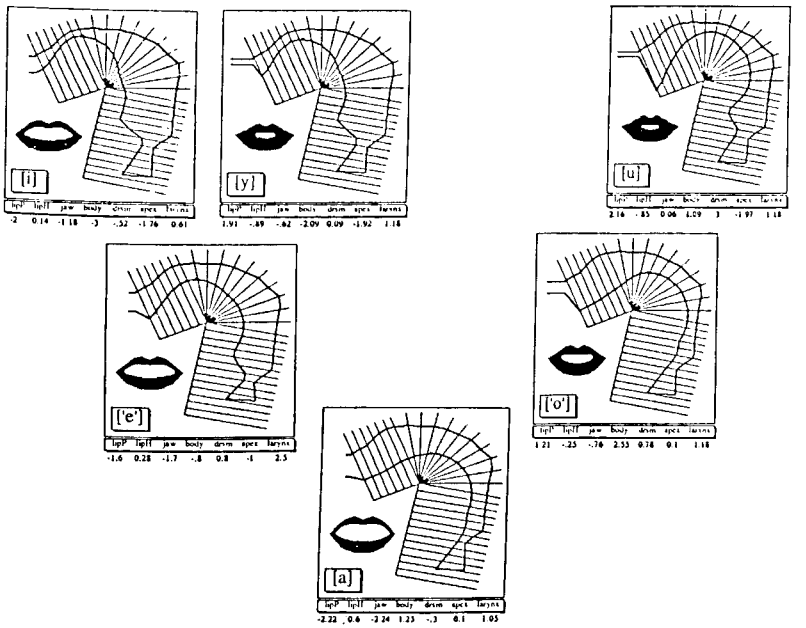


Figure 2. Sagittal views of vocal tract for the five most frequent vowels of world languages /i 'e' a 'o' u/. And the French vowel /y/.

OPTIMIZING ARTICULATION: THE CASE OF ARABIC PHARYNGEALS

Raquel Willerman
University of Washington, Seattle, WA

ABSTRACT

Jaw position in Arabic pharyngeals is investigated to determine whether the open jaw observed by Ghazeli (1977) [1] was due to coarticulation with the following low vowel or to an intrinsic open jaw target for pharyngeals. Both factors were found to contribute to the anomalous open jaw position (jaw position is typically closed for consonants). The data are discussed in terms of biomechanical principles which suggest that pharyngeal consonants are articulatorily complex.

INTRODUCTION

Ghazeli's (1977) [1] x-ray tracing of the vocal tract during production of the Arabic pharyngeal fricative in the word /ʕæli/ revealed an open jaw. This seems anomalous given that jaws are typically open for vowels and closed for consonants. However, it is unclear whether the open jaw seen in Ghazeli's tracing was due to coarticulation with the following low vowel or to an intrinsically open jaw target for pharyngeals. The following study investigates jaw position in Arabic pharyngeals.

SPEAKERS

Five male native Arabic speakers participated in the study: two Egyptians, two Moroccans, and one Lebanese.

SPEECH MATERIALS

Two pharyngeal consonants and two coronal consonants: /ʕ, ɦ, s, l/, were combined with the vowels, /a:, i:, u:/ into the twelve possible CV pairs. Each CV sequence is the beginning of a real Arabic word. The phonetic transcriptions and English glosses of these words are provided in Table 1.

Arabic	English gloss
/ʕa:da/	"he returned"
/ɦa:ða/	"he goaded"
/sa:d/	"predominated"
/la:dan/	"laudanum"
/ʕi:d/	"holiday"
/ɦi:n/	"time"
/si:na:/	"Sinai"
/li:n/	"softness"
/ʕu:d/	"lute"
/ɦu:t/	"whale"
/su:d/	"blacks"
/lu:ða/	"weakness"

Table 1. Arabic speech materials and their English glosses.

The twelve words in Table 1 constitute one trial, the words in each trial appearing in random order. There were three trials of normal speech and three trials of loud speech which were at least fifteen dB louder than the normal speech trials. Speakers monitored their amplitude by looking at a sound pressure meter.

The words were presented in Arabic script. Speakers were instructed to clench their teeth together before reading each word. All jaw measurements refer to a displacement from clench position.

EQUIPMENT

Jaw movements were tracked with a head-mounted strain-gauge cantilever system (Barlow, Cole, & Abbs, 1983) [2]. Jaw movement was sampled in two channels at 1kHz each, and an audio channel was sampled at 10kHz. The principal direction of movement in an x-y plane was determined and a rotation was performed on the signals. The single channel representing jaw movement in this principal direction is used in all further analyses.

ANALYSIS

Jaw position was measured at two points in each word: the midpoint of the first consonant and the midpoint of the following vowel. Midpoints were determined acoustically, using only waveform and spectrographic displays. No special attempt was made to get the lowest jaw position for either segment.

RESULTS

Pharyngeal jaw positions were highly dependent on jaw positions of the following vowel. Jaw positions for /l/ were significantly less dependent on the following vowel and jaw positions for /s/ were the least dependent. There appear to be consonant-specific tendencies to coarticulate. In addition, the tendency for pharyngeal consonants to coarticulate with a following vowel suggests that Ghazeli's tracing of an open jaw is at least partially due to coarticulation with the following low vowel /æ/.

However, we can also ask whether pharyngeal jaw positions are even more open than jaw positions of following vowels. To answer this question, vowel jaw positions were subtracted from consonant jaw positions. These jaw(C) - jaw(V) differences are given in Table 2. A positive jaw(C) - jaw(V) difference means that the consonant jaw position is more open (more displaced from clench) than the jaw position of the following vowel. A negative jaw(C) - jaw(V) difference means that the consonant jaw position is higher (less displaced from clench) than the following vowel.

	Normal		Loud	
	mean	(*p*)	mean	(*p*)
/ʕ/	1.51	(<.01)	1.32	(<.01)
/ɦ/	1.31	(<.01)	1.78	(<.01)
/s/	-3.41	(<.05)	-5.34	(<.01)
/l/	-.94	(<.05)	-2.38	(<.01)

Table 2. Mean jaw(C)-jaw(V) values in mm and p values averaged for five speakers and three vowels in normal and loud speech.

Table 2 gives mean jaw(C) - jaw(V) values in mm for five subjects across three vowels. Positive numbers in the pharyngeal rows, /ʕ, ɦ/, indicate that pharyngeal jaw positions were more open than jaw positions of the following vowel. Negative numbers for the /l, s/ rows indicate that jaw positions for these consonants were more closed than the following vowels. The *p* values are the results of one-tailed t-tests which ask whether the mean jaw(C) - jaw(V) values are significantly different from zero. In every case they are. Normal and loud speech pattern similarly, though jaw(C) - jaw(V) differences for /s/ and /l/ show a marked increase in loud speech.

The relationship between pharyngeal jaw position and each vowel is shown by the average jaw(C) - jaw(V) value of both pharyngeal consonants in each vowel context (Table 3). The pharyngeal consonants have significantly lower jaw positions than /i/ and /u/ vowels for both normal and loud speech. Pharyngeal jaw position is also lower than the jaw position of normal speech /a/. However, loud speech jaw positions for /a/ are virtually identical to loud speech pharyngeal jaw positions! In sum, the open jaw position during a pharyngeal consonant seen in Ghazeli's (1977) [1] tracings seems to be due to at least two factors: coarticulation with a following low vowel and a consonant- specific open jaw position which is at least as open as any following vowel.

	Normal		Loud	
	mean	(*p*)	mean	(*p*)
/a/	.70	(<.05)	-.02	(NS)
/i/	1.50	(<.01)	2.17	(<.00)
/u/	2.04	(<.00)	2.5	(<.00)

Table 3. Mean jaw(C)-jaw(V) in mm and p values from five speakers for pharyngeal consonants in each vowel context for normal and loud speech.

Formant frequencies of these pharyngeal consonants are typical: a high F1 which is close to F2 (Klatt & Stevens, 1969) [3]. For every /ʕV/ pair, the pharyngeal consonants have a higher F1 than the vowel that follows, a high F1 being correlated with a low jaw. This is illustrated in Figure 1 which shows averaged formant data from all five subjects for loud speech. The first and second formants of each vowel are graphed in an F1-F2 plane, yielding the familiar vowel triangle. Average formant values for the voiced pharyngeal fricative /ʕ/ in each vowel context are plotted in the same F1-F2 plane. Notice that the context-bound pharyngeals form a triangle around the /a/ vowel, as if /a/

formant values coincided with the ideal, context-free pharyngeal. Thus, both jaw targets and formant frequencies of pharyngeal consonants resemble those of the low vowel /ɑ/.

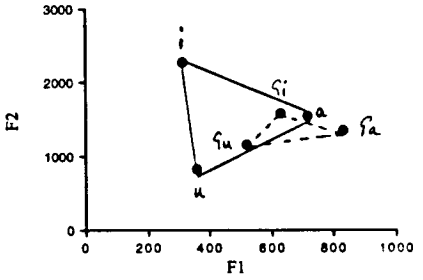


Figure 1. Formant data from loud speech.

DISCUSSION

Extreme displacements of the articulators may be biomechanically costly. Nelson (1983) [4] and Nelson, Perkell, and Westbury (1984) [5] observed that jaw movements are constrained by an economy of effort that can be quantified by peak velocity. In other words, as jaw movement times got shorter, movement distances also became smaller. This tradeoff between distance and time indicates a resistance to large increases in peak velocity. So why do Arabic speakers displace their jaws to such an extent during production of pharyngeal consonants?

The benefit of opening the jaw during pharyngeal consonants lies in the synergy between tongue and jaw movements. In addition to downward displacement in the vertical plane, jaw lowering is accompanied by rotation around the temporo-mandibular joint. This means that opening the jaw will displace the mandible in the dorsal direction as well. If the tongue remains in a neutral resting position with respect to the jaw, lowering the jaw will also back the tongue. This seems to be less costly than extreme displacements of the tongue alone. Arabic speakers may lower their jaws in order to reduce the extent of tongue displacement required for pharyngeals.

If pharyngeal jaw and tongue targets are similar to those of /ɑ/, shouldn't pharyngeals be as biomechanically costly as the vowel /ɑ/? The answer is no. Despite nearly identical articulatory topography, pharyngeal consonants and /ɑ/ are produced under different timing constraints. For a given speaking rate, consonants are generally shorter than vowels (Crystal & House, 1982) [6]. Although jaw targets for pharyngeals and /ɑ/ are nearly identical, the jaw has less time to reach its consonant target than its vowel target. More severe timing constraints on consonants means an increase in peak velocity for pharyngeals vis à vis /ɑ/.

CONCLUSION

This leaves Arabic speakers two choices for producing a pharyngeal consonant. One articulatory strategy is to exploit the synergy between tongue and jaw; that is, the speaker lowers the jaw during the consonant to bring the tongue closer to its pharyngeal target. In this case, the jaw must move fast in order to reach the pharyngeal jaw position in the relatively short amount of time that is allotted to consonants. This strategy is biomechanically costly in terms of having to apply more force to the jaw in order to achieve a faster rate.

On the other hand, the speaker can forgo tongue-jaw synergy. In this case, the speaker must "superpharyngealize" the tongue to reach the back and low constriction without help from the jaw. With a relatively high jaw, the tongue must move a greater distance from its resting position with respect to the jaw in order to reach the pharyngeal target. This is biomechanically more costly in terms of extreme displacement of the tongue from neutral.

Both scenarios were used by the Arabic speakers in this study. Articulatory strategies are inferred from the individual jaw(C)-jaw(V) values. Four out of five speakers had pharyngeal jaw positions which were consistently lower than the following vowel. This indicates that these speakers exploit the synergy between tongue and jaw, at the supposed biomechanic cost of increased rate of jaw movement. One speaker had pharyngeal jaw positions which were consistently higher than the following vowel. This indicates that the speaker uses a "superpharyngealized" shape of the tongue, at the biomechanic cost of increased displacement from neutral.

Yet, acoustic measurements from the pharyngeal consonant portions of the spectrograms showed that all five subjects attained the high F1 and closely hovering F2 that is correlated with pharyngeal constriction. Thus, phonetic reduction can be ruled out as an explanation for the singular speaker's not-so-low pharyngeal jaw position; and, superpharyngealization seems more likely.

Is it significant that four out of five speakers increased extent and possibly rate of jaw displacement in order to decrease extent of tongue displacement; whereas, only one speaker minimized jaw displacement at the cost of increased tongue displacement? For now, the fact that extreme displacements of the tongue are more often avoided through synergistic movements of the jaw seems to suggest that jaw displacement is relatively cheaper than tongue displacement. The idea that jaw movement may be virtually free of charge biomechanically seems more plausible upon considering the primary function of the jaw: to grind food between the teeth. The mandible has powerful masticatory muscles capable of delivering great force, much more force than is required for speech movements. Thus, the mandible is in a sense overpowered for speech purposes. This leads to the following asymmetry: tongue movements are biomechanically more costly than jaw movements. Further support for asymmetrical treatment of jaw and tongue comes from comparing the range of displacement used for speech to the total anatomically possible range of displacement. Jaw displacements used in speech range between 5-15mm from clench; whereas, displacements are much greater for loud speech and yawning. In contrast, the range of tongue displacements used in speech is near the total range of possible displacement.

By either articulatory strategy, tongue-jaw synergy or "superpharyngealization" of the tongue, pharyngeals are biomechanically costly and should be avoided in the world's languages. Indeed, Maddieson's (1984) [7] survey of 317 languages reveals that pharyngeals are relatively rare in phonetic inventories; they occur in only three to four percent of the languages which have fricatives. Yet, the results of this investigation show that speakers have ways of minimizing the effort required in pharyngeal production. Furthermore, this study indicates that different articulators may be weighted differently with respect to constraints on peak velocity and economy of effort. It is hoped that non-data-driven measures such as peak velocity can be used to construct a metric of articulatory complexity in order to assess the complexity of phonetic sound systems.

REFERENCES

[1] Ghazeli, S. (1970), *Back consonants and backing coarticulation in Arabic*, Unpublished doctoral dissertation, University of Texas, Austin.

[2] Barlow, S., Cole, K., & Abbs, J. (1983), A new head-mounted lip-jaw movement transduction system for the study of motor speech disorders, *Journal of Speech and Hearing Research*, vol. 26, pp. 283-288.

[3] Klatt, D., & Stevens, K. (1969), Pharyngeal consonants, *MIT Quarterly Progress Report*, vol. 93, pp. 207-216.

[4] Nelson, W. (1983), Physical principles for economies of skilled movements, *Biological Cybernetics*, vol. 46, pp. 135-147.

[5] Nelson, W., Perkell, J., & Westbury, J. (1984), Mandible movements during increasingly rapid articulations of single syllables: Preliminary observations, *Journal of the Acoustical Society of America*, vol. 75, pp. 945-951.

[6] Crystal, T., & House, A. (1982), Segmental durations in connected speech signals: Preliminary results. *Journal of the Acoustical Society of America*, vol. 72, pp. 705-716.

[7] Maddieson, I. (1984), *Patterns of sounds*, Cambridge: Cambridge University.

TEACHING PHONETICS

Peter Ladefoged
Phonetics Lab, Linguistics Department, UCLA, Los Angeles, CA 90095-1543

ABSTRACT

A brief overview of the origins of contemporary phonetic teaching is followed by an outline of the general format of the symposium.

GENERAL PHONETICS

A large proportion of the people who attend phonetics congresses are concerned with teaching phonetics. Accordingly, in this session, we will both present and seek ideas on what we should teach, and how we should teach it. There is no way in which we will be able to discuss all aspects of phonetics. Ours is an active field in which new discoveries are always being made, so that no one can ever be completely up to date. No doubt the presentations at this congress will provide material for many hours of future instruction, but most of this will be for advanced students. What we will attempt to do here is to consider the needs of students of general phonetics. What do they need to know, and how are they going to learn it?

HISTORICAL BACKGROUND

We should begin by noting that organized teaching of *general* phonetics is fairly new. There were several great phoneticians at the end of the nineteenth century, but very little teaching of the subject. Alexander Melville Bell was a university lecturer in elocution, a subject which included only part of phonetics as we now know it. Henry Sweet eventually became a university Reader in Phonetics, but, according to Daniel Jones, Sweet did not teach classes in general phonetics (Daniel Jones has noted (p.c.) that he was not Sweet's pupil but met him on a number of occasions). Paul Passy held a chair of General and Comparative Phonetics, but his main academic interests were in modern language teaching. Otto Jespersen was a professor of English Language and Literature, and devoted most of his time to non-phonetic issues. Perhaps the closest to general phoneticians were Wilhelm Viëtor, who was a Professor of Linguistics, and l'abbé Rousselot, who established the first university phonetics laboratory as well as working in dialect geography. But neither of them presented an overview of the field as a whole in regular organized classes of the kind that are now available.

Most of the European phoneticians of the first half of the twentieth century were concerned with teaching pronunciation, usually in connection with teaching a foreign language, although sometimes as teachers of the deaf or others with pathological problems. As David Abercrombie [1] has reported, even Daniel Jones, the most well known Professor of Phonetics throughout the first half of the present century, did not teach general phonetics. His students learnt about the subject only through studying the phonetics of English, the phonetics of French, and so on. Substantial teaching of general phonetics did not begin until the 1940's, with scholars such as Malmberg in Sweden, Fischer-Jørgensen in Denmark and Abercrombie in Scotland. There were older Departments of Phonetics, such as that at London, but they were concerned with teaching phonetic skills rather than with presenting a general overview of the field.

In the early part of the century, American scholars such as Sapir were also emphasizing practical phonetic skills in their training of anthropology students (Emeneau, p.c.). But the situation was slightly different from that in Europe in that the main impetus for phonetic training came from the study of American Indian languages rather than from studying languages such as French and German. Towards the end of the first half of the century Pike, a professor of Anthropology, formulated the first contemporary synthesis of phonetic issues [2] while being concerned with training missionaries and bible translators.

Thus in neither Europe nor the U.S. was there a subject, general phonetics, that was taught until the 1940's. Abercrombie's course at Edinburgh University in 1949 is one of the earliest year-long courses in general phonetics. Since then the subject has developed different emphases, but the basic topics that all students of phonetics should know remain much the same as those taught in Abercrombie's course: articulatory and acoustic phonetics, speech perception, experimental phonetic techniques, phonetic performance skills and the classification and symbolization of speech sounds.

Times are changing rapidly, and the impetus for much phonetic work is now coming from those needing to know more about speech communication, rather than from those studying foreign languages. Nowadays it is appropriate to add a knowledge of computer speech processing to those listed above. However, it is still true that the other areas—articulatory and acoustic phonetics, speech perception, experimental techniques, the classification and symbolization of speech sounds, and performance skills—remain (together with computer speech processing) the basis of the subject today. Anyone who hopes to solve current problems in speech communication needs a good grounding in all these areas.

A SYMPOSIUM ON TEACHING PHONETICS

It is possible to study phonetics in different ways in many institutions. This symposium will begin with an account by Gerrit Bloothooft of some of the possible ways of studying phonetics in Europe, with particular regard to the relation between phonetics and speech communication. We will then consider some fundamental aspects of the subject: teaching acoustic phonetics, which Jacqueline Vaissiere will discuss, problems of intonation, which Mary Beckman will present, and the development of phonetic skills, which I will consider. As we have seen, there are many other topics that are of importance to phoneticians, including fundamental notions of articulatory phonetics, the anatomy and physiology of speech, speech perception, speech synthesis and recognition, and phonetic transcription. These and other similar topics are all important matters, and we do not wish to seem neglectful of them. We hope that some of them will be dealt with in the general discussion.

Part of the delight of studying phonetics is that it has connections with so many different disciplines. In this Congress we have a very wide range of people, nearly all of whom think of themselves as phoneticians of some sort. What holds us all together? What is the core of the subject? How should we teach it? How should we help newcomers have as much fun as we all have?

REFERENCES

[1] Abercrombie, D. (1991). *Fifty Years in Phonetics,* Edinburgh: Edinburgh University Press.
[2] Pike, K. (1944). *Phonetics,* Ann Arbor, MI, University of Michigan Press.

TRAINING OF PHONETICS AND SPEECH COMMUNICATION IN EUROPE

Gerrit Bloothooft
Research Institute for Language and Speech
Utrecht University, The Netherlands

ABSTRACT

The benefits of international cooperation in training are presented for the European Erasmus programme 'Phonetics and Speech Communication' in which 29 institutes from 13 countries participate. The exchange of students and lecturers, summer schools, and information dissemination are discussed. Further developments will be directed towards distribution of teaching material, discussions on course contents and curricula, and broadening of the network.

INTRODUCTION

The day before the start of the XII International Congress of Phonetic Sciences in Aix-en-Provence in 1991, 22 people met to discuss opportunities for cooperation in training and mobility of students in Europe within the framework of the European Erasmus programme. This programme provides financial support for exchange of students and lecturers, for development of new curricula and study materials, and for intensive programmes such as summer schools. There proved to be a broad support for initiatives in this area (also exemplified by the warm support of ESCA, the European Speech Communication Association), and in the next four years the initial network of 11 sites grew to an almost full-coverage of 29 sites in 13 countries of the European Union [Sweden, Finland, Norway, Danmark, Germany, the Netherlands, the United Kingdom, Belgium, France, Spain, Portugal, Italy, Greece]. Starting with the exchange of students, the network has been successful in getting support for an annual European summer school (organized in cooperation with ELSNET, the European Network of Excellence in Language and Speech), and a programme for the exchange of lecturers. Furthermore, the network has encouraged thinking about future developments in education of students by means of cooperation in study programs. The annual information book of the network has been praised as an excellent example of distribution of information. And we did even better. With the help of a great many colleagues from inside and outside the network, we were able to publish the handbook 'European Studies in Phonetics and Speech Communication' at this XIII International Congress of Phonetic Sciences in Stockholm. Although we are a relatively small research community, or just because we are small, active and involved, we have proven to be able to stand in the frontiers of developments in training.

The organizers of this congress decided to schedule a plenary session on training. To my knowledge, this is the first time that this happens at a Phonetics Congress. There could be no better sign of recognition of the importance of training for the future of Phonetics. Besides, a lot of things are changing these days that have influenced and will influence the study of Phonetics.

A NETWORK FOR TRAINING

Before I give a more detailed account of the achievements and possibilities of our Erasmus programme 'Phonetics and Speech Communication', I want to give some general thoughts on the benefits of cooperation in training.

Phonetics is a discipline that is small in size but complex in contents. In a single department, Phonetics is mostly taught by a few people that color the curriculum according to their own specialization in research. Sometimes, local links to education in Linguistics, Speech Therapy, Psycholinguistics, Hearing Science, Computer Science, and Speech Technology are used to enrich the contents of a study in Phonetics. Yet, most departments will face practical limits to their possibilities in education. On the other hand, we have a great resource of expert lecturers in all fields of specialization in Phonetics in Europe as a whole. It would be great if there were opportunities to share this resource.

A network can be an excellent instrument to share knowledge and expertise in training. A network can also be a fine way to get to know each other and to learn from each other by discussions. There is a lot of freedom too. Every partner is free to decide whether to use the opportunities or not.

On the basis of these types of observations, we made two important decisions in the early days of our Erasmus network. The first was that we aimed at a single network for the whole of the European Union, and the second was that, in order to maximize opportunities, we did not want to be too restrictive in the definition of Phonetics and Speech Communication. To put it simply, we did not want to make a distinction between the major areas Phonetics, Speech Technology, and Speech Therapy, as far as the latter involved research and no professional treatment of patients.

These choices have worked out excellently. In other areas of study various partial networks exists, each with a limited number of partners. In case of exchange, this reduces the possibilities for students to find the host university of their choice. Our network now comprises 29 departments in 13 different countries, including both curricula in Phonetics and in Speech Technology, and specializations in Speech Therapy.

Students in Phonetics can choose out of several Speech Technology departments to learn more about algorithms, speech processing and speech systems, while reversely, speech technology students can go to Phonetics departments to follow courses in basic and advanced knowledge of Phonetics. And, of course, Phonetics students can also go to other Phonetics departments that specializes in the area of the students' interest. The same holds for Speech Therapy and Speech Technology students.

THE ERASMUS PROGRAMME

Organization

An Erasmus programme may come into existence if a number of departments express the wish to cooperate in training. The signature of the rector of a university on a letter of endorsement is sufficient for participation, and commits the university to support the actions of the network. Since internationalization is high on the agenda of universities nowadays, we did not experience any difficulties in getting this signature for our programme. Most universities have an international office that takes care of the registration of visitors, their accommodation, and the payments of grants to students. To my knowledge, these international offices are very capable and helpful, and they provide the basis for the success of the programme. They are a vital infrastructure that takes away a lot of work from the staff members in departments.

Because we have an Erasmus network with many partners, the coordination is in the hands of a steering committee of three people, Joaquim Llisterri, Valerie Hazan and myself. We are responsible for the continuity of the network and the gradual widening of its activities. Every year there is a general meeting of all local coordinators to discuss the ongoing exchanges, and to decide on new activities and the budget. These general meetings have a special flavor that will be remembered by the participants, are very important for personal contacts between coordinators.

The Erasmus programme has four chapters: (1) the exchange of students, (2) the exchange of lecturers, (3) the organization of intensive programmes, and (4) the development of new curricula. We have been active in the first three chapters.

Exchange of students

The number of student exchanges have to be estimated by a local coordinator more than one year in advance. This causes difficulties because students tend to change their minds rather easily. Once our programme has been approved for continuation, further arrangements between home and host universities can be made with respect to

the study programme, and practical matters like accommodation.

The student will get (very) moderate funding in addition to a basic bursary, but the student does not have to pay the fee for the host university, and the accommodation is arranged (although not for free). So, once a student has arrived, he/she will get a room, some extra money, and has access to all student facilities of the host university. The study programme may take between three months and a full year. The study may involve taught courses and/or thesis work and is open for students up to the PhD degree. Most students are going abroad during a later stage of their studies in which there is a freedom to choose optional courses. This avoids problems with the acknowledgement of study credits for obligatory courses. Although most teaching material is in English and English is spoken by most members of departments, students are encouraged to follow language courses to improve (social) contacts inside and outside the department. Language courses are funded by the Erasmus programme.

We are a small community and figures for exchange students are not impressive, about 25-30 students a year, which equals to about one exchange per institute per year. Still, this is not bad if we consider that the number of students that graduate each year in Phonetics rarely exceeds five students per institute. Most of the exchange students report very positively about their stay abroad. They learned a lot, made useful contacts, and also found the experience of being in a new social environment for a long time both difficult in the beginning but very rewarding in the end. Many wanted to stay longer!

Exchange of teaching staff

Traveling students or traveling lecturers? Both are possibilities to distribute knowledge, but the old, historical practice of traveling scholars does not fit well into present day teaching practice. Now and then, someone will get the opportunity to go abroad for a long period as a visiting scholar, do research and perhaps some teaching, but a regular exchange of teachers for a longer period faces many practical obstacles. There are financial an organizational barriers that make it difficult to visit another institute during more than a few weeks. Consequently, lectures do not fit in the normal teaching schedules, unless they are part of existing courses, or given as intensive courses that do not clash with other courses or exams. Since the third year of our network we have asked and got support for the exchange of lecturers. The financial support is modest, however, and covers travel costs only. Still, between five and ten exchanges of lecturers are supported each year, several of which are bilateral. The general meeting of the programme decides on what exchanges can be supported.

Summer schools

In summer schools and intensive courses, lecturers and students can be brought together around a certain topic during a few weeks in a stimulating atmosphere. Since three years, our Erasmus programme has organized the European Summer School on Language and Speech Communication, in cooperation with ELSNET. The summer schools are topic-oriented with 'Prosody' in London 1993, 'Corpus-Based Methods' in Utrecht 1994, and 'Multi-Linguality' this year in Edinburgh. Between 60 and 90 participants (students, staff, industry employees) attend the summer school. There is a technological orientation and an ELSNET flavor because of the attempt to bring Speech and Language (technology) students together. The high quality of the courses, the many practicals, and the social programme are highly appreciated by the participants. The difficulties in the annual organization are related to the financial uncertainties and the availability of a host department.

Information dissemination

Although spread of information is no separate chapter in an Erasmus programme, it is a prerequisite for an effective network. Students and staff have to know which courses are offered at what site in order to decide on the host university and the study programme for an exchange. we have published an annual information book that presents

(1) general information on the programme, (2) descriptions of the participating departments, and (3) a description of all courses presented within the network. The latter involves the impressive number of more than five hundred courses. The information book has been mentioned in several places as an excellent example for other Erasmus networks.

The success of the information book has been the basis for the project to transform the book into a handbook on studies of Phonetics and Speech Communication in Europe. The aim of the handbook is to raise interest for our field with prospective students, to widen the horizon for advanced students, and to stimulate exchange and further cooperation. We have asked experts from all over the world to give their views on past, present and future of their field of interest. With 25 contributions, this chapter gives a unique, most interesting and stimulating entry to Phonetics and Speech Communication. Because we did not want to limit our book to the Europe of the European Union, we have asked colleagues from all countries between the Atlantic and the Ural for help. This resulted in a board of 25 country-editors who made a general text for each country, including interesting stories on historical developments, explanation of university systems and other general information. They also organized the collection of descriptive texts for the different departments in each country. Finally, we made an attempt to give an overview of the contents of elements of studies in Phonetics and Speech Communication. I will come to that later.

With the publication of the handbook 'European Studies in Phonetics and Speech Communication', we have laid a solid foundation for further thoughts on education. The handbook excellently fits into the philosophy of the next phase of European cooperation in training, the Socrates programme, in which the Erasmus programme will be continued. For the next four years, we may expect consolidation of existing exchange programmes without much extra funding for new initiatives, and certainly not for annual meetings of coordinators. What is new is the possibility to organize think-tank conferences on the future of education. It is my hope that we will get the support to continue along that line too.

DEVELOPMENTS

After having given an overview of what has been done within our Erasmus programme during the last four years, I would like to look into the years to come. Although we have been very active in many areas of training and mobility, there are still several aspects that need our attention. These aspects concern questions on (1) *how* to teach, (2) *what* to teach, and (3) how to extend cooperation.

Teaching Phonetics

It is of great importance that we share experiences in training of our students. In the following contributions of Mary Beckman, Jacqueline Vaissiere, and Peter Ladefoged, you will experience fine examples of teaching styles that may inspire all of us for our own courses. You will also note that they all use example materials and demonstrations that we would like to use ourselves. We should work on ways to establish a wide-spread distribution of sound demonstrations, audio-visuals, and software. This software may vary between simple direct demonstrations and very complex programs for computer-aided instruction. At this congress, you may have seen an example of computer-aided instruction for main chapters in Phonetics, developed with contributions from many experts. All this can be used in addition to the many excellent books that we have to teach the basics of Phonetics. And besides all the materials and computer-based demonstrations, it would also be very stimulating if lecturers were recorded on video during presentation of courses to bring our students into contact with the personalities of the great instructors in our field.

The curriculum in Phonetics

Apart from learning from each other *how* to teach, there is the big question of *what* to teach. With this question we immediately stumble on a definition of Phonetics and on the qualifications we think that graduate students in Phonetics should have. To be clear, there is no one

who can prescribe how departments have to set up the curriculum. This is the responsibility of the departments themselves, the board of examiners and the head of the department. This is the domain of the academic freedom. But still, it is my opinion that it would be wise to approach the contents of a curriculum in Phonetics with an open mind, because there is a lot to gain if we could agree on aspects of education. These advantages concern practical matters like easier exchange of students and easier acknowledgement of study credits, but we can also envisage help to maintain the quality of education in order to keep in pace with the rapid developments in research and applications.

I think it will be very difficult to get a full agreement on a curriculum in Phonetics in all its aspects, but I also believe that this is not necessary. We should respect and cherish the existing different angles towards Phonetics, and have an open eye for local limitations. But exchange programmes for students and staff can widen the possibilities. Above that, electronic networks allow for *virtual* exchange: there have been demonstrations of courses presented over internet, while the first electronic class rooms now come into existence. In these cases, there is no need for the lecturer and the student to be in the same room, they may have remote locations.

Within the Erasmus network, we have not yet discussed the structure of curricula. We have chosen for a stepwise process, starting with a description of the elementary contents of a study, whether this concerns Phonetics, Speech Technology or Speech Therapy. Of course, such a description as a whole by far surpasses the contents of a single curriculum. We hope, however, that the collection of elements can be complete in such a way that for all existing curricula it can be shown what choices have been made for its contents. The results of our first attempt can be seen in the last chapter of our book 'European Studies in Phonetics and Speech Communication'. Even in this first phase, there is still a lot of work to do. The completeness of the set can be criticized, textual explanations of the contents of elements can be elaborated, an element can be enriched with more information on key books and key papers, links between elements can be established better, and so on. Ideally, the elements should also point to related courses throughout Europe. In all, this approach asks for an information system with hypercard-like properties, to be built on World Wide Web, which can be browsed by students to pick the courses that fit their interests and needs.

It may be that by discussing and working on the elements of study, more agreement will emerge on obligatory basics and optional advanced topics than we now anticipate. The step towards a challenging cooperation in terms of a European Degree in Phonetics, which is an option under the Socrates programme, may then prove not too difficult. Such a European Degree would imply agreement on the end-qualifications of students. It should be formalized what set of courses constitute the degree, with the flexibility that (some of the) courses may be presented at various universities in different countries. Exchangeability of courses and/or course contents should be accepted by the responsible boards at all departments involved. It is difficult to say how far away in time this vision of European cooperation is, because there are practical barriers as well. We have made an overview of all degrees in Phonetics and Speech Communication that can be obtained in the various countries in Europe. This shows that, in order to arrive at a European Degree we also have to overcome the differences in university systems and the place of the eduction of Phonetics in these systems. These formal difficulties in comparability of studies are beyond our scope and worked on at governmental levels.

Interactions between curricula

Our Erasmus network presents opportunities to share. For that reason we are open to departments that have the main orientation in Phonetics, in Speech Technology, or in Speech Therapy. These are not three independent main streams in the field. Several links exist between them. I would like to discuss two interactions: between Phonetics and Speech Technology, and between Language and Speech.

Phonetics and Speech Technology

Speech Technology has always been important to Phonetics in that it provides the tools that support phonetic experimentation. But since a few decades, Speech Technology has become a discipline on its own with the development of speech communication systems that have a considerable economic potential. Funding of research and consequently, jobs for students are now largely to be found in Speech Technology, whether we like it or not. In parallel, new curricula and specializations in Speech Technology have been developed. At Technical Universities or Computer Science departments, Speech Technology curricula tend to have little input from Phonetics. It would provide an interesting discussion whether this is optimal.

On the other hand, developments in Speech Technology have had an impact on studies of Phonetics. Most curricula in Phonetics have at least a few courses on speech synthesis, automatic speech recognition, and speech processing. In some of the larger departments of Phonetics, Speech Technology is a graduate specialization, or presented as a separate MSc course. These options are certainly in the interest of the careers of Phonetics students who also have an interest in technology.

In both cases, the opportunities provided by exchange within the Erasmus programme can enrich the study programme for students. For example, several students in the Utrecht Phonetics department specializing in Speech Technology, have had valuable exchange periods to technology-oriented departments in Sheffield, Aalborg, and Stockholm.

Language and Speech

Phonetics is often a part of a Department of Linguistics and students generally have easy access to courses in linguistics. It depends on the character of the individual institute whether these courses are incorporated as obligatory or optional components in the Phonetics curriculum. Far more complicated are the relations to the technological counterparts in linguistics: Computational Linguistics and Natural Language Processing. Driven by technological developments like in spoken dialogue systems and speech-to-speech translation systems, these areas now interact with Speech Technology, and we may think about their place in the education of speech students. Although this type of integration is the major priority of the European Network of Excellence in Language and Speech (ELSNET), it is not immediately obvious in what studies, at what level, and to what extent this integration should be realized in education and formalized in curricula. As could be expected, the first initiatives and experiments towards integration were taken at departments where specializations in both Speech Technology and Computational Linguistics already existed. In some of these cases, it is possible for a student to get an MA or to follow a one-year MSc course in Language and Speech Processing. I mention these developments to illustrate the continuous and important activities in areas that may be considered remote to Phonetics.

Cooperation on a global scale

I have discussed cooperation in training within the reference of the European Union. This by no means implies a limitation of our interests in cooperation to the borders of the European Union, but these processes need time and proper funding. The European Tempus programme offers support for mobility from and to Central and Eastern Europe, much like the Erasmus programme, but we do not yet have a joint programme for Phonetics and Speech Communication. I gladly refer to the support we have had from our colleagues from Central and Eastern Europe for the 'European Studies' book. There is a strong wish to cooperate at all levels, but the financial obstacles are enormous. More initiatives from our side are needed here in the near future.

Cooperation with countries outside Europe, with a special interest of students for the US and Japan, largely exists on an individual basis with support from bilateral cultural treaties or other sources. A regular contact on training issues is wholeheartedly welcomed and easily realized now that

world-wide communication facilities are at the desk of most colleagues.

However, enlarging the scale of cooperation evokes the dangers that the organization becomes too complex, that the active involvement of partners decreases, and that general agreement is difficult to reach. Up to now, the Erasmus programme did well, and I am most grateful to my helpful friends in the steering committee, Valerie Hazan and Joaquim Llisterri, and all other partners who gave an exemplary demonstration of the possibilities of distribution of labour during the preparation of our 'European Studies' book. This gives great expectations for the future. Nevertheless, I believe that training matters with a global dimension should be anchored in the international organizations in Phonetics and Speech Communication. This would best ensure the broad basis needed for continuous developments in training in order to bring our knowledge to the next generation.

TEACHING ACOUSTIC PHONETICS

J. Vaissière
CNRS-URA 1027, Institut de Phonétique,
19 rue de Bernardins, 75005 Paris, France

1. INTRODUCTION

Ladefoged [1] stated in the introduction to this symposium that we should think of a good phonetician as someone who has a good grasp of phonetic principles and understands the issues in speech production, perception and acoustics.

Since the acoustic signal is the central element in the speech communication chain, a rather deep knowledge of speech sounds, how to characterise them acoustically, how such characteristics can be produced in an human vocal tract, how they are perceived, plays a dominant role in a unified understanding of speech. Acoustic phonetic knowledge permits an understanding of the interdependence of issues in speech production, perception and acoustics. A solid understanding of the complexity of human speech relies on appropriate knowledge of the acoustic phonetics field.

My talk is addressed to teachers faced with the problem of teaching acoustic phonetics in very large class rooms. At my institution, undergraduate linguistics students obligatorily attend year-long phonetics courses three hours a week for the first two years and a minimum of 1 1/2 hour during the third year. The degree of motivation for studying phonetics is not uniform among the students. It is important that they must all accept that they have to learn a series of notions that they may perceive at first as too technical.

The talk is also addressed to teachers of phonetics who are seeking ideas for how to transmit in the most efficient possible way the necessary background to students and professionals from various disciplines (linguists, engineers, professors from foreign universities, medical doctors, physiologists, computer scientists, speech pathologists, etc.). Those students are motivated, they have access to phonetic laboratory facilities, and they will soon be engaging in multidisciplinary research involving phonetics. Intensive courses in acoustic phonetics allow them to find phonetic problems for themselves and to solve the problems using a limited mathematical background (see for example, a paper on Arabic fricatives [2] and on the singer's formant [3], presented in this congress). This sort of acoustic knowledge is like mathematics for an engineer. The demand of intensive phonetics courses is expanding globally.

2. WHAT TO TEACH?

All students should be given the necessary background to understand the links between

- **basic acoustic laws** underlying generation and propagation of sounds,
- **articulatory manoeuvres** for producing the speech sounds in an human vocal tract,
- **acoustic characteristics** of the resulting sounds, and
- the interpretation of such acoustic characteristics by the **perceptual mechanism.**

In other words,

1) Students should be given the background to understand the characteristics of the spectrographic representation of speech in relation to articulation.

2) To do so, students must understand how speech sounds are created by talkers.

3) So, students should know about the control of airflow (sound source generation) and of vocal tract shapes (acoustic characteristics of speech sounds)

4) Consequently, they need to understand the relationship between the vocal tract shapes (VT) and the acoustics (Fn) without or with only simple mathematics (computer programs can be helpful for the acoustic calculations).

To summarise, there are five basic topics:

1) - what is a **wave**?

Periodic, *quasi-periodic* and *non-periodic* waves (noise).

A *sinusoidal* wave is a periodic wave having a single frequency.

Periodic waves (vowels, nasals, etc.) are composed of '*harmonics*'. They are the sum of sinusoidal waves, with their frequencies n * Fo, where n = 1, 2, 3, ..., n, and Fo is the fundamental frequency in Hz.

Vowels are quasi-periodic and *fricatives* non-periodic.

2) - What is **resonance**?

A resonance manifests itself as a *damped oscillation* of a wave . It is characterised by its *natural frequency* and *damping*.

The *vocal tract* acts as a set of resonators. The resonance frequencies depend on the shape (area function) of the vocal tract.

3) **Simple acoustic tube**, and their relation to resonances:

Resonance modes like the *quarter wave length* resonances, *half wave length* resonances, and *Helmholtz* resonance.

4) Vocal tract acoustics calculations using **connected simple tubes** representing vowels.

Comparison of acoustic and perceptual calculations and that using a simulation program.

In order to obtain the "simple tubes" representation of the vocal tract, X-ray pictures (or MRIs) in the midsaggital plane are used.

5) About **sources:**

Vocal-fold vibration (measurements using the glottograms, EGG) etc.

Aerodynamics principles underlying noise generation.

3. HOW TO TEACH

In the following text are listed a set of traditional and less traditional "tools" for teaching acoustic phonetics.

a) Tuning fork, pendulum and anecdotes

Most of the available introductory books contain valuable information on how the courses can start by using traditional tools.

Sounds are acoustic vibrations propagated to the ear through air. What is a *vibration*, and what *propagation* means? A **pendulum**, such as the one that I will ask Mary Beckman to set up gently into motion, is a mass attached to a string pulled into oscillation. A simple watch or chronometer may be profitably used to introduce in front of the students the notions of *movement, amplitude* of movement, *periodicity*, influence of the length of the string on periodicity, *cycle, frequency, sinusoidal* oscillations, *damped* oscillations, representation of the displacement on a *graph* (displacement versus time, the magnitude of displacement versus frequency or resonance curve), etc.

Tuning forks allow students to *hear* sounds differing by *pitch* only, or by *loudness*. Two identical tuning forks can be used in a very effective manner for an introduction of the notion of *natural frequency* and *resonance*, as I will ask Peter Ladefoged to demonstrate in front of you. All demonstrations with tuning forks can be even made more fruitful by systematically providing the students with *waveforms*, *spectral slices* and *spectrograms* of the sounds.

A useful way to attract the attention of the class is to start a section with one of the traditional **anecdotes** (the resonant Tocama Narrows Bridge [4], the glass broken by the voice of a singer, etc.) and to let them find by themselves various examples of resonances in daily life: child swing, grand-father's clock, undesirable car vibrations at a certain speed, etc..

Peter Ladefoged's **book** 'Elements of acoustic phonetics' first published in 1962 and this year to be re-issued in a revised version with additional material on digital speech processing [5], offers a coherent 'parcours pédagogique" for teaching the basic notions at the most elementary stage.

b) Spectrogram decoding as a basic complementary course

The best way I found to prepare the students for a unified understanding of speech is extensive practice in **spectrogram decoding.** At the very beginning, the students learn to characterise non speech, familiar sounds, and synthesised sounds, with varying intensity, pitch, formant frequencies. Then they are taught how to recognise the five vowels /oe/, /i/, /a/, /u/, /y/ and nasal /a/ in connected speech. The perturbation theory is then used to introduce the influence of the consonants on the neutral vowel [oe], and the formant directions in the transitions in /peop/n, /toet/, /koek/ and /roer/, etc. At the end of the second year of progressive learning, most of the students become decent spectrogram readers. The student is taught to address, at the very beginning, the same detailed attention to how each French sound is produced, its pertinent acoustic characteristics and its corresponding perceptual sensations, using a speech editor to extract the sound from connected speech. Needless to say, teaching the basics of Fant's theory is facilitated when addressing fairly good spectrogram readers.

c) Easy access to speech analysis facilities for the teacher and the student.

Computer speech analysis allows phonetics teachers to do a quick printing of high quality spectrographic representations, waveforms and spectra, which are distributed to the undergraduate students who don't have direct access to speech analysis facilities.

The availability of an inexpensive analysis software (such as SNORRI) [6] producing very high quality spectrograms and affordable Audio Cards may allow in the near future to equip a number of PCs outside of our laboratory, giving to some undergraduate students access to speech analysis facilities.

e) Speech analysis programs as pedagogical tools

Demonstrations with analog Sonagraph, pitch detector and filter banks have a strong pedagogical impact on students, and should be used whenever possible in parallel. Video recordings of the demonstrations using the analog devices can be displayed in the classroom.

Since analysis of the speech signals employs digital techniques, the acquisition of knowledge of digital signal processing, in particular **filtering** and **sampling**, has become necessary for all students. One way of teaching is to let them perform first *inadequate* filtering and sampling, and then to teach the basic knowledge necessary to do the right thing.

e) Pencil, paper and eraser for the students

Needless to say, a very important part of the teaching should be done with pencil, paper and eraser:
- hand measurement of the fundamental frequency directly from the signal in the time domain, or using the tenth harmonics in the frequency domain,
- estimation of the area function from a few typical VT sagittal profiles,
- modelling of simplified VT configurations corresponding to [i], [a], [u] and [y] by a small number of connected tubes, hand calculation of the natural resonances of each isolated tube, identifying the resonance mode, and determination of the associations between resonance cavities and formants.
- hand calculations of natural frequencies in coupled resonators representing different place of constriction along the vocal tract for consonants,
- etc.

A VT acoustics simulation should be used only after the students have a good understanding of what is going on, for example,
- listening to the sounds corresponding to the hand calculations they have done,
- estimating the effect of radiation impedance and yielding walls, and
- comparing modelling with simple tubes with a more realistic model of VT shapes,
-etc.

Too early use of computer simulation (a source of fun, indeed) may hinder correct understanding.

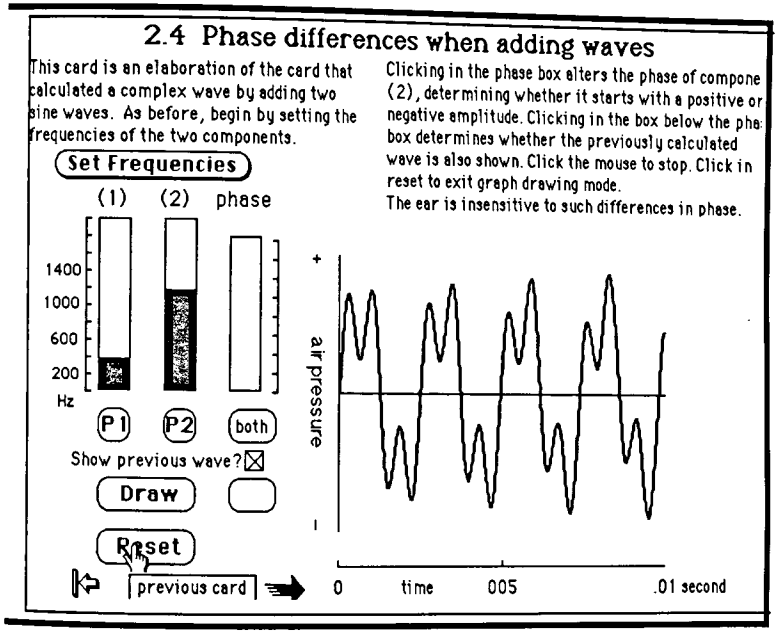


Figure 1: One of the sHyperCard tsack "Acoustic Phonetics", illustrating the effects of adding waves and of phase differences. It is possible to set the frequencies of the two components, sine wave P1 and P2, to listen to each component separately, and to observe the complex wanves and to listen to the resulting sound. It is then possible to alter the phase of component P2 in relation to P1, and to compare the resulting sound with the previous sound.

f) "Acoustics Phonetics" HyperCard teaching stack for the instructor and the students

Not all the necessary notions are easily taught to beginners. Let us take some examples.

A critical point is to let the undergraduate students understand how any waveform can be analysed as the sum of two or a greater number of sine waves (*Fourier analysis*). One way is to let them add pure tones (sine waves) on graph paper to draw a complex wave shape, and then to propose to the students the reverse problem: the decomposition by hand of a complex wave into its two sine components.

I have recently tested with success one of the lesson in 'Acoustic Phonetics' HyperCard stack [7]. The card allows the learners to hear two sine waves separately and the corresponding sounds (see Figure 1). A short demonstration will be done in front of you by the author of the program, Peter Ladefoged: selecting particular frequency for the first sine wave P1, hearing P1, choosing particular frequency for the second wave P2, listening to P2, combining both P1 and P2, observing the resulting complex waveform and listening to it.

This kind of demonstration seems to be very effective as a first approach to Fourier analysis. Let us take a second example. Intuitively, the students assume that two waves that are very different on paper should sound very different, and have trouble accepting that the human ear is not particularly sensitive to (constant) *phase*. The previous HyperCard program allows one to change the phase relation between P1 and P2, to listen to the resulting sound, and then to compare it with previous sound.:

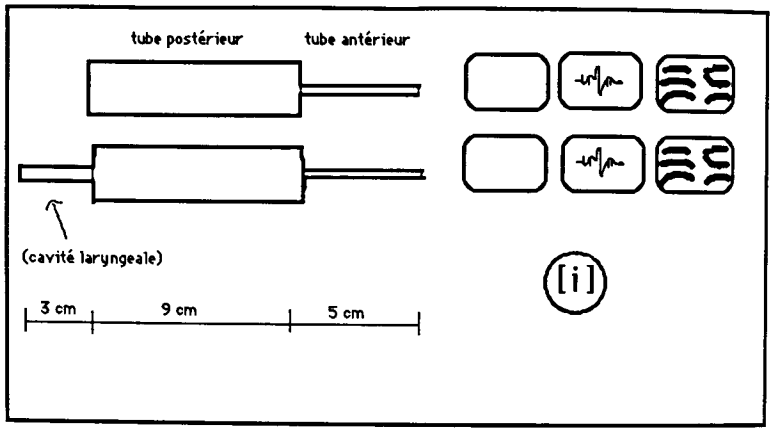


Figure 2: The vowel [i] synthesised with two (9 cm long and 12 cm2 in area, and 5 cm long and 1 cm2 in area) or three tubes. By clicking on the appropriate box, it is possible to listen to the synthesised sounds, or to observe the corresponding signals and spectrograms .The demonstration aims to show the improvement by adding a third tube (3 cm long, and 2 cm2 in area) simulating the laryngeal cavity.

The comparison done by the ear speaks for itself. Practical experience helps one to believe what the books say.

d) Synthesis and simulation programs

A course on acoustic phonetics should provide a good understanding of the links between (i) production and the acoustic properties of the signal, and (ii), the acoustic properties of the signal and perception.

Fortunately, the first link, i.e. the link between VT configurations and acoustic properties of the signal is rather well understood. **Fant**'s famous book "Acoustic Theory of speech production" [8] is an authoritative. guideline for the teacher of phonetics. Note that much of the mathematical specification contained in Fant's book cannot be acquired by average linguistic students (and luckily, it is not strictly necessary). Lehiste's "Readings in Acoustic Phonetics" is also an excellent source of inspiration for setting up the content of the courses [9]. Our research students are invited to use also intensively Stevens's many articles ([10] for example).

The other link, i.e. between the acoustic characteristics of the sounds and what is effectively **perceived** by the listener(s) is less well understood. Of particular interest are the acoustic explanations of

- **compensating** and **reinforcing** movements (two simultaneous changes in VT, such as lip rounding and larynx height can reduce or enhance acoustic effect,
- sound **confusions**, and
- sound **changes**.

Both links can be studied in parallel, thanks to modern computer technologies which provide quick calculations and sound facilities. Vocal tract acoustics simulation programs make it feasible to manipulate one articulatory parameter at a time (tongue position, jaw position, lip height and width, lip protrusion, etc.). Formant synthesisers make it feasible to manipulate one acoustic parameter at a time (formant frequency, bandwidth, etc.) Thanks to both type of these programs, it is possible to investigate the links between VT (articulation) and Fn (acoustics) and the auditory responses (perception) altogether.

To summarise, the programs I used most frequently in acoustic phonetics courses are:

a) a version of Klatt's **formant synthesiser** [11] on PC

b) **Two-tube** model, **Fant's model**, a variable number of tubes model, with optional on/off radiation load, and optional rigid/yielding walls (programs written by Maeda, for PCs). The insertion of a noise source is under development.

c) Maeda's **vocal tract acoustics simulation** program on PC (with nasal cavities)

d) and recently "Acoustic Phonetics" HyperCard Stack from Ladefoged [7], on Macintosh.

In addition, vocal tract acoustic simulation programs developed at ICP at Grenoble [12] have been implemented on Macintosh and there is also an UCLA vocal tract model available (also on Macintosh). Note that we still lack at our institute a good simulation program on aerodynamics, and on the glottal source. Suggestions and eventual offers from the audience are really welcome.

It is important to note that the same programs are used:
- for demonstration in front of the class-room for all students,
- for a series of assignments given to the students having access to the lab facilities (some of the exercises are inspired by the excellent course initially set up by Dennis Klatt at MIT) [13] and
- as a research tool (cf [2] and [3]).

If there is enough time, I would like to show you four other demonstrations of what can be done with simulation programs.

First, any **formula** becomes attractive and easier to internalise if each of its variables can be changed at a time in front of the students. For example, concerning with the quarter wave length resonance equation, we can change the sound velocity of the air simulating speech in helium, and manipulate the length of the tube simulating the neutral vowel /oe/ produced by an infant or giant vocal tracts.

Second, what is the perceptual effect on formant patterns of a **laryngeal cavity** to the specification of a transfer function of a vowel calculated by computer simulation? This can be demonstrated by synthesizing sounds with and without the laryngeal cavity (Fig. 2). The students can observe the appearance of a supplementary formant around 2800 Hz on [i] spectrum and they can hear the differences. As you can judge by yourself, the new [i] (with added "laryngeal cavity") sounds nicer, more clear than the sound corresponding to a two tubes representation.

Third, simulation may help the students to understand the basic idea underlying most of a **theory**, which may seem, as first, rather abstract. The following 'figure parlance' serves as an introduction to *Stevens's quantal theory* of speech [10]. You are invited to listen to simulated /a/ sounds with two-tubes model and transcribe what you hear on a sheet of paper. In the critical region where the change in formant-cavity affiliation occurs, a stability in the sound quality is perceived but in the two outside regions, the same amount of displacement leads to a change in the quality of the vowels.

Fourth, the following card demonstrates the relation between the degree of velar lowering and the degree of perceived nasality for the vowels /i/, /a/ and /u/ [14]. Such demonstration can be used to draw a link with the **phonology** course on the distinctive use of nasalisation in the vocalic systems.

g) Portable computer for the teacher

In Rousselot's time, the very few students of phonetics in Paris were acquiring the necessary notions by doing experiments in the phonetic lab, under the supervision of their teacher. Thanks now to availability of portable computers, such as the one I am using now in front of you, it is possible to bring **part of the lab** into larger classrooms. A small portable computer can quickly become your most comprehensive assistant for teaching, and offer you technical assistance with rapidity, patience, precious memory, and constancy.

A portable computer can advantageusly replace in many cases a **tape-recorder.** Any sound can be stored in the computer, quickly accessed, segmented using a simple speech editor such as SoundEdit on Macintosh, and repeated at will, allowing one to set up spontaneous identification and discrimination experiments in the class-room. Note that the speech files can now be easily transferred between PC and Macintosh computers and stored on disquettes eventually distributed to the students as a complement to the written version of their courses.

New technics allows also to project the display of the computer onto a large screen via a **transparency** projector. Traditional silent pictures can be advantageously replaced with transparencies with sound possibilities.

k) The use of the French vocalic systems

Finally, the use of the French vocalic system seems pedagogically well adapted for acoustic phonetics courses.

According to my own experience decoding French spectrograms is easier than decoding English spectrograms (at least in carefully spoken isolated sentences). The use of rich French and well tempered vocalic system makes it possible to illustrate distinctive labialisation and nasalisation.

Two vowels are particularly precious /y/ and nasal open back vowel /ɔ̃/. The inadequacy of an **F1 versus F2-F1** representation for the vowels is easily demonstrated to the students: /i/ and /y/ can have about the same F1 and F2, but still sound very different and are almost never confused on spectrograms. **Exchange of cavity affiliation** for a formant is also easily understood by the [i]-[y] pair. Nasal vowels are useful to illustrate the acoustic and perceptual consequences of **side cavities.** The existence of the nasal vowel [ɔ̃], which is [+nasal] and [+round] is particularly adapted for studying **co-ordinated anticipatory coarticulation** (the lips and the velum).

3. CONCLUSION

A number of good standard phonetics books, Fant's book, classical articles, traditional pendulum and tuning forks, paper, pencil and eraser, speech analysis facilities, a portable computer, the availability of pedagogical multimedia computer programs, extensive use of computer simulation, etc. seem to facilitate the task of teaching acoustic phonetics to all types of students. Furthermore the use of synthesis and vocal tract acoustic simulation programs seem indispensable for teaching the link between articulation, acoustics and speech perception. The attempt by Peter Ladefoged setting up a multimedia course on Acoustic Phonetics on HyperCard, with self-training and sound facilities seems to me a new way of teaching that should be pursued: listening is very important for phoneticians, and traditional books are too silent for allowing a deep understanding of certain aspects they describe. Transforming classical figures taken from the literature as "speaking pictures" and systematically attaching speech files to the Word version of research reports seems to me very useful.

It is important to make the best use of technological potential. The continuous technical progress, data base on CD-ROM, multimedia (text + sound + images), extended network possibilities, broadcast and interactive television, will probably change the way of teaching and of learning in the years to come. In particular, a complete platform for improving the teaching of all aspects of phonetics has become technologically feasible. It has still to be done. It may be very difficult to obtain funding for such a project. It is also difficult to find the necessary time to collect all the free data available. If a collaboration via e-mail and Interet between enthusiastic teachers of phonetics and engineers willing to share their talents and resources could be established, some progress can be made before the next congress of phonetic sciences.

But already now, and thanks to generous donors, it is possible, without any expense or with a moderate expense to improve teaching environment for the large mass of students who don't have access to laboratories facilities. It seems also possible to provide the students who want to engage in multidisciplinary research involving acoustic phonetics with the tools they need to solve the problems, using a limited mathematical background.

ACKNOWLEDGEMENTS

Thanks again to the generous donors, engineers and phoneticians alike, Shinji Maeda from CNRS, Gérard Bailly and Louis-Jean Boé from ICP in Grenoble, Peter Ladefoged at UCLA, Yves Laprie from Nancy for SNORRI, and also to Ken Stevens for his inspiring teaching. Thanks also to my students, for the fun I have to prepare courses for them.

REFERENCES

[1] Ladefoged, P. (1995), "Teaching Phonetics", Introductory comments, *ICPhS 95.*

[2] Yeou, M., & Maeda, S. (1995). "Uvular and pharyngeal consonants are approximants not fricatives: An acoustic Modelling study", *ICPhS* 95, . Stockholm.

[3] Pillot, C. (1995). "Production and perception of the singer's formant". *ICPhS 95,* . Stockholm.

[4] Spectacular Disastrous, Warner Brother's video.

[5] Ladefoged, P. (1962), *Elements in acoustic phonetics",* University of Chicago Press, Chicago (last revised version, 1995).

[6] Laprie, Y. (1989), "SNORRI: an Interactive Tool for Speech Analysis", *Eurospeech 89* Paris, Vol. 1, pp 613-616.

[7] Ladefoged, P. *Acoustic phonetics,* HyperCard Stack.

[8] Fant, G., (19xx), *Acoustic theory of speech production,* The Hague.

[9] Lehiste, L. (1970), *Readings in Acoustic Phonetics,* MIT Press, Cambridge.

[10] Stevens, K., (1989), "On the quantal nature of speech", J. Phonetics, 17, 3-45.

[11] Bailly, G. and Tran A. (1989), "Compost: a rule-compiler for speech synthesis", *Eurospeech 89* Paris, Vol. 1, pp. 136-139.

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

[13] Stevens, K.N., and Perkell, J., (1993), *Laboratory on the physiology, acoustics and perception of speech,* Course 6542J, 24966J, Massachusetts Institute of Technology.

[14] Maeda, S. (1993), "Acoustics of Vowel Nasalization and articulatory shifts in French Nasal Vowels", in Nasals, nasalization, and the velum, Huffman, M.K. and Krakow, R.A. (eds)., Academic Press, San Diego, pp. 147-166.

PROBLEMS OF INTONATION

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

DELIMITING THE PROBLEM

My advertised role in this symposium is to talk about "problems of intonation" in teaching phonetics. The first step is to delimit what we mean by intonation, and to understand why there are problems. So, to repeat my earlier definition, let me delimit the object of study as: "**all** aspects of the perceived pitch pattern that the speaker intends for the hearer to use in understanding the utterance, or that the hearer does use whether intentionally controlled by the speaker or not" [1]. Intonation is one of the most difficult aspects of speech to teach about, for the same reasons that it is one of the most difficult aspects of language sound structure to model.

The problem of meaning

One difficulty stems from the kinds of meaning that many intonational categories have, which are like the meanings of such words as *the* versus *a* in English, or the meaning of choosing the indicative versus subjunctive form in Portugese. They can signal often very subtle facts about the relationship between an utterance and its larger discourse context.

For example, in the American English intonation system, there is a categorical contrast between high tone (H*) and low tone (L*) pitch accents which in different discourse contexts can be interpreted as the difference between a statement and a yes-no question, between an imparting of new information and a gentle reminder of old information, between an affirmative repeating and an incredulous echoing of what the other speaker has just said, or between the literal use of a word such as *now* and its use as a discourse marker for a shift in discourse topic. What do all these usages have in common? Pierrehumbert and Hirschberg [2] have suggested that H* means commitment — the speaker intends the hearer to add a particular proposition or entity to the background of mutually believed information at that point in the discourse — whereas L* means lack of commitment, either because the speaker intends the hearer to provide the correct information (yes-no question), or because the speaker knows that the hearer already has the information (gentle reminder), or because the speaker questions the other speaker's intent to add the information (incredulous echo), or because the speaker intends that the word should not be taken as part of the substantive information content of the narrative (discourse marker usage).

This is a meaning difference that is inherently more difficult to characterize than many of the lexical differences that we use to support our analyses of the consonants and vowels of a language. A speaker of English need only point to exemplar objects to characterize the meanings of *mat* and *bat*, and thus get across that these are different morphemes. And it is not difficult to find a slew of other morphemes such as *mitt* versus *bit* or *moat* versus *boat* to support an analysis that localizes the difference to the beginnings of the morphemes in order to claim that [m] and [b] are discretely different consonant sounds that are differentiated in onset position in English.

By contrast, there is fervent argument among experts concerning the analogous questions in the analysis of the intonational morphology of English, and even those ultimate experts — children acquiring English as their native language — apparently find these kinds of meaning differences more difficult to learn to control (see [3], [4]). It is hard to get the phonetic skills relevant to studying intonation without also giving oneself at least a nodding acquaintance with the related literature in semantics, pragmatics, and discourse structure.

The alphabetic (dis)advantage

Intonation is difficult to teach also because our modeling of it begins from scratch. We do not start with an already highly practiced phonological theory, as we do when teaching our students how to analyze consonant and vowel contrasts.

Perhaps because the orthographies of English, French, and German are conservative, we tend to concentrate on the difficulties that our students' literacy imposes, to bemoan how difficult it is to get students to transcribe a form how they hear it and not how they know it is spelled. However, research on speakers who have learned only a logographic writing system [5] (or who find it difficult to learn to read an alphabetic one [6]) makes it clear that learning to read in an alphabetic writing system imparts (or requires) a very sophisticated meta-linguistic phonological awareness, an awareness that we take full advantage of in our standard method for teaching about vowel and consonant contrasts. We simply assume the linear segmental phonological analysis of the International Phonetic Alphabet, and concentrate on teaching other things — on developing the ability to hear and mimic consonant and vowel contrasts not in our students' native repertoires, and on imparting basic facts and physical principles that our students need to know in order to understand how the contrasts are produced and perceived.

However, this method is not without cost. Our blithe assumption of the alphabetic model creates the Platonist illusion that consonants and vowels exist in nature already segmented that way, that there are countable "sounds" out there for the phonetician to study independent of the phonological analysis. We imagine that we can get all the skills that we need in order to know about how structure is transmitted in the speech stream without developing our ability to imagine and evaluate alternative models of the structure transmitted. This self-delusion then handicaps us when we try to teach about "suprasegmentals" — about any feature of language sound structure that defies this particular phonological model, as intonation so clearly does. So we teach our students about the physiology of the larynx and the physical relationship between vocal fold tension and fundamental frequency, and we are baffled that it is so difficult to make them understand the dimensions of intonational variation available to them in their native languages, or to seewhy ostensibly the same pitch contour can function so differently in different parts of an utterance or in utterances from another language.

The difficulty of prosody

This difficulty is compounded by the fact that so many aspects of intonation are crucially linked to prosodic constructs such as the syllable, stress, and phrasing. For example, we cannot explore such intonational constructs as pitch accents and boundary tones without recourse to prosodic units such as the syllable and the intonational phrase, and we cannot explore the different functions of pitch accents in English versus Japanese without recourse to the prosodic construct of stress. These prosodic constructs are themselves difficult to model for our students, again because they do not fit the alphabetic model, but also for a more fundamental reason. I think prosody is difficult by its very nature, because prosody is all about the organizational structure of speech (as opposed to its contrastive structure), and I think organizational features are harder for the human mind to grasp.

Let me try to explain this idea with a metaphor. Think of a jacquard weave in which the design is made not just by varying whether the warp or the woof is the surface thread, but also by varying the thickness along the length of each thread and the saturation and color of the dye. These are all features that can be modeled to some extent without specifying how any particular variation functions at any place in the weave. To predict the effect of some change in the dye, for example, the expert weavers need to know about the physical characteristics of the dye and how color is perceived and so on. But now consider the shapes that the weavers are trying to build in the surface of the cloth. To model these, the weavers will need to know about such things as edge detection and so on, but also they need to specify what aspects of the thread they can vary at any place in the weave to define the desired edge.

Features of vowel timbre, consonant place and manner, phonation type, pitch level, and so on, are easier because they are all thread features. We can talk about them to a first approximation without thinking about how they function in any particular part of the organizational structure of a given language. For

example, we can talk about segmental effects on the pitch contour and the psychoacoustics of pitch perception and even the categorical intonational features H ("high tone") versus L ("low tone") without specifying whether these features function to differentiate monosyllabic verb morphemes (as in Yoruba) or differentiate single-tone pitch accents (as in English).

Prosodic constructs such as the syllable, stress, and phrasing, on the other hand, are like the shapes in the cloth that the weavers make by alternating woof and warp, thick and thin, or dark and light colored thread. They are features of the organization itself which depend crucially on what thread features can alternate where. They are about that difficult question precluded by the Platonist illusion: What are the possible units of counting for a particular language? What are the segmentations that the speaker can intend for the listener to parse in the signal?

OVERCOMING THE PROBLEM

Of course many other aspects of language sound structure are also very difficult. The standard acoustic model of vowel contrasts is extremely difficult for the unfortunately large number of our students who lack the background in physics and the necessary mathematical skills to grasp the basic principles of resonance. But (as Vassiere points out in her paper) there are dramatic examples — such as the film clip of the Tacoma Narrows Bridge bouncing rhythmically in the breeze as it gets ready to break apart — to give an initial intuitive feeling for the phenomenon, which then can be used to motivate a more precise handwaving.

An important element to success in teaching the skills relevant to intonation, then, is to exercise just this kind of showmanship. In the rest of this paper, then, I will try to give a few of the examples from English and other languages that have worked for me in developing strategies for overcoming each of the three main sources of difficulty for intonation.

Start with some easier meanings

One common strategy for overcoming the problem of meaning is to first teach about a language in which different tone patterns differentiate ordinary lexemes, before teaching about the difficult-to-characterize pragmatic morphemes that tone patterns differentiate in English, German, or Italian. If the language chosen for this purpose has an intonational system such as that of Swedish [7] or Osaka Japanese [8], a system which allows a relatively easy progression from hearing the differences in short citation forms to decomposing longer forms into aspects that are part of the lexical specification of the words and aspects marking more elusive discourse functions, this method can provide a more gentle introduction into the terrain. It allows the students to concentrate first on some of the other skills that they will need in order to overcome the alphabetic disadvantage, before venturing into the crags and crevices of pragmatics and discourse structure. The method only works, however, if the teacher is a fluent speaker of the language, or if the teacher can bring into the classroom native speaker consultants who have strong intuitions about pragmatic felicity and about any relevant prosodic constructs.

The alternative strategy is to choose to teach first the intonational system of a language which all of the students must know very well to be in the class — the language of instruction — and to choose the initial examples to be as vivid and salient as possible. In some places, such as the linguistics department of the Ohio State University, where I teach, many of the students will even be native speakers of the language. The first examples then can be situations in which failing to differentiate two intonation patterns can have an embarrasingly funny effect, as in the two utterances:

(1) Mary does intonation.
L+H* L- L%

(2) Mary does intonation.
L*+H L- H%

which I have used to illustrate the difference between the L+H* and L*+H accents in English. I offer the two as alternative responses to the claim that "Only crazy people do intonation." The students in the class who are native speakers of English laugh at the second response (which implies that the teacher is crazy), and those who are not become very motivated to learn to control the contrast between the two tunes.

Sometimes the most salient examples involve a difference in interpretation that might be signaled also by a syntactic difference. The L* H- H% pattern in American English, for example, is often called the "yes-no question contour" in differentiation to the "declarative contour" H* L- L%. It is convenient to use a yes-no question and a declarative sentence with these tunes, in order to build two parallel series of utterances showing how the H- versus L- phrase accents extend over longer and longer regions as the nuclear pitch accent is moved earlier and earlier. However, the teacher must be careful also to provide examples of utterances where H* L- L% occurs on a yes-no question and where L* H- H% is used as a statement (e.g. the statement of incredulous disbelief), lest some students get the mistaken impression that some tunes are "syntactic" and therefore more worthy of their attention (or less worthy, depending on their feelings about areas of grammar other than phonology and phonetics). The example I like to use is the joke about the man forced to read a public confession, who renders the confession with great feeling, dividing he sentences into many intonational phrases:

(3) I || was wrong, || and Stalin || was right. || I || should be vilified. || ...

and putting a L* H- H% pattern on each. It will be obvious from the context of the story that the man cannot be asking his audience a series of yes-no questions.

Because jokes and funny stories in general have this advantage of building an explicit fixed context into the performance of the example, they can also be used to help get across an important lesson. In order to investigate intonational form, one must carefully observe (or even actively manipulate) the discourse context for the speaker. As part of inculcating good observational habits, students will need to be debunked of the notion that asking the speaker to recite a particular sentence or phrase for the field linguist somehow provides a "neutral" context, and that the prosodic organizations and associated intonation contours produced in this context have special status as "neutral" or "default" patterns. We need to make it clear that all this method accomplishes is to make it impossible to observe the context which the speaker then implicitly imagines in order to be able to recite the form. If this makes the students feel a bit daunted at the prospect of having to learn how to judge discourse contexts as well as to hear tunes and stresses and phrasings, better to have them feel daunted than to send them into the field thinking that they will be able to model intonation in a new language without acquiring all of the requisite observational skills.

Use the F0 contour as the narrow phonetic representation

When teaching an introductory general course in phonetics, one may be tempted to try to shoehorn intonation into the alphabetic model by teaching the IPA symbols for "high tone", "falling word accent", "global rise", "downstep", and so on. I think this is a mistake.

Of course, students specializing in a particular language or language area may want to learn some of these symbols in order to have access to the literature and fieldwork notes of other linguists. For example, students specializing in Chinese or Tibeto-Burman linguistics may want to learn the nicely iconic Chao tone letters [9], which are the basis for the current IPA symbols for "tones & word accents". On the other hand, they will also need to learn the less iconic superscript numbers transliterating the tone letters, since this is the transcription that they will encounter in reading this literature, and students specializing in Bantu languages will need to learn the even less iconic non-IPA diacritics that many Africanists use [10]. But such specialist reading knowledge is quite different from the phonetic skills that one needs to teach to enable the students to observe intonation patterns and analyze intonational systems *in vivo*. Since the segmentations imposed by the alphabetic analysis are so wrong for intonation, I think it is better to avoid any symbolic transcription until one is confident of the phonological categories. I think it is far better to sidestep the IPA entirely and expend the same effort on teaching students to look at raw F0 contours while making stylized drawings of the pitch contours that they hear. In other words, we should train the students to use the F0 contour in lieu of an initial "narrow transcription" in doing fieldwork.

Since this may be construed as a radically anti-IPA stance, let me explain it further by considering where the narrow transcription fits in a standard strategy for teaching observational skills relevant to analyzing consonant and vowel systems, the strategy that structures the sequence of chapters and topics in Peter Ladefoged's textbook [11], which we use in our introductory course at Ohio State University. The first step in the sequence is analogous to the second strategy I suggested above for overcoming the problem of intonational meaning. We begin by teaching the students to transcribe the consonant and vowel phonemes of the language of instruction, which for many students will be the native language. Then we make them produce narrow transcriptions of that language, before we go on to teach them about other languages. We use that intervening step of narrow transcription to reduce the phonological interference from the first language that will otherwise hinder their observations when they go to do fieldwork. We do this because it is far less threatening than simply putting them into the field, where they would have to worry about the meanings of forms and a host of other things (such as the cultural norms about eye contact and so on), at the same time as they are trying to learn to hear contrasts not found in the native language. In other words, by treating allophonic variation in consonant and vowel quality in the native language as if it were alphabetic contrast, we give the students a Zen exercise to help them turn off the attentional skills built up over a lifetime of learning to perceive the phonemes of the native language. This works because we assume that for any language they will encounter in the field, it would not be disastrously wrong to posit as a first working hypothesis, a linear segmental analysis of the language's system of consonant and vowel contrasts.

Applying the same strategy to the analysis of intonation is a recipe for making blindfolds. It may be safe to assume that an alphabetic tone-sequence analysis will work for any new intonation system encountered in the field, but *a priori* we cannot assume any of the other facts we need to have in order to uncover the tonal analysis, answers to questions such as: Do the different tones contrast paradigmatically in composing pragmatic morphemes (as in English) or lexical specifications (as in Cantonese), or is their primary role the syntagmatic marking of prosodic constructs such as stress (as in Danish) and accentual phrase edges (as in most dialects of Korean)? What are the relevant prosodic constructs for anchoring the tones to the consonant and vowel features in an utterance, and are there contrasts in temporal alignment between the tones and the anchor site? How do these facts about the relationship between tone contrasts and meanings and about the anchoring of tones to other features of contrastive structure translate into density of tonal specification? Are some tone sequences or some prosodic constructs associated with a systematic manipulation of the more global pitch pattern, such as downstep or pitch range reset, which can obscure the local tonal values? Intonation systems can differ so markedly along all of these dimensions that prematurely adopting any symbolic transcription obstructs the observation of variation in pitch necessary to getting the phonological analysis. To give just one example, if Janet Pierrehumbert and I had not decided to use a nonsymbolic phonetic representation, the F0 contour, we would not have been able to observe the relationship between F0 slope and phrase length, and so on, observations that challenged the traditional narrow phonetic transcription with its mora-by-mora specification of high versus low tone and suggested the analysis of the Japanese intonation system proposed in [12].

This is not to say that I do not advocate teaching symbolic labels for intonational categories. Far from it. I use the ToBI labels H*, L*, L+H*, and so on, every time I teach about English intonation, because the literature from the past half century of work on the intonational categories and meanings of American and Southern British English makes me confident that this analysis will work for the dialects spoken by the majority of my students. My confidence is bolstered by the high degree of intertranscriber agreement documented for this transcription system [13], and by the fact that the system has been taught successfully to non-native speakers

employed to label spontaneous dialogue [14] and has even been used by a non-native speaker to teach about English in a graduate phonetics seminar on intonation [15]. A phonetician teaching intonation to a class of Swedish speakers or Japanese speakers or Spanish speakers or Mandarin Chinese speakers similarly would be justified in teaching a symbolic transcription of the intonation patterns, because we know what the categories are. But when there is no established body of research on intonational form and meaning for a language, we would do better not to use that language as the obvious first extended example even if all the sudents in the class are native speakers. In such a case, we might teach the students first about a language that does have an established phonological analysis. We might give them the necessary fieldwork skills to control the pragmatic context in eliciting forms and to interpret the F0 track, and encourage them to embark on the research toward such an analysis for their own language. But we should not fudge by offering them a symbolic "narrow phonetic transcription" of intonation to stand in until the real work is done.

What skills do our students need in order to use the F0 track in lieu of a narrow transcription? The most important thing at first is that they will need to learn about microprosody. One might be tempted to "doctor" F0 contours for the students at early stages by digitally smoothing and manually whiting out any apparent "perturbation" remaining after the smoothing, but I think this is a mistake. No non-intelligent smoothing procedure can do what the human mind does, and the students need to train their eyes to do what their ears' minds do: to parse over the effects on intonation of consonant and vowel features when they are "listening" for tonal contrasts. We do better to start this training immediately, by giving them examples of the same tone pattern anchored to different consonants and vowels. When they have mastered the art of visually "listening" through the microprosody, then we can tell them about the places where F0 is not a good measure of pitch. They need to know to look for the manifestation of vocal fry and other non-modal phonation types. At this point it is good to set them to designing little experiments. We can ask them how they might go about comparing slopes of two different kinds of F0 rise in the language being studied, or how they might test competing hyptheses about the timing of a tone relative to some prosodic landmark.

Grapple with prosody

Which brings us to the most difficult area of all. There is no way to tackle intonational analysis without grappling with the problem of prosody. We must teach our students about the nature of prosody and give them the skills to uncover the prosodic constructs that native speakers of a language use to structure the intonational categories and other contrastive features.

Again, the best strategy here follows the same lines as the best strategy for teaching about tone features. We can start with a language that all of the students know fluently, where we can tap their intuitions about relevant prosodic constructs such as how many syllables are in a word or how many intonational phrases are in a longer utterance. For each prosodic construct, it is good to start with clear cases, where native speakers (and linguists) tend to agree on the number or location of relevant prosodic landmarks, before proceeding to cases where the analysis is more unsure or simply unknowable. For example, in teaching about stress in English, it is best not to start with words such as *thirteen* and *Malay*, where different students may produce different nuclear accent placements in the citation form, and which would lead us directly into the nettles patch of stress shift. (Note that these are the words that some pronunciation dictionaries transcribe as having two primary stresses.) On the other hand, we also should not gloss over the difficulty of prosody, but attempt from the beginning to get across the idea that it prosody about organizational structure. It is a mistake to think we can make it easier for them by first offering misleading definitions in terms of contrastive structure. For example, we are unwise to say: "Just as tone can be defined in terms of relative pitch, stress is the relative loudness of a syllable." A stressed syllable may indeed sound louder than an unstressed syllable at any

level of the stress hierarchy, but this phonetic equation will lead the students straight down the dead-end track of looking for stress in the RMS amplitude contour.

I find that it is easiest to grapple with prosody if I first cover the phenomena of "coarticulation". That is, I place prosody in the pedagogical sequence where I can more easily show how prosodic structure is realized substantively by often very subtle differences in coordination among different contrastive features. I then can relate questions about prosodic constructs to questions that have already come up in the discussion of segmental features and coarticulation, such as: "What is the difference between an affricate and a cluster of stop plus fricative, anyway, and why isn't the [ts] at the end of *cats* an affricate?" In answering these questions I will already have stressed a point that is crucial for understanding higher levels of structural organization — namely, that what we transcribe as ostensibly the same set of contrastive features organized in ostensibly the same sequence can constitute different structures in different prosodic positions within a language and across languages. So in English [tʃ] is an affricate when the fricative is performed as the release phase of the postalveolar stop (as in the phrase *catch it*), but it is a sequence of stop and fricative when the [ʃ] is phased later to allow the [t] to have its own alveolar target. This is a good example for talking about syllable boundaries, too, and leads them to look for the distributional cues that support the structural analysis of [ts] as an affricate in German.

With examples like this, one can emphasize that segments are not Platonic entities existing out there in nature separate from the phonologies of actual speakers of a language, and that the students cannot go about finding out whether [tʃ] is an affricate or a cluster in the same way that they might go about finding out whether initial voiceless stops in Indian English are aspirated or not. At this point they will be ready to be told the same fact about syllables, stresses, phrases, and so on. If there are native speakers of both English and Japanese in the class, for example, one can demonstrate that [ski] is a single syllable for the English speakers, but definitely two for the Japanese speakers. This then can lead into difficult cases in English, where reduction makes the syllable count difficult. For example, we can get them to design experiments to see how much a native speaker can reduce the first vowel in *supports* be before the word comes to be indistinguishable from *sports*. This in turn can illuminate one's definition of the lowest level of stress contrast in English.

In teaching about stress, I find it also helpful to give examples of languages where this most basic level of prominence contrast is marked by rather different phonotactic constraints. For example, since the language of class instruction is usually English for me, I can tell them about Swedish, where stress can be defined at the lowest level in terms of phonemic length contrasts (a stressed syllable has a phonemically long vowel or is closed by a long consonant), and about Mandarin Chinese, where it can be defined in terms of the tonal specification (i.e. an unstressed syllable is one with the so-called "neutral tone"). Then I can show how the seemingly very different phonological hallmarks for English and Mandarin are associated with nearly identical facts about phonetic coordination and undershoot, with unaspirated stops being voiced foot internally in both languages, and unstressed vowels often being reduced to a syllabic realization of a neighboring consonant, and so on. Also, this lets me make the important point that in English, as in Swedish, pitch accents can be anchored only to syllables that are stressed at this more basic segmentally defined level, and that pitch accent then defines another level of stress contrast above the basic level.

If the class includes many students who are taking phonetics as part of a program in speech pathology, another way to get across some of these points is to assign them to transcribe the productions of the very young children whom they see in the speech clinic. Since this is a standard part of the diagnosis for phonological disorder, they will see the relevance of this exercise immediately. The productions of very young children are often very difficult to transcribe because they have not yet achieved the precision in temporal coordination that supports a segmental analysis for adult speech. Pointing out this source of difficulty makes it easier to highlight the link between segmentation at all levels of the prosodic hierarchy and timing control.

Once the students have more background in phonetics and phonology, one can also have them exercise their analytic imaginations in a similar way by reading about nonsegmental analyses of consonant and vowel contrasts, such as Browman and Goldstein's Articulatory Phonology (e.g., [16]) and the Neo-Firthians (e.g. [17]) and the arguments about the segment in Steriade's recent work [18].

Above all, I think we must be humble about our current state of knowledge about the phonetics of prosody. The most helpful answer often will be simply: "We don't know, yet." But let us be sure to phrase it that way, to make the "we" inclusive, inviting the students into this grand enterprise.

REFERENCES

[1] Beckman, M. E. (1995). "Local shapes and global trends," This conference.

[2] Pierrehumbert, J. & Hirschberg, J. (1990). "The meaning of intonation contours in the interpretation of discourse," In P. R. Cohen, J. Morgan, & M. E. Pollack (Eds.), *Intentions in Communication*, pp. 271-311. Cambridge, MA: MIT Press.

[3] Cruttendon, A. (1974). "An experiment involving comprehension of intonation in children from 7 to 10," *Journal of Child Language*, vol. 1, pp. 221-232.

[4] Cruttendon, A. (1985). "Intonation comprehension in ten-year-olds," *Journal of Child Language*, vol. 12, pp. 643-661.

[5] Read, C., Zhang, Y. Nie, H., & Ding, B. (1986). "The ability to manipulate speech sounds depends on knowing alphabetic spelling," *Cognition*, vol. 24, pp. 31-34.

[6] Share, D., Jorm, A., Maclean, M., & Mathews, R. (1984). "Sources of individual differences in reading acquisition," *Journal of Educational Philosophy*, vol. 76, pp. 1309-1324..

[7] Bruce, G. (1977). *Swedish Word Accents in Sentence Perspective*. Lund: Gleerup.

[8] Kori, S. (1987). "The tonal behavior of Osaka Japanese: an interim report," *Ohio State University Working Papers in Linguistics*, no. 36, pp. 31-58.

[9] Chao, Y.-R. (1930). "A system of tone letters," *Le Maître Phonétique*, No. 30, pp. 24-27.

[10] Maddieson, I. (1990). "The transcription of tone in the IPA," *Journal of the International Phonetic Association*, vol. 20, pp. 28-32.

[11] Ladefoged, P. (1993). *A Course in Phonetics*. 3rd ed. London: Harcourt Brace Jovanovich.

[12] Pierrehumbert, J. B., & Beckman, M. E. (1988). *Japanese Tone Structure*. Cambridge, MA: MIT Press.

[13] Pitrelli, J., Beckman, M. E., Hirschberg, J. (1994). "Evaluation of prosodic transcription labeling in the ToBI framework," *ICSLP'94*, pp. 123-126.

[14] Black, A., & Campbell, N. (1995). "Predicting the intonation of discourse segments from examples in dialogue speech," ESCA Tutorial and Research Workshop on Spoken Dialogue Systems, Aalborg University.

[15] Jun, S-A. (1994). UCLA seminar.

[16] Browman, C., & Goldstein, L. (1989). "Articulatory gestures as phonological units," *Phonology*, 201-251.

[17] Local, J. K. (1992). "Modelling assimilation in non-segmental rule-free synthesis," In G. J. Docherty & D. R. Ladd, eds., *Papers in Laboratory Phonology II: Gesture, Segment, Prosody*, pp. 190-223. Cambridge, UK: Cambridge University Press.

[18] Steriade, D. (1994). "On representing segmenthood," Paper presented at FLSM 4, University of Illinois, Urbana, IL.

DEVELOPING PHONETIC SKILLS

Peter Ladefoged
Phonetics Lab, Linguistics Department, UCLA, Los Angeles, CA 90095-1543

In popular usage, when someone is said to be a good phonetician, it usually means that they can hear and describe small differences among speech sounds, and that they can produce a large number of different articulatory gestures. This is rather like saying that a good biologist is someone who is skilled at using a microscope. It would be better to think of a good phonetician as someone who has a good grasp of phonetic principles and understands the issues in speech production, perception and acoustics. But, be that as it may, it is undoubtedly useful for anyone in the field to have at least the basic phonetic skills in this popular sense, just as it is useful for most (but not all) biologists to be proficient users of microscopes.

What are the sounds that a student of phonetics should be able to produce and distinguish? There is an easy answer for those aspiring to be known as fully competent general phoneticians. They should be able to produce and distinguish all the sounds represented by the symbols of the International Phonetic Alphabet. There are two advantages in this answer. Firstly, the IPA symbols and their associated diacritics can be used to describe the vast majority of sounds in every known language and every known dialect. Secondly the organization of the symbols on the set of IPA charts constitutes, on a single page, a complete theory of phonetic description. It is not a theory with which I entirely agree, but it is a working, and fairly universally accepted, set of classificatory terms, arranged in a hierarchical structure. Each symbol stands for a certain combination of these terms. Knowing the sounds represented by all these symbols enables one to communicate to the widest possible phonetically trained audience.

Learning all the sounds represented by IPA symbols is probably an unnecessarily exotic goal for those who are, or hope to be, concerned with a more particular aspect of phonetics, such as the pronunciation or synthesis of a particular language. A suitable goal for students of phonetics of this kind (and a first goal for more general phoneticians) is to become skilled observers of their own language. This might well begin by their being taught to make a broad transcription of their own speech. Instructors might begin by asking students to transcribe lists of words that present points for discussion (*Chocolate*. 'Do you have three vowels or two? What is the quality of the last vowel?'). Then move on to short phrases with possible assimilations (*In this shop*. 'Do you have a dental nasal in the first word? Does anything happen to the consonant at the end of the second word?'). The next step is to transcribe other voices; and from a pedagogical point of view it is best to try to transcribe an instructor's speech rather than a recorded utterance. An instructor can give instant feedback, peering over what the student has written and making comments such as: 'You wrote ['ɪnpʊt] whereas I said ['ɪmpʊt]; can you hear the difference? How do you say the word *input*?' The aim is always to get people to become good observers, first of their own speech, and then of others.

In general, as instruction proceeds, students should be asked to use a progressively more narrow transcription. But there are many difficult decisions to be made concerning the level of detail required. Students transcribing field tapes need different preparation from those transcribing mother/child speech. Those evaluating speech synthesis systems have yet another task.

Whenever possible, students need to be told to look as well as to listen while transcribing speech. The importance of visual clues can be demonstrated by what happens when observers see a picture of a person saying one thing while a recording of another utterance is played; for some utterances they report that they heard what they saw rather than what the sound that was actually reproduced. In normal utterances, watching what people say gives information not only about simple things, such as whether they are saying ['ɪnpʊt] or ['ɪmpʊt], but also about the substitution of velar consonants for alveolar consonants in phrases such as *I can go* pronounced as ['aɪ kn̩ 'goʊ].

Part of teaching is a matter of breaking presuppositions. People expect to hear words pronounced more or less as they know they themselves would say them, and they are often influenced by the way they are spelled. One way of avoiding both these sources of expectations is to use nonsense words in dictation exercises. Asking students to write down a form which is a possible English word such as [skɑnzɪm] trains them to listen in an objective fashion. It can also lead to interesting discussions of why people tend to write down [skɑnzɪn], and of the voicing status of the velar stop.

Another way of highlighting aspects of utterances that often go unremarked is by asking students to say a phrase backwards. Many ways of recording an utterance onto a computer include a provision for playing back the recorded utterance either forwards as normal, or backwards. If an utterance that has been said backwards is recorded and then played backwards, it should come out forwards. But as students quickly find out, reversing 'Mary had a little lamb' ['mɛəɹɪ hæd ə 'lɪtl 'læm] and saying ['mæl l'tɪl ə dæh ɪ'ɹɜɛm] does not work out. The differences between initial and final allophones, the allophonic duration differences, and the stress and intonation all have to be taken into account.

Implicit in much of what has been said above is the notion that a student of phonetics must be able to produce as well as to hear small differences among speech sounds. The links between perception and production are very tight, so that, as Johnson, Ladefoged and Lindau (1993) have noted, we need an auditory theory of speech production just as much as a motor theory of speech perception. Training students to produce sounds is an important part of getting people to be good phonetic observers. It is certainly true that if you can produce a difference between two similar sounds, then you find it easier to hear the difference.

As part of their listening technique, many phoneticians try to repeat the utterance they are listening to. On first hearing a new phrase, I personally find it useful to try to say as much of it as possible immediately afterwards. Similarly, in fieldwork situations, when one cannot quite decide between two alternative possibilities, it is often advisable to repeat them both, asking the language consultant, for example, 'Did you say [k!ʰa] or [k!a]?' (The English phrase can often be avoided by simply holding up one finger and saying the first possibility, and then two fingers and saying the second, while looking questioning.)

This leads us to consider how to teach people to make sounds that are not in their normal repertoire. When teaching phonetic performance skills, the first thing to remember is that some people are naturally good phonetic performers and others are not; but everyone can get better with practice. It is probably like singing. Some people say that they can't sing at all, and that they never sing to themselves, even when alone in the shower or bathtub. These are likely to be people who were brought up in non-musical households, in which singing was never considered a necessary or even an appropriate thing to do. But with lots of practice (and encouragement) they could still learn to sing, perhaps not like Pavarotti, but with the possibility of performing "Happy birthday to you" without embarrassment.

Learning to be an accomplished phonetician is like learning to sing professionally; it takes a great deal of work. Most beginning students of phonetics need to spend at least an hour a day for a year or two, listening to sounds and producing them. In addition, they need to work for as much time as possible with a teacher who can correct them. If a skilled teacher is not available, then working with another student is the next best thing. An outside observer, even one with no more skills than one's own, will often be able to spot performance errors and offer feedback. Another good technique is to work with speakers of another language, provided that they are willing to be sharply critical of one's attempts to produce the sounds of their language. In fieldwork situations I have often found working with teen aged children to be especially profitable. They enjoy the role reversal implicit in

their being the strict teacher and the outsider being the student.

There is no simple way of learning to produce a set of articulatory gestures that are not part of one's native tongue. Often one just has to try things such as moving the tongue a little bit closer or further from the roof of the mouth, adding more voicing, or removing nasalization. This is where an experienced teacher is invaluable in being able to direct one appropriately. Speakers of other languages can often do little more than shake their heads in despair as one vainly tries to imitate them. Then one has to remember all the phonetic possibilities, and ask one's self 'Is the degree of voicing correct? Have I got exactly the right place of articulation? Does this speaker distinguish between apical and laminal sounds? Should it be more or less fricative? And so on, through the whole set of features in one's phonetic theory (including, of course, suprasegmental features such as tone and length to which the speaker might be sensitive).

When it comes to learning to produce what are (for most speakers of Indo-European languages) more exotic sounds such as ejectives, implosives and clicks, help from an instructor is particularly valuable. But teachers have to learn to moderate their own performance. For example, when teaching ejectives I have heard instructors produce loud, ringing examples of [p'a, t'a, k'a] which their students find confusing. Unskilled students who hear these sounds and try to produce something that sounds to them similar often produce an energetically pronounced stop with a great deal of aspiration. It is far better for the instructor to proceed more gently, making less forceful ejectives. A good technique is to start from a glottal stop in a known word such as *butter* [bʌʔə] in some forms of British English, or *button* [bʌʔən] in most forms of American English, and then superimpose an alveolar stop articulation, without building up much oral pressure. Once students get the idea of making and releasing a stop closure while making an intervocalic glottal stop, they can usually extend this gesture into one in which the glottis moves upwards (even if only slightly) and compresses the air in the oral cavity.

Similar problems arise in teaching people to make voiced implosives. When instructors make a too emphatic voiced implosive, students often respond by producing a prenasalized stop, which sounds similar to them. In this case I usually begin by trying to get students to feel the downward movement of the glottis that occurs in a fully voiced stop, and then move on to the implosive. It also helps to show them the pressure changes that occur. An ordinary drinking straw can be held with one end between the lips and the other just below the surface of a colored soft drink. It is then possible to see the pressure in the mouth decreasing and sucking liquid up into the straw when an implosive is produced.

Producing clicks in real words provides further challenges, which will be considered in the oral presentation of this paper. Most people can pronounce clicks in isolation—they are used as non-linguistic vocal gestures in a wide variety of cultures. The first difficulty that most people have in using these sounds in a language is in integrating them into the stream of speech. Next, and what is probably more difficult for many people, is to ensure that each click has what is technically known as the correct accompaniment. To understand this point it is necessary to realize that all clicks involve multiple articulations. As is shown in figure 1, there is a velar closure and another closure further forward in the mouth (on the alveolar ridge in the click illustrated). The release of the forward closure produces the sound of the click. Accompanying it is the sound associated with the velar closure, which may be, among other things, voiced or voiceless, oral or nasal, and plosive or affricate. In the oral presentation of this paper I will try to teach people to produce the Xhosa words shown in Table 1 below.

Of course, spending a few minutes in a Congress session is not enough time to learn to become a good practical phonetician. But I hope I have encouraged all of you to go home and spend some time every day, listening to recordings and producing each of the sounds on the IPA chart.

Table 1. Words illustrating contrasting clicks in Xhosa.

	DENTAL	ALVEOLAR	LATERAL
VOICELESS	**ukúk|ola**	**ukúk!oba**	**úk‖olo**
	'to grind fine'	'to break stones'	'peace'
ASPIRATED	**úkuk|ʰóla**	**ukúk!ʰola**	**ukúk‖ʰoba**
	'to pick up'	'perfume'	'to arm oneself'
BREATHY VOICED	**úkug|o͡ba**	**ukúg!oba**	**ukúg‖oba**
	'to be joyful'	'to scoop'	'to stir up mud'
VOICED NASAL	**ukúŋ|oma**	**ukúŋ!ola**	**ukúŋ‖iba**
	'to admire'	'to climb up'	'to put on clothes'
BREATHY VOICED NASAL	**ukúŋ|ola**	**ukúŋ!ala**	**ukúŋ‖oŋ‖a**
	'to be dirty'	'to go straight'	'to lie on back knees up'

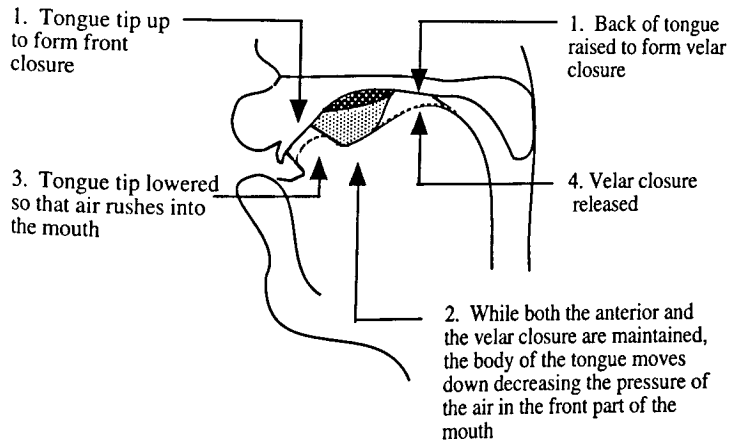


Figure 1. The actions required for producing a click.

A VIEW OF THE FUTURE OF PHONETICS

Björn Lindblom
Department of Linguistics, Stockholm University,
S-10691 Stockholm Sweden

ABSTRACT

To produce new knowledge and promote applications serving practical needs fundamental research is necessary. However, as hard times strike and research funding is cut, sponsors in government and other sectors tend to demand useful results *without* expensive "digressions" into basic science. Should the future of phonetics be entrusted to applied areas? Will phoneticians succeed in convincing sponsors of the intrinsic merits and practical necessity of their own research? The future of phonetics is in whose hands? Phoneticians still have a choice.

INTRODUCTION

Phonetician - a jack of all trades, a master of none? Or a person holding the key to a more profound understanding not only of speech, but of human language as a whole. Phonetics - a science in its own right? Intellectually, yes. There are plenty of good questions around from which to build a future phonetics. But will there be anybody to ask them in the future? For, no matter how forcefully articulated, the long-term priorities of fundamental research continually face the threat of being overruled by short-term definitions of social "needs" and of being replaced by the short-sighted agenda of "immediate usefulness". However, since answers to the core questions of phonetics have timeless and cross-cultural intrinsic value and provide the knowledge resources without which future practical applications will not be possible, prospects are good that, armed with good questions, good methods and a critical awareness of the role of external "market forces", phoneticians of the next century will be ready to meet the challenges and will find themselves contributing to one of the most central and dynamic of scientific enterprises: *Understanding human language*.

"WHAT IS A PHONETICIAN?"

At the opening of the XIIth ICPhS at Tallinn, that question was raised by Ladefoged [1] who noted that "communication engineering, physical acoustics, psychology, anatomy, physiology, linguistics, applied linguistics, computer science and poetry" are part of our lives as phoneticians.

> " ... we are phoneticians, we, the people who come to phonetics congresses, and know something about some of these diverse disciplines. None of us can know enough about all of them, *which is why being a complete phonetician is an impossible task.* But every four years we can get together and pool our knowledge. This is phonetics." (Ladefoged 1988; italics ours).

Ladefoged is right in saying that a complete mastery of all the disciplines that overlap with phonetics is an impossible task for any single individual. But is such broad knowledge really a relevant goal? Is it not the case that our interest in adjacent fields is limited to those aspects that help us answer the questions we ask? Phoneticians seek facts and insights about how speech is produced, perceived and acquired. And about how the world's sound patterns are related to the on-line phenomena of speaking, listening and learning. It seems clear that those and other questions are highly interdisciplinary presupposing bits of

knowledge coming from anthropology, biology, cognitive science, computer science and engineering, linguistics, literature and music theory, mathematics, neuroscience, philosophy, physics, sociology and several other fields. The student of sign language is in an analogous sitation.

Is a phonetician a jack of all those trades, but a master of none of them? Or a person with an agenda defined by the questions (s)he asks? Somone who makes *selective* use of information from a variety of sources? Who uses only what helps explain certain facts and makes certain measurements possible?

According to the second possibility, a phonetician is a person seeking an understanding of the issues most relevant to developing phonetic theory and who aims at acquiring *enough knowledge* about other fields to be able to extract relevant information from them and put it to productive use. "Being a complete phonetician" would still be a remote goal and a forbidding task for the individual, but not one that we could not easily cope with, given good questions, good methods and lots of colleagues to argue and interact with.

If we opt for the latter definition, what are the questions? That, of course, is one of the issues to be debated at the ICPhS 95. I would like to offer here a short sketch of my own line of reasoning and the priorities that it gives rise to.

WHY PHONETICS BELONGS IN A BIOLOGICAL FRAMEWORK

To structuralists like Saussure, the form of language was a set of social conventions shared by the members of a speech community. During the second half of the 20th century, a significant event was the appearance of Chomsky's *Syntactic Structures* which made explaning why children acquire their mother tongues the ultimate goal of linguistics. Chomsky's writings have undoubtedly been major factors in turning the focus of linguistic theory from the descriptive to the explanatory, from the group to the individual and, thus, from the social to the biological.

Seeing language as a fundamentally biological phenomenon is particularly compelling in the light of language typology and language acquisition. Language is unique to our species. There is no known human culture without language. On the surface, the world's languages vary greatly in terms of their grammar and phonetics, but behind all the geographical, historical and seemingly diverse facts, a great many structural similarities have been identified. Looking at acquisition we note that children learn to speak (or to use sign) spontaneously without conscious effort or explicit instruction. They do so in a period of time which is remarkably short in view of the complexity of what they acquire and considering the incomplete and often degraded input that reaches their ears (the "poverty of input" argument). Children who grow up in linguistically deprived environments give especially vivid examples of the alleged "information-poor" input and the "spontaneity" of the process. For instance, children surrounded by speakers of "pidgin", lack normal adult models. Nevertheless they develop "creole" languages that are more complex and more similar to normal adult languages. Also, there are reports on deaf children whose hearing parents do not master sign language well. On their own these children apparently acquire a sign grammar that is more elaborate than that of the input and closer to the normal adult model.

Facts such as these inevitably lead to the conclusion that human language could not possibly be something that a few of our ancestors thought of, and which then caught on and spread across the globe. Language is not a "cultural invention". It must be seen rather as a biologically based behavior unique to man.

IF BIOLOGY, WHAT KIND OF BIOLOGY?

To many syntacticians and psychologists, language form is complex and arbitrary, and, although all languages appear to be cut from the same cloth, their formal idiosyncracies, so the argument goes, defy functional explanation.

Leading phonologists [2] concur with this "view from syntax". Briefly stated, their claim about sound patterns is that, when everything associated with language use (production, perception, learning, memory, social factors etc) has been accounted for, there will remain a large core of phenomena, "...'Language per se'...", the innate language faculty, "which is not reducible to features of other kinds.... It is exactly this area ... that ought to occupy the central concern of linguists if they wish to arrive at an adequate conception of the essential and special nature of human Language" (Anderson 1981:495).

While fully accepting that learning language is a biological process, many behavioral scientists have not embraced the notion that language form is beyond functional explanation. Among them are phoneticians like ourselves. Our perspective on sound structure brings out the obvious - but by no means trivial - facts that all phonetic forms must be pronounceable, and that phonetic forms that differ in meaning must meet the condition of perceptual distinctiveness. Less obvious, but nevertheless true, is the fact that as these conditions, pronounceablitity and distinctiveness, interact during the development of a lexical system, they are capable of giving rise to structures of considerable complexity in completely unsupervised, self-organized ways.

The 'formalist' and the 'functionalist' views contrasted here both attribute a strong biological component to language learning. Both views share the assumption that language acquisition results from an interaction between two components: innate "predispositions" on the one hand, and experience of the ambient language, on the other. What exactly is the nature of these two components? This is where the two approaches differ in two major ways: They disagree on how the linguistic facts should be interpreted (the *arbitrary* vs *natural* issue), and on the nature of the innate "predispositions" (the *modular* vs *non-modular* issue, that is "specific to language", or "not specific to language").

To the formalist, languages are underlyingly similar but built in arbitrary and basically unnatural ways. The reason children learn language, despite its formal idiosyncracies, is that they are equipped with a 'language organ', a specialized "module" in their brains. In Chomskyan terminology: Universal Grammar, a prespecification of possible grammars from which children select their native languages by 'parameter setting'.

To the functionalist, on the other hand, language form, especially phonology, is natural, and hence normal children learn it effortlessly. As shown by a huge literature on speech development and child phonology, children develop sound structure as a result of an interaction between the linguistic input and "innate behavioral predispositions".

What is the difference, then, between the two approaches? Are Universal Grammar and "innate behavioral predispositions" two different names for the same thing? The answer is provided by how the two approaches take their stance on the naturalness and the modularity issues.

While the formalist says no to naturalness and yes to modularity, the functionalist's responses are the opposite. The functionalist assumes that, on the path to the adult phonological system, the child gets significant help from what she finds pronounceable (neuro-motoric constraints on the production of speech), what appears salient and distinctive in the speech

stimulus (auditory and perceptual constraints). Clearly, the mechanisms of hearing and the respiratory, phonatory and articulatory apparatus, are products of man's "innate endowment". But, importantly for functionalist methodology, those mechanisms are not "modular" (specific to language), since they subserve a number of other functions as well (listening to non-speech, breathing, processing food etc). (The analogous argument applies to the production and perception of sign). It is precisely at this point that the "view from syntax" diverges drastically from the "view from phonetics".

Accordingly, the functionalist hypothesis says that, by making natural movements and sounds that are adapted to production constraints, the child "fortuitously" stumbles over aspects of the adult phonology from which further, more differentiated development can then occur.

An illustration: At about six months of age children begin to produce "canonical babbling": [bababa], [dadada] etc. A simplified, but instructive account of this behavior might be given as follows. (It might be termed the *easy-way-sounds-OK* approach, or, in Swedish, *görs-lätt-hörs-rätt modellen*). The child who hears others communicate tries to participate by making articulatory gestures that are as motorically "natural" (=biomechanically low-cost) and as acoustically "salient" as possible.

Result: A vocalization with articulators in near-neutral positions combined with a mandibular open-close oscillatory movement. By doing this, the child ends up with an utterance that is not yet language, but which resembles it very strongly: [bababa], [dadada] etc. The syllable-like aspects of canonical babbling are "emergents", that is novel features arising as fortuitous consequences of a search strategy set up to scan the space of motoric possibilities from low to greater production "complexity" [3].

The point is that, in this case, children appear to get significant help, not from prespecified, "specific-to-language" information in Universal Grammar, but from general behavioral processes such as "adaptation" and "emergence". According to this interpretation the striking thing about canonical babbling is not that it shows the child coming closer to language, but rather language (phonology) being of a form that is close to the child. From the child's point of view is, in a sense, located "just around the corner".

Restating and generalizing: Is *language as a whole* learnable because it is eminently natural and reachable via processes of "adaptation" and "emergence"? Or is adult linguistic competence so hopelessly remote from where the child starts that it needs help from "specific-to-language" specializations in our genetic endowment (cf Universal Grammar)? Broadening the perspective further: To what extent should the contents of the phonetic systems that are found in the world's languages, and that are acquired by the world's children, be seen as "formal, largely prespecified, idiosyncracies". Alternatively, to what extent should they be seen as natural, behaviorally derived "adaptive emergents"?

The case for Universal Grammar rests largely on arguments from syntax. More familiar with speech processes and sound structures, phoneticians view things differently: Presumably, most of us believe that it is no accident that, in the world's languages, we find close matches between the facts of sound structure on the one hand and the phenomena of on-line speech on the other. A parsimonious (and an, in principle, uncontroversial) interpretation of such observations would be that phonological units and processes are adapted to their use in speaking, listening and learning. Implication: *Why should syntax be different*?

WHY PHONETICS HAS A PRIVILEGED ROLE

Phonetics is in a particularly good position for applying the program of contemporary biology to language. If it does, prospects are favorable for arriving at a more complete and profound explanatory theory not only of speech, but eventually of human language as a whole. Phonetics could lead the way in such an undertaking, because phoneticians have more direct access to the stuff that explanations are made of, namely facts and principles whose empirical motivation is independent of the data to be explained. Phonetics can invoke knowledge which is relevant to speech but which was acquired independently of it, often in adjacent fields, such as information on the general mechanisms of hearing and motor control, a circumstance that gives phonetics a situation that is unique compared with that of other domains of linguistic inquiry (cf syntax), and perhaps also that of many areas of biology. From that perspective being a jack of all trades turns out to be an asset, not a handicap.

THE INFLUENCE OF "MARKET FORCES" ON RESEARCH PRIORITIES

In his opening plenary address of this congress, Kohler asks: Is phonetics a language science in its own right? Indeed it is, he concludes, by virtue of the paradigm of phonetic phonology and phonetic explanation [4]. The present remarks are compatible with his views. In fact, they go further in suggesting that phonetics may even hold the key to tomorrow's linguistics.

Both Kohler's discussion and our own have a strong programmatic touch. They are as it were *in-principle scenarios* for phonetics. How viable would those (and other possible) scenarios be when confronted with the real world?

We, the people who get together at phonetics congresses, ask the questions that define our field! That may indeed be so, but what determines the questions we ask? Purely intellectual, intra-disciplinary reasons? In principle, *yes*, but, in practice, *only to some extent*. We are all shaped by the niches where we find it possible to survive academically and otherwise. Hence, even the most idealistic thinkers among us must continually adapt to a broad range of academic, economic, sociological and political factors. Internationally for many phoneticians, survival today means work oriented towards practical needs. On teachers of phonetics, there is increasing pressure to adapt curricula to the current needs of the students who are more likely to become active in applied areas than in fundamental research - that is, of course, if they get jobs at all.

So while we are free *in principle* to ask whatever questions we want and to give phonetics the directions that we ourselves favor, we are reminded that, *in practice*, it is ultimately society at all levels that significantly influences how we ask our questions. The contents of our subject matter is shaped by local and global "market forces" whether we like it or not. If there is no demand for fundamental knowledge, it is unlikely to emerge, or, if it does emerge despite all odds, it will do so much more slowly.

What is wrong with that? Why not entrust the future of phonetics to applied areas and let our fundamental understanding of speech processes develop as a spin-off from various applications, reaching us, as it were like crumbs from the rich man's table?

Our answer must be no. The following three objections should be borne in mind.

First, working in applied areas we are under absolutely no obligation to promote basic science, to solve problems so that we learn more about human speech. There is no such constraint as "basic knowledge and theoretically solid science first, then practical applications". The only objective in applications is that of solving limited and well-defined practical problems in a manner satisfying all performance criteria.

Consider an example from speech technology. Finding out how speech is produced, structured acoustically and perceived is relevant both to the phonetician and the speech technologist. However, phoneticians study human behavior, whereas speech technologists construct machines. Are these tasks basically the same? Yes and no.

Suppose we were to study birds and airplanes rather than humans and speaking machines. Obviously jumbo jets do not flap their wings. Consequently, birds and planes are built according to entirely different performance criteria. There is a parallel here with human and machine speech production. If human ears cannot tell the difference between synthetic and natural speech, but the resulting signals are made in totally different ways, should we refuse to have a certain telephone service installed that sounds all right, but happens to use speech produced by totally ad hoc and non-biological rules? Clearly that would be like waiting to fly until jumbo jets begin to flap their wings. If the telephone service is good enough from the customer's point of view, commercial forces will most certainly impose it on us whether it represents a good model of human speech or not.

Despite the possibility of potentially fruitful interactions with technology and other areas, the overall conclusion is clear: In applied phonetics, we never dig deeper than necessary to solve practical problems. In applied projects the long-term task of explaining speech represents an irrelevant detour. Shortcuts are acceptable and welcome.

Our second objection derives from those conclusions: Using applied phonetics to increase fundamental knowledge offers neither the most direct or fastest route nor any guarantees.

The third and most important objection concerns a fact that is often overlooked in current discussions of research and development. Most technical applications of today were made possible by fundamental research begun a very long time ago.

In that context, our previous metaphorical use of birds and airplanes is somewhat misleading. It gives the justification for the fact that, in applied work, the first priority is solving practical problems, not contributing to basic science.

However, before we accept that conclusion we must stop to consider *how practical problems get solved at all*. The knowledge that goes into a solution must never be taken for granted nor trivialized.

The much more significant implication of the bird-airplane metaphor is therefore this: Although planes are heavier than birds and fly faster, engineers could not have built them successfully without a thorough understanding of *aerodynamics*.

We do not need to be experts on the history of physics to realize that aerodynamics was not invented overnight. Normally, the knowledge that is being put to various practical use today took centuries to accumulate. In our own time, Gunnar Fant and others developed a theory of speech and showed how to apply it to make synthetic speech. Without wanting to detract from the considerable achievements of these pioneers, we should recognize that their efforts were anchored in a thorough understanding of acoustics, a branch of physics with a long history and with a body of knowledge to which Sir Isaac Newton (1642-1727), Jean Baptiste Joseph Fourier (1760-1830), Lord Rayleigh and many others made significant contributions [5].

The formation of knowledge embodied in scientific theories can be compared to the formation of fossil fuels. They need time to develop. We know that burning fossil fuels leads to a *depletion of resources* and poses a serious growing threat to life on this earth. Many people are therefore hard at work to promote the use of renewable energy

sources and advocate a society based on the philosophy of *recycling*.

Analogously, research funding policies that favor applied over basic research represent a kind of "depletion of resources" which must be balanced by the long-term support of general and fundamental science. On paper, that would seem to be an obvious responsibility of both state and private organizations. However, as we all know, in practice, maintaining the balance between "depletion" and "renewal" in scientific research is not achieved automatically. It presupposes a strong and active participation by the researchers themselves.

CONCLUDING REMARKS

The future strength of phonetics rests on the recognition of two main facts:

First, understanding human spoken language is understanding an important part of ourselves and of our place in nature and society. Pursuing such an undertaking successfully within the framework of general science will result in a rational account of language and speech and will show how man is to some extent unique, but basically a product of the same processes of continuous biological evolution that made all other organisms. The impact of such an account will eventually be enormous as education and communications technology spread it across the globe and to all the cultures of the world. The fact that phonetics has a privileged position in that undertaking makes phonetic research a priority of high timeless and cross-cultural intrinsic value.

Second, technological, educational, clinical and other applications cannot do without a fundamental understanding of human spoken language. Some of our colleagues would no doubt disagree. Numerous proceedings from speech technology conferences convey a strong sense of optimism about the power of computational and statistical methods that should provide shortcuts to the much slower, step-by step and experimentally based search for insights about the way humans process spoken language. The tacit hope seems to be that, before long, we will see systems that achieve speaker independent recognition of connected speech and that do so successfully although they make minimal use of phonetic, linguistic and other behavioral knowledge.

What is probability of success of such efforts? Given the complexity of spoken language, we can safely assume that such systems may score impressively on limited tasks, but are extremely unlikely to ever come near complete success in emulating human performance *unless they are based on comprehensive models of human behavior*. Assuming otherwise would seem to severely underestimate the immensity and complexity of human language. It resembles betting against other events of infinitesimally low probability, e.g. life having arisen several times in several places in the universe.

Favored by sponsors, gambling on shortcuts will no doubt continue to attract people and cost a lot of money, although it appears singularly untempting to the informed phonetician. Supporting, and doing, fundamental research seems like a much safer strategy in making phonetics useful.

REFERENCES

[1] Ladefoged P (1988): "A view of phonetics", *UCLA Working Papers in Phonetics 70*, also plenary address, International Congress of Phonetic Sciences XII, Tallinn.

[2] Anderson S R (1981): "Why phonology isn't natural", *Linguistic Inquiry* 12:493-539.

[3] Willerman R (1994): *The phonetics of pronouns: Articulatory bases of markedness*, doctoral dissertation, University of Texas at Austin.

[4] Kohler K J (1995): "Phonetics - A language science in its own right?", plenary address, International Congress of Phonetic Sciences XIII, Stockholm.

[5] Hunt F V (1992): *Origins in acoustics: The science of sound from antiquity to the age of Newton*, New York:Acoustical Society of America through the American Institute of Physics.

ACKNOWLEDGEMENTS

The author would like to thank Sue Brownlee, Catharina Kylander and Rolf Lindgren for suggesting valuable revisions of a draft of this paper.

A REALIST PERSPECTIVE ON SOME RELATIONS AMONG SPEAKING, LISTENING AND SPEECH LEARNING

Carol A. Fowler
Haskins Laboratories, New Haven CT USA

ABSTRACT

In a realist theory of speech perception, listeners perceive gestures of the vocal tract. These gestures can be shown to *be* the phonological components of utterances. Accordingly, by perceiving gestures, listeners perceive the talker's phonological message. I suggest that this tight coupling between what talkers do and what listeners perceive fosters listeners' imitation of talker's gestures and that this, in turn, fosters phonetic learning.

A REALIST PROGRAM OF RESEARCH ON SPEECH

In a realist theory of speech, speaking is a true expression of the phonological message that a talker intends to convey to a listener. That is, the phonological structures that a listener must perceive in order to recognize the speakers' words are the linguistically significant actions of the vocal tract that constitute speaking.

For their part, realist listeners hear the phonological message. They use structure in the acoustic signal, and sometimes in the optic array, not as structures to be perceived in themselves, but as information for their causal source in the world. In speech, the causal source of structure in the acoustic speech signal, and sometimes in the optic array, is, at bottom, the articulating vocal tract. As noted, however, appropriately described, the articulations of the vocal tract achieve the phonological constituents of spoken utterances. Accordingly, when listeners perceive what speakers do, they hear the talker's phonological message.

Studying the relation of speaking to listening

In any theory of speech, the relation of speaking to listening is an intimate one. Speaking and listening to speech jointly constitute the primary means by which linguistic communication can take place. However, the nature of the intimate relation of speaking to listening is different in a realist theory than in other theories, and that difference fosters a difference in the research programs that the different theoretical approaches are disposed to develop.

In most alternatives to a realist theory, phonological elements of spoken messages have their primary reality as covert, mental categories. Although these mental categories may be referred to as "phonological representations," they are not, in fact, considered to represent anything themselves. To the contrary, speakers represent *phonological categories* to listeners by moving their articulators so as to structure acoustic speech signals appropriately. In these views, articulation is a flawed vehicle for representing phonological segments, because coarticulation prevents iconic representation of their discrete, context-free character.

From this theoretical perspective, there is little to learn about listening by studying speaking aside from studying the acoustic signal that speaking creates, and speaking has, in fact, not been a central topic of investigation among most perception researchers. However, in the same way that the realist theorist, James Gibson devoted the first major section of his final book, *The ecological approach to visual perception* [1], to a description of the to-be-perceived ecological niche of a visual perceiver, so the realist investigator of speech must study what speakers do in order to understand what speaking makes available to be perceived. Therefore, studying speech production constitutes one important part of a realist research program the ultimate goal of which is to understand speech perception.

Research on speech production over the last approximately 15 years has provided an important new perspective on speaking that, however, has not changed the way that many perception researchers write about production. Investigators still cite with approval Hockett's striking metaphor in which coarticulated phonemes are likened to

smashed Easter eggs having passed through a wringer [2], or they refer to coarticulation as distortion [3]. However, in my opinion, findings on speech production show clearly that these characterizations are mistaken. Indeed, the recent findings make a central claim of a realist theory of speech perception plausible. It is that phonological primitives of spoken utterances are linguistically-significant actions of the vocal tract.

The major findings are these. Just as other components of the body do for every action that we perform, articulators of the vocal tract form transient coalitions during speech. These "synergies" are physiologically implemented couplings of articulators that are organized to achieve a task [4] or goal. Synergies are best detected in experiments in which an articulator is unexpectedly perturbed while it is moving in some direction. If the articulator is the jaw, and it is tugged down unexpectedly while it is raising for a bilabial closure [5], extra activity in a muscle of the upper lip can be detected within 20-30 ms of the perturbation, with the result that the upper lip lowers more on perturbed than on unperturbed trials, and the extra lowering compensates for the unexpectedly low position of the jaw. Biliabial closure is achieved despite the perturbation. Responses to perturbations are functional--that is, they are specifically responses that compensate for the perturbation [6-9]. That the latency of the responses is so short indicates that responses cannot arise far from the site of the perturbation. Synergies are physiological systems.

Of course, the function of synergies cannot be to counteract unexpected, externally applied tugs on the articulators. A plausible function is to compensate for internally applied tugs on the articulators arising in coarticulated speech. A low vowel coarticulated with a /b/ may tug the jaw down as it is raising for bilabial closure. The closure synergy ensures on-line compensation for that perturbation.

Synergies have two corresponding functional aspects. Their primary function is to achieve a linguistically-significant gesture. In doing that, however, in addition, they compensate for perturbing actions of coarticulating gestures. This second aspect can be elaborated further, thanks to the work on "coarticulation resistance" conducted largely by Daniel Recasens [10-13]. Recasens' work shows that synergies provide selective barriers to some coarticulatory actions. A consonantal gesture that requires a constriction to be made by the tongue dorsum will prevent or considerably reduce coarticulation by vowel gestures that also use the tongue dorsum. A consonantal gesture that does not require the tongue will not block vocalic gestures of the tongue dorsum. In short, synergies prevent coarticulation from being the destructive or distorting force that it has been characterized as being in the literature.

The import of these findings for a realist theory of speech perception has to do with the realist claim that phonological constituents of utterances are public actions of the vocal tract, not covert categories in the mind that those actions imperfectly represent. If we describe vocal-tract actions, not at the level at which we track movements of individual articulators, but at the physiologically real, coarser-grained level at which gestures are achieved (or, for some phonological segments, such as /m/ or /p/, the level at which gestural constellations are achieved), we see the context-independence required of the commutable phonological components of spoken words. Except in the most casual speech, bilabial closure is invariantly achieved when a speaker intends to produce a bilabial stop despite variation in the contributions to closure by the jaw, the lower lip and the upper lip. Coarticulation does not eliminate--indeed, synergies prevent it from eliminating--context-independence at the gestural level of description of vocal-tract actions for speech. These findings illustrate the importance for a theory of speech perception of understanding speech production.

Studying acoustic speech signals and listeners' attention to them

According to the realist theory, perception has a universal function that it must, therefore, serve the speech perceiver as well as the visual, auditory,

haptic, gustatory and olfactory perceiver. The function is to acquaint perceiver/actors with components of their ecological niche. This is an evolved function that natural selection has shaped perceptual systems to serve. Universally, then, perceivers use the structure in media that stimulate their sense organs--light for seeing, air for hearing, etc.--not as *objects* of perception, but as information for the part of the niche that caused the structure. Therefore, the second component of a realist research program is to discover how structure, largely in air, can specify gestures to perceivers.

This component of the research program, currently neglected, has to lag that on speech production, because we can only look for informative structure in the acoustic signal once we have identified the gestures and can determine how they should causally structure the air.

A final component of the realist research program is to study the perceiver's use of the acoustic speech signal. Realist perceivers should betray their use of the signal to recover gestures in two complementary ways. They should "parse" such unitary dimensions of the signal as its fundamental frequency (F0) into parts if distinct linguistic gestures have had converging effects on them. Complementarily, they should count the constellation of sometimes diverse acoustic consequences of a single gesture as a constellation that specifies the gesture.

In fact, listeners exhibit both symptoms of realist perceiving. For example, listeners judge intonational accents on high vowels (with high intrinsic F0) as lower or less prominent than accents, having the same F0, on low vowels (with low intrinsic F0) [14]. That is, they behave as if they have parsed F0 due to vowel intrinsic F0 from that due to production of the intonational melody. In turn, listeners hear the F0 that they ascribe to vowel intrinsic F0, not as pitch on the vowel, but as vowel height [15]. That is, vowel intrinsic F0 serves as part of the constellation of acoustic consequences of vowel production that provides information for the vowel.

Listeners' use of constellations to specify gestures is further indexed by their very poor discrimination, under some conditions, of syllables that differ in two ways as compared to their discrimination of syllables that differ in just one of those two ways [16]. This occurs, for example, when two syllables differ in the duration of a silent interval between [s] and [lIt] and/or in the presence or absence of labial transitions before the [lIt] sequence. Syllables with transitions are identified as "split" if the duration of silence is sufficiently long and as "'slit" if it is not. In the absence of transitions, more silence is required to shift the percept from "slit" to "split." If one syllable of a pair has a long silent interval, but no transitions whereas the second syllable of the pair has a shorter silence and transitions, listeners can fail to discriminate the pair members even though syllables differing only in presence or absence of transitions are highly discriminable. Out of context, of course, an interval of silence and transitions are highly discriminable. However, when they specify the same gesture (in the example, labial closure) in the context of a syllable, perceivers of gestures discriminate them poorly.

Studying the relation of speaking and listening to learning

There is a kind of symmetry in communicative events at the phonetic or phonological level of description: To speak is to engage in a kind of activity having linguistic significance that speakers share with members of their language community, and to listen is to perceive that activity by a speaker and to detect its significance. To listen successfully, then, is to achieve "parity" [17] in communication.

In this final section of the paper, I will suggest tentatively that some phonetic learning, which happens throughout a speaker's life, occurs due to this tight coupling between speaking and listening, which engenders a dispositon of listeners to imitate the speech they hear.

There is a striking set of findings in the literature that the discovers of the findings and I agree implies that speech listeners hear the actions of a speaker's vocal tract [18-20]. I am referring to findings by Kozhevnikov and colleagues and by Porter and colleagues regarding the latency with which listeners can imitate speakers. I will use these findings as foundations for drawing some inferences about a possible role for imitation in speech learning.

In general, in the literature on reaction time, it is well-known that "simple" reaction times are shorter than "choice" reaction times by 100 ms or more [21]. A simple reaction time procedure involves detection. For example, a subject might be instructed to press a button any time that a light flashes whether the light is red or blue or green. The subject merely has to detect the light and hit his or her one response button. In a comparable choice response task, the subject must hit a different button depending on the color of the light that flashes. Accordingly, the choice task involves not only detection, but also a mapping between the color of the light and the appropriate response button. It is not surprising that choice response times are longer than simple response times.

However, they are not always significantly longer. They were longer by a statistically nonsignificant 12 ms in the research of Porter and Lubker. In that research, the simple response task was to shift from producing the vowel [a] to another vowel [o] whenever a model speaker's vowel shifted from [a] to any of three vowels including [o]. In the choice task, listeners shifted from [a] to whatever vowel the model speaker had shifted to. Average simple response times for an [a] to [o] shift were 168 ms; corresponding average choice times were 180 ms, a nonsignificant difference.

I draw two inferences from these findings. One, following Porter and colleagues, I infer that listeners perceive vocal tract actions. The choice task involves almost no choice at all if listeners perceive the gestures of the talker, because perceiving the talker's gestures is, essentially, receiving instructions for the required response. If, instead, listeners hear the acoustic signal, the task is still a choice task: Listeners have to determine which gestures of their vocal tract will produce an acoustic signal corresponding to the one they heard.

The other inference is not warranted by these findings alone, but it is, I think, suggested by these findings considered in the context of relevant others. The inference is that imitation is dispositional, and this disposition to imitate, I propose, leads to some speech learning.

Consider two findings, one on infants and one on adults that suggest a disposition to imitate. Meltzoff and colleagues [22, 23] find that newborn infants imitate the facial expressions of adults. For example, infants will protrude their own tongue in the presence an adult protruding his or her tongue. This, of course, is especially remarkable, because the newborn cannot see its own imitation. The tendency to imitate does not go away. McHugo and colleagues [24] recorded from muscles of the face of subjects viewing a videotape of Ronald Reagan on the presidential campaign trail. Regardless of the viewers' opinion of Reagan, pro or con, when Reagan frowned on film, the corrugator muscle of the forehead, which is associated with frowning, was active. When Reagan smiled, the zygomaticus muscle at the lips was active.

Are listeners likewise disposed to imitate speakers? There is, to my knowledge, no strong evidence on this point. There is, however, evidence of vocal "accommodation"[25] whereby people speaking together may converge in their vocal intensity, speaking rate or frequency of pausing.

Recent research by Michele Sancier and me is in its preliminary stages. However, it has provided a striking outcome that suggests dispositional vocal imitation leading to speech learning in a speaker well past the critical stage of language acquisiton.

Our research was inspired both by the foregoing evidence that humans are disposed to imitate and anecdotal evidence that geographically dislocated adults may show dialectal drift toward the ambient speech of their new language community. A sample anecdote involves a young woman from Tennessee who attended a college in New England. Returning from Tennessee after the Christmas break of her freshman year, she announced to her

New England friends, in what sounded to them as a marked southern accent, that her family and friends back home had told her that she had lost her southern accent. Another example is of a colleague of Michele and mine at Haskins Laboratories who is a native British English speaker. He reported to me that his relatives in English consider him to speak with a "ghastly American accent." However, in reporting this to me, he pronounced the adjective [gastli], using a vowel that sounded British to my ears, rather than the American [ae]. In both of these cases, and many others that colleagues have reported to us, it is apparent that some drift toward the ambient speech of the language community must be occurring well after the end of the ostensible critical period for language acquisition.

When this drift occurs in speakers of a different dialect of a common language, there may be more than one source of the drift. It may occur, as we suppose, because listeners are disposed to imitate the gestures they hear. Or, instead or in addition, it may occur for social reasons. Fitting in vocally may facilitate fitting in socially.

To avoid that second possible source of gestural drift, we have been looking at the speech of a bilingual speaker. She is a native speaker of Brazilian Portuguese who attends the University of Connecticut. She is a fluent, but accented, speaker of English. She spends the academic year in Connecticut, where she speaks English almost exclusively and the summer and occasional Christmas breaks in Brazil where she speaks Portuguese almost exclusively. If this individual shows drift of her gestures in *Portuguese* toward the gestures of her English-speaking language community when she is in Connecticut, the reason is unlikely to be social. It is no social advangate for her to produce American-accented Portuguese. Accordingly, we interpret crosslinguistic gestural drift as evidence of speech learning based on a disposition of listeners to imitate perceived gestures.

The speaker had her own anecdote that led us to focus first (and, so far, only) on the voice onset times (VOTs) of her voiceless consonants. When she goes home to Brazil, her father accuses her (not in these words) of producing excessively aspirated voiceless stops. If her father is correct, then her unaspirated Portuguese voiceless stops are drifting in the direction of the aspirated voiceless stops of English.

To look for evidence of drift, we recorded the speaker on six occasions. First, we recorded her twice, approximately 24 hours apart, after she had been in Connecticut for five months and just before she left for a visit to Brazil. Next, we recorded her in two sessions one within hours of her return to Connecticut after her two month stay in Brazil and a second session one day after that. Finally, we recorded her in two sessions after she had been in Connecticut for four months.

In each session, a speaker of English read 12 sentences to her. After each sentence, our subject produced its Portuguese translation. This procedure was repeated four times so that four translations of each sentence were obtained. Compatibly, a speaker of Portuguese read 12 sentences four times each to our subject, who provided their English translations. Figure 1 provides the data on her Portuguese and English /p/s and /t/s. In the displays, findings are collapsed over each pair of sessions recorded 24 hours apart. (Generally, these sessions did not provide statistically distinguishable measures of VOT.)

Two findings are notable. First, analogous to some findings of other studies of the speech of bilinguals [26], when the speaker produces English voiceless stops, her VOTs are longer than those of her Portuguese stops. (Notice that the vertical scales for the graphs of the speaker's Portuguese and English speech both span 30 ms of VOT; however, the graph of Portuguese speech displays VOTs between 0 and 30 ms in duration, whereas that of English speech displays VOTs between 35 and 65 ms in duration.)

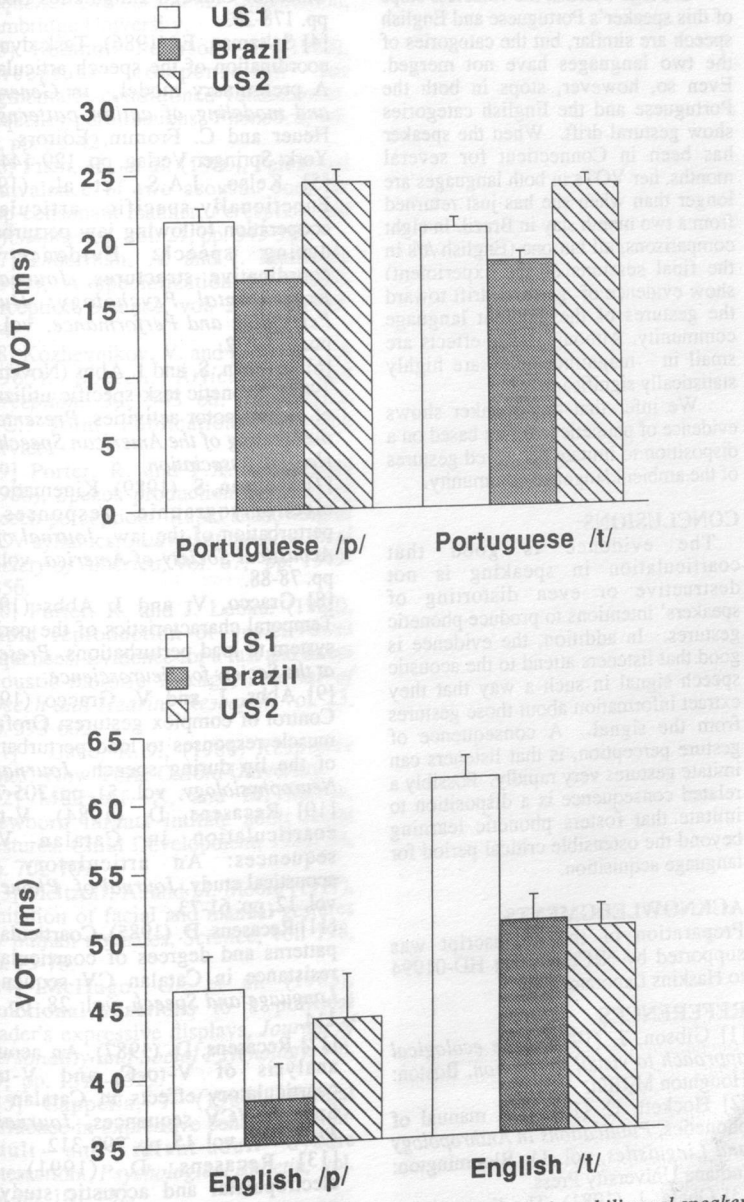


Figure 1: Voice onset times (VOTs) and standard error bars for a bilingual speaker of Portuguese and English measured at three points in time: After a several month stay in the United States (US1), after two months in Brazil, and again after several months in the US.

In Flege's terms, the voiceless stops of this speaker's Portuguese and English speech are similar, but the categories of the two languages have not merged. Even so, however, stops in both the Portuguese and the English categories show gestural drift. When the speaker has been in Connecticut for several months, her VOTs in both languages are longer than when she has just returned from a two month stay in Brazil. In eight comparisons, all but one (English /t/s in the final sessions of the experiment) show evidence of gestural drift toward the gestures of the ambient language community. Although these effects are small in magnitude, they are highly statistically significant.

We infer that this speaker shows evidence of phonetic learning based on a disposition to imitate perceived gestures of the ambient language community.

CONCLUSIONS

The evidence is good that coarticulation in speaking is not destructive or even distorting of speakers' intentions to produce phonetic gestures. In addition, the evidence is good that listeners attend to the acoustic speech signal in such a way that they extract information about those gestures from the signal. A consequence of gesture perception, is that listeners can imitate gestures very rapidly. Possibly a related consequence is a disposition to imitate that fosters phonetic learning beyond the ostensible critical period for language acquisition.

ACKNOWLEDGMENTS

Preparation of the manuscript was supported by NICHD grant HD-01994 to Haskins Laboratories.

REFERENCES

[1] Gibson, J. (1979), *The ecological approach to visual perception.* Boston: Houghton Mifflin.

[2] Hockett, C. (1955), A manual of phonetics, *Publications in Anthropology and Linguistics*, vol. 11, Bloomington: Indiana University Press.

[3] Ohala, J. (1981), The listener as a source of sound change, in *Papers from the parasession on language and behavior*, C. Masek, et al., Editor. Chicago: Chicago Linguistics Society, pp. 178-203.

[4] Saltzman, E. (1986), Task dynamic coordination of the speech articulators: A preliminary model, in *Generation and modeling of action patterns*, H. Heuer and C. Fromm, Editors. New York: Springer-Verlag, pp. 129-144.

[5] Kelso, J.A.S., et al. (1984), Functionally-specific articulatory cooperation following jaw perturbation during speech: Evidence for coordinative structures, *Journal of Experimental Psychology: Human Perception and Performance*, vol. 10, pp. 812-832.

[6] Shaiman, S. and J. Abbs (November, 1987), Phonetic task-specific utilization of sensorimotor activities. *Presented at the Meeting of the American Speech and Hearing Association.*

[7] Shaiman, S. (1989), Kinematic and electromyographic responses to perturbation of the jaw, *Journal of the Acoustical Society of America*, vol. 86, pp. 78-88.

[8] Gracco, V. and J. Abbs (1982), Temporal characteristics of the perioral system to load perturbations. *Presented at the Society for Neuroscience.*

[9] Abbs, J. and V. Gracco (1984), Control of complex gestures: Orofacial muscle responses to load perturbations of the lip during speech. *Journal of Neurophysiology*, vol.. 51, pp. 705-723.

[10] Recasens, D. (1984), V-to-C coarticulation in Catalan VCV sequences: An articulatory and acoustical study, *Journal of Phonetics*, vol. 12, pp. 61-73.

[11] Recasens, D. (1985), Coarticulatory patterns and degrees of coarticulation resistance in Catalan CV sequences, *Language and Speech*, vol. 28, pp. 97-114.

[12] Recasens, D. (1987), An acoustic analysis of V-to-C and V-to-V coarticulatory effects in Catalan and Spanish VCV sequences, *Journal of Phonetics*, vol. 15, pp. 299-312.

[13] Recasens, D. (1991), An electropalatal and acoustic study of consonant-to-vowel coarticulation, *Journal of Phonetics*, vol. 19, pp. 177-196.

[14] Silverman, K. (1987), *The structure and processing of fundamental frequency contours*, PhD Dissertation, Cambridge University:

[15] Reinholt Peterson, N. (1986), Perceptual compensation for segmentally-conditioned fundamental-frequency perturbations, *Phonetica*, vol. 43, pp. 31-42.

[16] Fitch, H., et al. (1980), Perceptual equivalence of two acoustic cues for stop-consonant manner, *Perception and Psychophysics*, vol. 27, pp. 343-350.

[17] Liberman, A. and I. Mattingly (1989), A specialization for speech perception, *Science*, vol. 243, pp. 489-494.

[18] Kozhevnikov, V. and L. Chistovich (1965), *Speech: Articulation and Perception.* 30, 543 ed. Washington D.C.: Joint Publications Research Service.

[19] Porter, R. and F.X. Castellanos (1980), Speech production measures of speech perception: Rapid shadowing of VCV syllables, *Journal of the Acoustical Society of America*, vol. 67, pp. 1349-1356.

[20] Porter, R. and J. Lubker (1980), Rapid reproduction of vowel-vowel sequences: Evidence for a fast and direct acoustic-motoric linkage, *Journal of Speech and Hearing Research*, vol. 23, pp. 593-602.

[21] Luce, R.D., (1986) *Response times.* New York: Oxford University.

[22] Meltzoff, A. and M. Moore, Newborn infants imitate adult facial gestures. Child Development, 1983. 54, pp. 702-709.

[23] Meltzoff, A. and M. Moore (1977), Imitation of facial and manual gestures by human neonates, *Science*, vol. 198, pp. 75-78.

[24] McHugo, G., et al. (1985), Emotional reactions to a political leader's expressive displays, *Journal of Personality and Social Psychology*, vol. 49, pp. 1513-1529.

[25] Cappella, J. (1981), Mutual influence in expressive behavior: Adult-adult and infant-adult dyadic interaction., *Psychological Bulletin*, vol. 89, pp. 101-132.

[26] Flege, J.E. (1987), The production of 'new' and 'similar' phones in a foreign language: Evidence for the effect of equivalence classification, *Journal of Phonetics*, vol. 15, pp. 47-65.

TOWARD A NEURODEVELOPMENTAL MODEL OF PHONETICS

R. Kent
University of Wisconsin-Madison, Madison, Wisconsin

The papers presented at this Congress give strong witness to the diversification and expansion of the phonetic sciences. These are signs that the phonetic sciences are thriving and growing. But diversification and expansion also come with a certain risk to the unity of the phonetic sciences. Unity is important not only for scholarly collegiality, but also for the larger and deeper insights that a wide embrace of knowledge permits.

Where do the phonetic sciences stand as we approach the close of both a century and a millenium? My comments are necessarily selective as the field is too large to be satisfactorily summarized in a few pages. The areas of research to be addressed here are: cross-language comparisons, speech disorders, speech development, and the relation between production and perception. The objective is to show the interrelationship among these areas in a general theory.

CROSS-LANGUAGE STUDIES

One powerful leverage in phonetic research is the cross-language investigation, which has been used quite profitably to study proclivities in production and perception. One objective in these studies is to determine universal patterns or tendencies. Another is the study of two or more languages that are selected according to a specific criterion, such as presence or absence of a phonetic feature. The compilation of data from a large number of languages is perhaps the phonetic sciences' equivalent to "big science," such as the human genome study in molecular biology. In both cross-language research and genome research, one important long-range goal is a characterization of essential human traits. Large-scale cross-language studies of the kind that explore universal patterns have been a patchwork of smaller studies conducted by individual investigators or research teams working in different nations. There is nothing wrong with this enterprise, but it is likely that international coordination of some kind would facilitate the effort. There are, after all, several thousand languages to be studied. Successful models of cross-language research include the UPSID database [1] and a recent report on Long-Term-Average Speech Spectra (LTASS) for 17 languages [2]. Quantitative studies such as the latter could address several acoustic and physiologic aspects of speech to define the universal parameters of phonation and articulation. The phonetic properties of individual languages could be described within these universal boundaries.

In the aggregate, studies of speakers of different languages reveal universal patterns in the phonetic structure of speech. Some recent studies give shape to a tentative conclusion, namely that acoustic properties of speech show some evidence of universal organization, but the congruence among languages is not sufficient to suggest a universal template of acoustic-phonetic patterns. One form of acoustic data available for several languages is the formant-frequency pattern for vowels. Krull and Lindblom [3] reported that vowels labeled with the same IPA symbol are to some degree tuned to individual languages. If so, then the IPA symbols are only generally indicative of acoustic similarity among sounds from different languages. Tuning for individual languages is not necessarily evidence against the basic principles of Stevens' [4] quantal theory of speech production, which holds that nonlinearities in the articulatory-acoustic relation determine preferences in phonetic selection. But it does argue against the hardest version of such a theory in which quantal relations in themselves would determine universal selections. It appears that the formant frequencies for individual vowels in a language are adjusted to reflect the structure of the vowel system for that language. Additional evidence for a language-tuning effect was reported by Sussman et al. [5] in their derivation of locus equations for bilabial, alveolar, and velar stops in five different languages. Although the locus equations were generally similar across languages, the differences were large enough to discourage a conclusion of universal determinism.

A unifying hypothesis for further research is that different languages divide the acoustic space for vowels into similar, but not identical, phonetic regions. Languages leave a subtle imprint on the otherwise universal articulatory-acoustic relations identified in quantal theory and similar approaches [4, 6]. This idea is consistent with evidence showing the persistence of foreign accent on vowels [7] and with accounts of the influence of early language exposure on vowel perception in infants [8]. Early language exposure appears to be a powerful determinant of phonetic organization.

DISORDERS AND DEVELOPMENT

One test of a theory's power is its capacity to address problems in several interrelated domains. Phonetic theories generally have been formulated to account for the abilities of competent adult speakers. But these theories may also be called upon to account for speech errors and for the development of speech. The following remarks focus on these two areas which often have been marginalized in phonetic theory but are assuming a more central role in the contemporary phonetic sciences.

Errors in Speech Production

The errors of speech consist of two general kinds of (possibly related) errors: (1) the lapses made by normal talkers or listeners, and (2) the clinical errors made by individuals with disorders of speech or hearing. Phonetics has been largely concerned with normal error-free performance, but the investigation of errors has been a productive pursuit. Error-free adult speech is difficult to analyze into its formative units partly because of its relatively seamless nature. But the errors of speech help to make the seams transparent. These "accidents of nature" can be exploited to gain a glimpse into the systems that underlie speech production and perception.

Sequencing errors have been studied for insights into the "slippage" points of speech production. Analyses of these errors help in understanding how segments, syllables and other units are organized, if units such as these are useful at all. But there is a worrisome possibility that the corpora of speech errors collected over the years contain a major flaw that could lead to the collapse of the entire literature. The flaw was identified by Mowrey and MacKay [9], who identified three potentially serious limitations in the traditional perceptual classification of speech errors. First, listener normalization, an integral aspect of the perceptual process, may override the detection of subtle errors. Second, errors at a fine phonetic level often cannot be detected reliably. Third, transcription techniques are not available to code highly or subtly anomalous sounds. The remedy is to use a suitable technology and a suitable phonetic representation that can register

fine variations in speech motor behavior.

Similar objections have been raised against some traditional perceptual analyses of neurogenic communication disorders such as verbal apraxia, dysarthria, and aphasia. Verbal apraxia is an object lesson in the misuse of perceptual techniques of speech analysis. The early influential descriptions of this disorder emphasized the predominance of phonemic substitution errors. These errors were interpreted to mean that the disorder was one of phonology or of phonemic selection. But subsequent studies, carried out with more refined acoustic, physiologic, and perceptual methods, have shown that the earlier analyses missed important errors at the phonetic or motoric levels [10,11,12].

The closer examination of errors in normal speech and neurogenic speech disorders points to a conceptualization in terms of gestures. These component movements are vulnerable to errors in timing and coordination. Some of these errors may give rise to apparent phonemic errors such as substitution or addition. But more subtle problems affecting individual movements and their combination may be the best focus for understanding the full range of errors in speech. Gestures offer an economical description of the multi-articulate nature of speech and can account for a variety of errors. The next phase in the study of neurogenic speech disorders may reflect a shift from reliance on global phonetic descriptors (such as broad transcription) to movement-based accounts (such as derivations of gestural scores).

Development of Speech

"There is overwhelming evidence that the emergence of coordinated movements is intimately tied both to the growth of the musculoskeletal system and to the development of the brain" [13, p. 966].

The foregoing quotation summarizes a developmental perspective on skilled movement. Speech is one form of skilled movement. As such, a central task in the understanding of speech production is to show how the coordinated movements of speech relate to the growth of the physiological system of speech and to the development of the neural systems that regulate speech movements. Seen in this way, the understanding of speech development is key to speech production and perception as a human faculty. Knowing how speech develops may be the fundamental discovery that unifies the various facets of phonetic study. The child who is developing speech is faced with two problems: (a) learning to perceive the phonetic code of the languages, and (b) learning to produce the motor patterns of the language in accord with the perceptual code.

Action theory has been highly influential in recent formulations of motor control in both developing and mature organisms. Action theory applied to speech has in the main addressed the adult speaker [14]. Action theory has found widespread application to virtually every kind of motion performed by muscular systems, including locomotion, reaching, and prehension. Action theory has emphasized particularly (a) the use of coordinative structures to solve the degrees of freedom problem common to multi-articulate systems, and (b) the biomechanical task specificity of motor responses. Action theory succeeds quite well in these two domains but it has had much less to say about the neural systems that regulate behavior.

A neurodevelopmental alternative to action theory is Edelman's theory of neuronal group selection [15,16,17]. Briefly, this theory states that repertoires of interconnected neuronal groups are established developmentally. Synaptic strengths are modified as the result of experiences including learning. Initially, large repertoires of variant neural

circuits are established by selectional mechanisms in the developing embyro. These are called primary repertoires. Continuing selectional events serve to enhance neuronal responses that have adaptive value for the organism. These are called secondary repertoires. An important concept in Edelman's theory is "re-entrant signalling," an exchange of signals along parallel and reciprocal connections among neuronal groups. The theory of neuronal group selection has been used to account for perception, motor responses, language, and consciousness. This theory places an emphasis on neural processes that is either lacking or vaguely described in accounts of action theory. The selectionist account offers a promising framework for a neural theory of phonetics, some aspects of which will be addressed in the balance of this paper.

PRODUCTION AND PERCEPTION

One of the most recalcitrant problems in the phonetic sciences is the unification of production and perception. But there is little agreement on how unity can be acheived. One indication of the sluggish progress is the nearly separate development of theories of speech perception and speech production. There are some notable attempts to bring perception and production under the same theoretical umbrella, especially the motor theory of speech perception [18,19]. Another attempt at unification is the theory of event perception applied to speech [20], which found theoretical resonance with action theory [14].

Ultimately, the integration of production and perception should be evident in the same neural mechanisms that explain phonetic development. One possibility for sensory-motor interaction is the proximity of motor and sensory neurons in the CNS representation of various parts of the body. Huang et al. [21] found that the auditory area of Crus II of the cerebellar hemisphere in both rat and cat is surrounded by orofacial somatosensory receptive fields. Furthermore, the cerebellar granule cells in the posterior vermis and the hemispheres exhibited phasic responses to auditory stimuli. This could mean that the cerebellum is involved in event timing, a possibility supported by clinical studies showing that a primary feature of the speech of persons with cerebellar disease is disordered timing. The speech disorder, known as ataxic dysarthria, is commonly described as having altered patterns of syllable timing or stress [22,23]. Moreover, an hypothesis of cerebellar regulation of event timing has been advanced by Keele and Ivry [24] who proposed that a major cerebellar function is to provide temporal computations that underlie a variety of perceptual and motor tasks. One role of the cerebellum in speech may be as a neural timekeeper for the registration of sensory and motor information.

Some cortical neurons also respond to more than one type of stimulation. Bruce, Desimone, and Gross [25] reported that the majority of neurons in the superior temporal sulcus of macaques were sensitive to more than one modality. Only 41% were exclusively visual. The remainder responded to visual-auditory, visual-somesthetic, or visual-auditory-somesthetic stimulation. Neurons that respond to more than one modality may be the means for a convergence of bisensory or multisensory peripheral stimulation. Separate sensory channels can dissociate stimuli by modality, but multisensory convergence in the CNS allows integration across modalities [26]. This convergence is highly relevant to speech, which has a plurimodal sensory foundation of audition, taction, kinesthesia, and baroreception [27].

Sensory-motor integration also could

be based on neural representations formed in a complex neural circuitry. Song birds are a good animal model for speech for several reasons, especially in that many avian species learn their songs through exposure and practice. Song learning apparently involves a comparison of the young bird's calls with those of the adult conspecific birds. Williams [28] described the neural circuitry underlying birdsong development as having multiple loops, multiple modalities, multiple representations of song, and multiple neural mechanisms. She identified two primary branches of the control system. The first branch descended from the control center to the motoneurons for respiration and vocalization. The nuclei of this branch continue to grow during song development and seem to be indispensable for adult song. The nuclei of this branch have long and variable latencies. The second branch forms a recursive loop between two nuclei of the descending branch. It contains one nucleus that reaches its maximum size in early song learning and then declines. This branch has short latencies with little variability. Auditory activity can be recorded in both branches, including the motoneurones controlling the vocal organ. Williams proposed that the short-latency recurrent branch allows for comparisons of different song representations. If these results can be extended to speech development, the implications include the following.

(1) Different parts of the neural circuitry reach their maximum sizes at different ages. This maturational variation could be related to the concept of a "sensitive period" for speech development [29,30]. Potential for language learning may be related to the maturational gradients of the complex neural circuitry.

(2) The neural circuitry that supports speech sound learning may contain auditory, motor, tactile, and kinesthetic representations. The existence of these different representations may explain the robustness of speech production in the face of various attempts to reduce or modify sensory feedback [27,31].

(3) Recurrent branches provide a means for the comparison of these different representations of speech, although the use of these comparative mechanisms varies with stage of development and task demands. Sensory disruption is more damaging to the speech of children than to the speech of adults [27], which may reflect the adults' capability to switch easily and effectively between various representations. Speech development may involve the acquisition of different neural representations and the facility to select among them as task conditions change.

This perspective is consistent with Edelman's [15,16,17] theory of neuronal group selection, and particularly with his idea of re-entrant signaling. Speech may be represented as a number of neuronal "maps" that combine different kinds of sensory and motor information. In its global nature, speech is defined by the totality of these maps and their interactions. More narrowly, speech can be defined by interactions among selected maps. Therefore, speech is auditory-visual (as the McGurk effect [32] demonstrates), and tactuo-motor (as in the haptic communication system employed by users of Tadoma, who can understand speech from tactile cues gathered from a hand placed over the talker's face and neck [33]). This idea also accords with models of auditory processing that emphasize the temporal properties of the global neural response [34]. A primary advantage of re-entrant signaling is that it provides a temporal coherence of related sensory and motor information. This coherence is a fundamental neural correlate of phonetic events. It also offers a useful interpretation of speech rhythm as the basic temporal plan for the coordination of sensory and motor information in speech production and perception.

POINTS OF CONVERGENCE

Principles that apply to different domains in the phonetic sciences have good potential for the delineation of universal tendencies. One important convergence is that the sounds prominent in babbling also are prominent in the world's languages [30,35]. Furthermore, these sounds also tend to be preserved in neurogenic speech disorders [36,37]. These universal patterns can be explained by a combination of factors including: the developmental anatomy of the vocal tract, nonlinearities in the articulatory-acoustic relation, evolutionary influences expressed as primary movement repertoires, and perceptual contrast.

Another property of speech that is evident from several vantage points is what Fujimura [38] referred to as the "inherently multidimensional principle of speech coordination" (p. 218). This principle is a kernel concept in theories of speech production [14] but is also recognized in various nonlinear phonologies [38] and articulatory phonology [39]. This property of the motor speech system appears to offer valuable insights into speech development [40] and speech disorders [41], and may help to resolve some the problems that have been described with corpora of segment errors in normal and disordered speech [9,12].
Multidimensionality can be expressed as the regulation of component gestures. The concept of gesture is becoming increasingly central in discussions of speech motor control, phonology, speech development in children, and speech disorders. Smith [31] noted, "The convergence of theories onto the notion of a basic articulatory gesture suggests that it is an idea with intrinsic merit" (p. 261). But she added: "Despite the general appeal of the idea, a disproportionate amount of theory development has occurred in relation to the data available that clearly support the notion" (p. 261). This lack of proportion between theory and data has been a handicap, but the new technologies and the increasing availability of databases [38] may help to strike the balance in the next millenium.

ACKNOWLEDGMENTS

This research was supported by NIH research grants DC00319 and DC0082 from the National Institute on Deafness and Other Communication Disorders, National Institutes of Health.

REFERENCES

[1] Maddieson, I. (1984), Patterns of sound. Cambridge, England: Cambridge University Press.
[2] Byrne, D., Dillon, H., Tran, K., Arlinger, S., et al. (1994). An international comparison of long-term average speech spectra. Journal of the Acoustical Society of America, vol. 96, pp. 2108-2120.
[3] Krull, D., & Lindblom, B. (1992). Comparing vowel formant data cross-linguistically. PERILUS (Phonetic Experimental Research, Institute of Linguistics, University of Stockholm), No. XV, pp. 7-15.
[4] Stevens, K. (1989). On the quantal nature of speech. Journal of Phonetics, vol. 17, pp. 3-45.
[5] Sussman, H. M., Hoemeke, K. A., & Ahmed, F. S. (1993). A cross-linguistic investigation of locus equations as a phonetic descriptor for place of articulation. Journal of the Acoustical Society of America, vol. 94, pp. 1256-1268.
[6] Badin, P., Perrier, P., Boe, L.-J., & Abry, C. (1990). Vocalic nomograms: Acoustic and articulatory considerations

upon formant convergences. Journal of the Acoustical Society of America, vol. 87, pp. 1290-1300.
[7] Munro, M.J., Flege, J.E., & MacKay, I.A.R. (in press). The effects of second language learning on the production of English vowels. Applied Psycholinguistics.
[8] Kuhl, P.K., Williams, K.A., Lacerda, F., Stevens, K.N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. Science, vol. 255, pp. 606-608.
[9] Mowrey, R.A., & MacKay, I.R.A. (1990). Phonological primitives: Electomyographic speech error evidence. Journal of the Acoustical Society of America, vol. 88, pp. 1299-1312.
[10] Itoh, M. Sasanuma, S., Tatsumi, I.F., Murakimi, S., Fukusako, Y., & Suzuki, T. (1982). Voice onset time characteristics in apraxia of speech. Brain and Language, vol. 17, 193-210.
[11] Ziegler, W. (1987). Phonetic realization of phonological contrasts in aphasic patients. In. J.H. Ryalls (Ed.), Phonetic approaches to speech production in aphasia and related disorders (pp. 163-179). Boston: College-Hill.
[12] McNeil, M.R., & Kent, R.D. (1990). Motoric characteristics of aphasis and apraxic speech. In G.R. Hammond (Ed.), Advances in psychology: Cerebral control of speech and limb movements (pp. 349-387). Amsterdam: North Holland.
[13] Sporns, O., & Edelman, G.M. (1993). Solving Bernstein's problem: A proposal for the development of coordinated movement by selection. Child Development, vol. 64, pp. 960-981.
[14] Kelso, J.A.S., Saltzman, E.L., & Tuller, B. (1986). The dynamical persective on speech production: Theory and data. Journal of Phonetics, vol. 14, pp. 29-59.
[15] Edelman, G. (1993). Neural Darwinism: Selection and Reentrant signaling in higher brain function. Neuron, vol. 10, pp. 115-125.
[16] Montague, P.R., Gally, J.A., & Edelman, G.R. (1991). Spatial signaling in the development and function of neural connections. Cerebral Cortex, vol. 1, pp. 199-220.
[17] Friston, K.J., Tononi, G., Reeke, G.N., Jr., Sporns, O., & Edelman, G.M. (1994). Value-dependent selection in the brain: Simulation in a synthetic neural model. Neuroscience, vol. 59, 229-243.
[18] Liberman, A.M., Cooper, F.S., Shankweiler, D., & Studdert-Kennedy, M. (1967). Perception of the speech code. Psychological Review, vol. 74, pp. 4311-4361.
[19] Liberman, A.M., & Mattingly, I.G. (1985). The motor theory of speech perception revised. Cognition, vol. 21, pp. 1-36.
[20] Fowler, C.A. (1986). An event approach to the study of speech perception from a direct-realist perspective. Journal of Phonetics, vol. 14, pp. 3-28.
[21] Huang, C., Hsiao, C.F., Yang, B., & Mu, H. (1991). Auditory receptive area in the cerebellar hemisphere is surrounded by somatosensory areas. Brain Research, vol. 541, pp. 251-256.
[22] Darley, F.L., Aronson, A.E., & Brown, J.R. (1969). Cluster of deviant speech dimensions in the dysarthrias. Journal of Speech and Hearing Research, vol. 12, pp. 462-496.
[23] Kent, R.D., Netsell, R. and Abbs, J. (1979). Acoustic characteristics of dysarthria associated with cerebellar disease. Journal of Speech and Hearing Research, vol. 22, pp. 627-648.
[24] Keele, S.W., and Ivry, R. (1990). Does the cerebellum provide a common computation for diverse tasks? In A. Diamond (Ed.), The development and neural bases of higher cognitive functions. Annals of the New York Academy of Sciences, vol. 608, pp. 179-211.
[25] Bruce, C., Desimone, R., & Gross, C.G. (1981). Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. Journal of Neurophysiology, vol. 46, pp. 369-384.
[26] Stein, B.R., & Meredith, M.A. (1990). Multisensory integration: Neural and behavioral solutions for dealing with stimuli from different sensory modalities. In A. Diamond (Ed.), The development and neural bases of higher cognitive functions, Annals of the New York Academy of Sciences, vol. 608, pp. 51-70.
[27] Kent, R.D., Martin, R.E., & Sufit, R.L. (1990). Oral sensation: A review and clinical prospective. In H. Winitz (Ed.), Human communication and its disorders: A review, Vol. 3 (pp. 135-192). Norwood, NJ: Ablex.
[28] Williams. H. (1989). Multiple representations and auditory-motor interactions in the avian song system. In M. Davis, B.L. Jacobs & R.I. Schoenfeld (Eds.), Modulation of defined vertebrate neural circuits. Annals of the New York Academy of Sciences, vol. 563, pp. 148-164.
[29] Lenneberg, E. (1967). Biological foundations of language. New York: Wiley.
[30] Locke, J. (1993). The child's path to spoken language. Cambridge, MA: Harvard University Press.
[31] Smith, A. (1992). The control of orofacial movements in speech. Critical Reviews in Oral Biology and Medicine, 3, 233-267.
[32] McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. Nature, 264, 746-749.
[33] Reed, C.M., Durlach, N.I., Braida, L.D., & Schultz, M.C. (1989). Analytic study of the Tadoma method: Effects of hand position on segmental speech perception. Journal of Speech and Hearing Research, vol. 32, pp. 921-929.
[34] Ghitza, O. (1988). Temporal non-place information in the auditory-nerve firing patterns as a front-end for speech recognition in a noisy environment. Journal of Phonetics, vol. 16, pp. 109-124.
[35] MacNeilage, P.F. (1994). Prolegomena to a theory of the sound pattern of the first spoken language. Phonetica, vol. 51, pp. 184-194.
[36] Klich, R.J., Ireland, J.V., & Weidner, W.E. (1979). Articulatory and phonological aspects of consonant substitutions in apraxia of speech. Cortex, vol. 15, pp. 451-470.
[37] Marquardt, T.P., Reinhart, J.B., & Peterson, H.A. (1979). Markedness analysis of phonemic substitution errors in apraxia of speech. Journal of Communication Disorders, vol. 12, pp. 481-494.
[38] Fujimura, O. (1990). Methods and goals of speech production research. Language and Speech, vol. 33, pp. 195-258.
[39] Browman, C.P., & Goldstein, L. (1992). Articulatory phonology: An overview. Phonetica, vol. 49, pp. 155-180.
[40] Thelen, E. (1991). Motor aspects of emergent speech: A dynamic approach. In N.A. Krasnegor et al. (Eds.), Biological and behavioral determinants of language development (pp. 339-362). Hillsdale, NJ: Erlbaum.
[41] Weismer, G., Tjaden, K., & Kent, R.D. (in press). Can articulatory behavior in motor speech disorders be accounted for by theories of normal speech production? Journal of Phonetics.

SPEECH RECOGNITION BY MACHINES

Bishnu S. Atal
AT&T Bell Laboratories, Murray Hill, NJ, U.S.A.

ABSTRACT

Voice communication with machines is no longer a dream but a reality. The tremendous progress that has been accomplished has come about as a result of solving successfully some of the fundamental problems caused by the immense variability present in the speech signal. This paper will discuss important issues in automatic recognition of speech, the major advances made in this field, the current state of the technology, and future developments.

INTRODUCTION

Speech is a natural form of communication for humans and thus the problem of understanding or "recognizing" speech by machines has challenged scientists for many years. We do not yet understand in any detail how humans understand speech. But, considerable advances in automatic speech recognition and understanding by machines have taken place [1-2].

Research in automatic speech recognition (ASR) since the 1970s has produced solutions for increasingly difficult tasks, from the correct recognition of a few isolated words from a single speaker to recognition of fluent speech from virtually any speaker. Progress in automatic recognition of speech is continuing and the research frontiers are shifting towards the solution of an even harder problem -- unconstrained dialogue with machines. The availability of high-speed processors and high density memories at reasonable cost in digital computers, along with large databases of recorded speech, has made it possible to develop sophisticated signal representations, pattern-matching techniques, and language models in support of automatic speech recognition.

COPING WITH ACOUSTIC VARIABILITY

Speech is the acoustic form of language. Speech recognition is essentially a process of recognizing acoustic patterns of the spoken language. Human communication by voice appears to be so simple that we often forget how variable these acoustic patterns are. Vast differences occur in the spoken utterance dependent on context, speaking style, speaker, dialect, speaking environment, microphone characteristics, etc. The major obstacle to achieving high accuracy in speech recognition is the large variability present in the speech signal, with only a small part that is important for carrying the linguistic information. A large part of this variability is due to various redundancies introduced by the human speech production process to achieve reliable speech communication in noisy and reverberant acoustic environments. Other sources of variability are introduced by differences in the vocal systems of speakers, differences in speaking rates, and the influence of neighboring sounds on the acoustic realization of a particular sound due to sluggish articulatory movements. Automatic methods of speech recognition must be able to handle this large variability in a proper manner. We illustrate here a few examples of the variability inherent in the speech signal.

Consider a simple case of the same word spoken by the same speaker on two different occasions. Acoustics realizations of the two utterances are in general not identical due to variations in the speaking rate or the speaking style. The

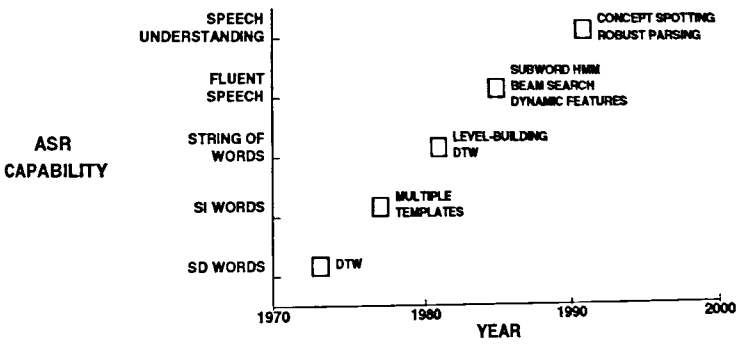


Figure 1. Major advances in automatic speech recognition from 1970 onwards that made it possible to achieve steady progress, from the simple problem of speaker-dependent recognition of isolated words to the more difficult problem of recognizing spontaneous speech encountered in dialogues.

speech recognition procedure must be able to compare the two utterances and conclude that they are same. Variations in speaking rate cause nonlinear distortions of the time axis of speech patterns. Linear scaling of the time axis is generally not sufficient to cope with speaking-rate variability. Nonlinear time normalization using dynamic programming is necessary to achieve time alignment between unknown and reference utterances.

The same word spoken by two different speakers will in general have different acoustic characteristics, due to the differences in their vocal tracts and speaking styles. A speaker-dependent recognizer uses the utterance of a single speaker to learn the speech patterns of that speaker. In contrast, a speaker-independent recognizer is trained on speech from many speakers and is used to recognize speech from speakers that may be outside of the training population.

Recognition of continuous speech introduces additional problems. In isolated words or speech where words are separated by distinct pauses, the beginnings and ends of words are clearly marked. In continuous speech, word boundaries are blurred and words evolve smoothly in time with no acoustic separation. Automatic methods of segmenting continuous speech into words therefore had to be devised. Machine recognition of continuous speech with a large vocabulary requires that syntactic and semantic constraints be incorporated in the recognition process.

The progress in automatic speech recognition has come about as a result of solving successfully problems created by the large variability present in speech signals. Figure 1 shows some of the major advances that were made during the past twenty years and were important in achieving this progress. Dynamic time warping (DTW) was the most important step taken in early 1970s to handle variations in speaking rates [3]. Clustering of speech patterns into multiple templates [4] for each word made it possible to recognize words spoken by any speaker (speaker-independent speech recognition). Recognition of individual words from a string of connected words required development of level-building

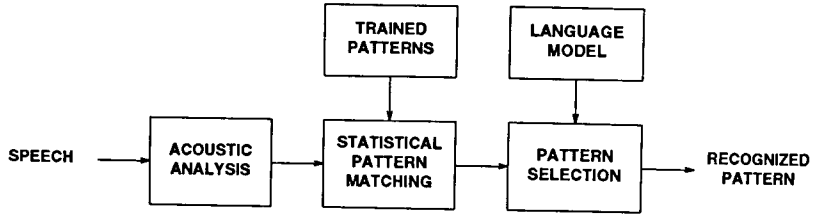


Figure 2. A block diagram showing the basic steps of the pattern recognition approach.

ASR algorithms [5]. The introduction of Hidden Markov Models (HMM) techniques [6] in early 1980s was a giant step that set the stage for big success in ASR. Recognition of continuous fluent speech was made possible by breaking words into phone-like subword units, use of bigram and higher-order language models, and the development of efficient beam search techniques [7]. Finally, introduction of concept spotting [8] and robust parsing techniques moved ASR from simple recognition of words to understanding meaning conveyed by a group of words.

AUTOMATIC SPEECH RECOGNITION PROCESS

There are at present three principal approaches to speech recognition: The first is based on statistical techniques of pattern recognition [6] that utilizes a training set of speech data to learn important information about the speech signal. The second approach, commonly known as acoustic-phonetic approach, uses knowledge of the relationship between acoustic and phonetic structures of the language. The third approach uses artificial neural networks. Best results in speech recognition have so far been achieved by using the statistical pattern recognition approach supplemented by the knowledge of acoustic-phonetic relationships in speech.

The basic steps of the pattern recognition approach are illustrated in the block diagram of Fig. 2. The speech signal is analyzed to provide a parametric representation at the acoustic level. These parameters ("features") are then compared to a stored set of patterns derived from a large collection of speech utterances from many speakers using a HMM-based training procedure. This comparison provides a set of scores representing the similarity between the unknown pattern and each of the stored patterns. The last step augments these scores with other knowledge about the speech utterance, such as the language, the context, and semantics, to yield the best recognition results.

Acoustic Features of Speech

The selection of proper acoustic or spectral features is crucial for achieving high performance in speech recognition. The short-time spectral envelope of speech, obtained either by filtering or linear prediction analysis, is still considered to be the most effective representation for speech recognition, especially if rendered on a critical-band ("Bark") scale. The spectra are computed sequentially in time at intervals of 10 to 20 ms and are usually converted into cepstral coefficients. The cepstrum is defined as the inverse Fourier transform of the logarithm of the power spectrum.

The cepstral coefficients are instantaneous (static) features. One of the most important advances in the acoustic representation of speech has been the introduction of dynamic features [9], such as first- and second-order derivatives of the

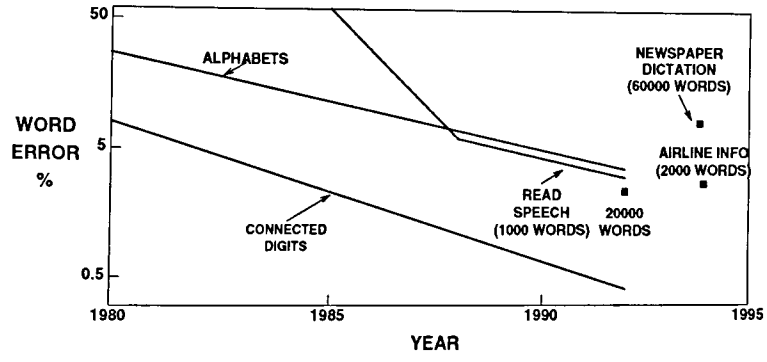


Figure 3. Reduction in the word error rate for different automatic speech recognition tasks between 1980 and 1995.

cepstrum. The static and dynamic features are generally combined to form a larger feature set; a smaller set can then be obtained by proper "pruning" from the larger set.

Training and Pattern Matching

Because of its statistical nature and its simple algorithmic structure for handling the large variability in speech signals, the HMM approach has found widespread use for automatic speech recognition [10]. Most of the successful systems today are based on this approach. An HMM representation can be used to model a sound, a word, a phrase, or a long utterance. During the training phase, the HMMs are trained from an ensemble of observation vectors coming from spoken utterances and stored for each of the basic pattern to be recognized. During recognition, the unknown acoustic patterns are compared with a set of stored reference patterns ("templates") established from the training data to provide a set of similarity scores between the test and reference patterns.

Pattern Selection

In the recognition phase, the utterance is decoded by determining the optimal sequence of HMM states and the corresponding speech units based on the observed sequence of acoustic feature vectors in the utterance. Search procedures based on dynamic programming methods are used to find the sequence of states with the maximum likelihood. Additional information based on the syntax and semantics of the source language is included in the recognition process to produce admissible outputs.

CURRENT CAPABILITIES

The performance of ASR systems continues to improve steadily. Figure 3 shows the word error rate for various test materials and the steady decrease in the error rate achieved since 1980. The performance of current ASR systems degrades considerably in the presence of noise, reverberation, or distortion and for conversational speech.

There are many factors that influence the performance of automatic speech recognition systems. The most important of these are the size of the vocabulary and the speaking style. Figure 4 shows examples of ASR tasks that can be handled by automatic methods for different vocabulary sizes and speaking styles. Generally, the number of confused words increases with the vocabulary size. Current systems can properly recognize a

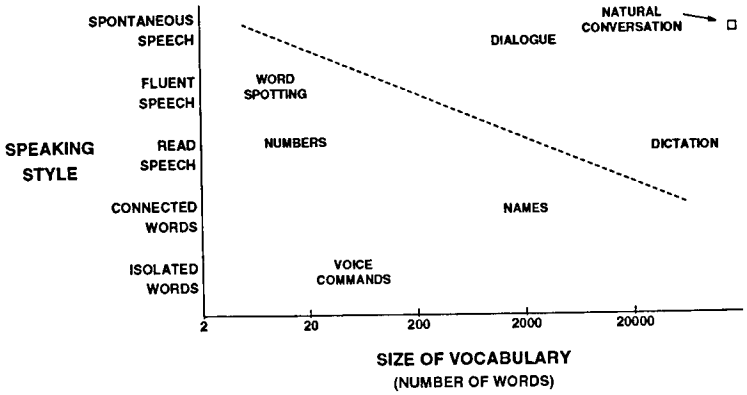


Figure 4. Different speech recognition tasks shown in a space of two dimensions: speaking style and size of vocabulary. Tasks that can be handled by current technology are shown to the left of the diagonal line. Items to the right of the line need more research to bring performance to a useful level.

vocabulary of as many as a several thousand words, while the speaking style can vary over a wide range, from distinct isolated words to spontaneous running speech.

Examples of speech recognition applications that can be handled by the current technology are shown on the left side of the diagonal line in Fig. 4. These include recognition of voice commands for computers, names, digit strings, and keyword spotting. The items on the right of the diagonal line in Fig. 4 are examples of the speech recognition tasks that work well in laboratory environments but which need more research to become useful for real-world applications. Automatic recognition of fluent speech with a large vocabulary is not feasible unless constraints on syntax and semantics are introduced. The present capability in handling natural languages and in following a dialogue is limited because we do not understand how to model the variety of expressions that natural languages use to convey concepts and meanings.

CHALLENGING ISSUES IN SPEECH RESEARCH

For speech technology to be used widely, it is necessary that the major roadblocks faced by the current technology be removed. Some of the key issues that pose major challenges in speech research are listed below:

- *Robust performance*. Can the recognizer work well for different speakers and in the presence of the noise, reverberation, and spectral distortion that are often present in real communication channels?
- *Automatic learning of new words and sounds*. In real applications, the users will often speak words or sounds that are not in the vocabulary of the recognizer. Can it learn to recognize such new words or sounds automatically?
- *Grammar for spoken language*. The grammar for spoken language is quite different from that which is used in carefully constructed written text. How does the system learn this grammar?
- *Flexibility*. Unless it is flexible, speech technology will have limited applications. What restrictions are there on the vocabulary? Can it handle spontaneous speech and natural spoken language?

A number of methods have been proposed to deal with the problem of robustness. The proposed methods include signal enhancement, noise compensation, spectral equalization, robust distortion measures, and novel speech representations. These methods provide partial answers valid for specific situations, but do not provide a satisfactory answer to the problem. Clean, carefully articulated, fluent speech is highly redundant, with the signal carrying significantly more information than is necessary to recognize words with high accuracy. However, the challenge is to realize the highest possible accuracy when the signal is corrupted with noise or other distortions and part of the information is lost.

FUTURE DEVELOPMENTS

The advances in digital technology is rapidly changing the fabric of telecommunications and the way we access information. A new mode of interacting with computers through voice is emerging. When combined with video, the voice mode offers an easy natural communication interface with computers. Speech recognition technology is a key component of such an interface. Human speech communication is a complex process and it will require scientific understanding of many other issues beyond acoustics and pattern recognition to mimic this process in computers. Speech science is expanding its frontiers to answer the basic question of how we put words together to express ideas in the spoken language.

REFERENCES

[1] Rabiner, L. and Juang, B.-H. (1993), *Fundamentals of speech recognition*, Englewood Cliffs: Prentice Hall.

[2] Makhoul, J. and Schwartz, R. (1994), "State of the art in continuous speech recognition", in D. Roe and J. Wilpon (eds.), *Voice communication between humans and machines*, Washington: National Academy Press, pp. 165-198.

[3] Itakura, F. (1975), "Minimum prediction residual principle applied to speech recognition", *IEEE Trans. ASSP*, vol. ASSP-23, pp. 57-72.

[4] Rabiner, L. Levinson, S., Rosenberg, A., Wilpon, J. (1979), "Speaker independent recognition of isolated words using clustering techniques", *IEEE Trans. ASSP*, vol. ASSP-27, pp. 336-349.

[5] Meyers, C. and Rabiner, L. (1981), "A level-building dynamic time warping algorithm for connected word recognition", *IEEE Trans. ASSP*, vol. ASSP-29, pp. 284-297.

[6] Bahl, L., Jelenik, F., and Mercer, R. (1983), "A maximum likelihood approach to continuous speech recognition", *IEEE Trans. Patt. Anal. Machine Intell.*, vol. PAMI-5, pp. 179-190.

[7] Chow, Y. et al. (1987), "BYBLOS: The BBN continuous speech recognition system", *Proc. IEEE-ICASSP*, Dallas, TX, pp. 89-92.

[8] Pieraccini, R. and Levin, E. (1992), "Stochastic representation of semantic structure for speech understanding", *Speech Communications*, vol. 11, pp. 283-288.

[9] Furui, S. [1986], "Speaker-independent isolated word recognition using dynamic features of speech

ICPhS 95 Stockholm Session 93.4 Vol. 1 Page 493

spectrum", *IEEE Trans. ASSP*, vol. ASSP-34, pp. 52-59.

[10] Rabiner, L. (1989), "A tutorial on hidden Markov models and selected applications in speech recognition", *Proc. IEEE*, vol. 77, pp. 257-286.

Author index - ICPhS 95 - Stockholm